Invited Sessions I - 11 December 2024, 14:30 - 16:00

Session 08	**Recent Development on Econometrics and Time Series Analysis** *Organizer/Chair: Guodong Li, The University of Hong Kong*	E22-2002	11 Dec, 14:30 - 16:00

A Composite Likelihood-based Approach for Change-point Detection in Spatio-temporal Processes

Chun-Yip Yau, The Chinese University of Hong Kong

This paper develops a unified and computationally efficient method for change- point inference in non-stationary spatio-temporal processes. By modeling a non-stationary spatio-temporal process as a piecewise stationary spatio-temporal process, we consider simultaneous estimation of the number and locations of change-points, and model parameters in each segment. A composite likelihood-based criterion is developed for change-point and parameter estimation. Under the framework of increasing domain asymptotics, theoretical results including consistency and distribution of the estimators are derived under mild conditions. In contrast to classical results in fixed dimensional time series that the localization error of change-point estimator is O_p(1), exact recovery of true change-points can be achieved in the spatio-temporal setting. More surprisingly, the consistency of change-point estimation can be achieved without any penalty term in the criterion function. In addition, we further establish consistency of the change-point estimator under the infill asymptotics framework where the time domain is increasing while the spatial sampling domain is fixed. A computationally efficient pruned dynamic programming algorithm is developed for the challenging criterion optimization problem. Extensive simulation studies and an application to the U.S. precipitation data are provided to demonstrate the effectiveness and practicality of the proposed method.

High Dimensional Spatio-Temporal Autoregressive Models for Matrix-Valued Time Series

Baojun Dou, City University of Hong Kong

This paper explores the modeling of spatio-temporal matrix time series data, driven by the need to predict daily trading volume curves for various assets. These predictions are critical inputs for execution algorithms like volume-weighted average price (VWAP) and volume inline, which major execution brokers use to manage large buy or sell orders for institutional clients. We examine two subclasses of models: the first draws from spatial econometric conventions with predetermined weight matrices, while the second avoids the challenges associated with predetermined weights by treating them as unknown parameters with banded sparse structures derived from practical applications. To address the inherent endogeneity, we utilize the iterated least squares method based on Yule-Walker equations for estimation. Theoretical foundations and asymptotic results for the proposed methods are established for both fixed and high dimensions. The methodology is further demonstrated through both simulated and real-world datasets.

Testing Exogeneity in Moderately High-dimensional Linear Regressions

Chen Wang, The University of Hong Kong

Exogeneity is a cornerstone assumption in regression analysis to achieve efficient inference. This paper studies testing exogeneity in high-dimensional linear regression models, where the number of regressors is asymptotically proportional to the sample size. We establish that the widely used Durbin-Wu-Hausman (DWH) test can suffer from severe size distortions in this context. To overcome such a limitation, we propose a novel test statistic based on the DWH test statistic, which maintains correct sizes in both fixed and high dimensions. Furthermore, we extend the proposed test to regressions that *further include high-dimensional exogeneity. Monte Carlo experiments illustrate the finite-sample performance of our proposed test.*

Factor Augmented Forecasting Subject to Structural Breaks in the Factor Structure

Ze-Yu Zhong, Monash University

This paper investigates the impact of structural breaks in the factor structure on factor-augmented forecasting. We decompose the break in the factor loading matrix into rotational and shift components. To effectively utilise the pre-break data and maintain robustness against shift breaks, we propose a novel factor estimator that minimises the L2 distance between pre- and post-break loading matrices through the rotation of factor estimates. We call this estimator the "rotated factors" and analyse its asymptotic properties, along with two competing factor estimators, in the presence of different types of breaks. To leverage the respective advantages of each factor estimator in an automatic data driven way, we introduce a method that averages over sets of factor estimates using a leave-h-out cross-validation criterion. Simulations demonstrate that combining different factor estimates through the proposed cross-validation averaging approach leads to improved forecasting performance compared to existing methods. Furthermore, we evaluate the effectiveness of our methods in an empirical application with U.S. macroeconomic data and emphasise the importance of incorporating structural breaks into factor-augmented forecasting models.

Session 23	**High-Frequency Econometrics** *Organizer/Chair: Yi Ding, University of Macau;* *Merrick Li, The Chinese University of Hong Kong*	E22-2007	11 Dec, 14:30 - 16:00

High Frequency Factor Analysis with Partially Observable Factors

Dachuan Chen, Singapore Management University

This paper considers a novel factor structure -- Partially Observable Factor Model-- where both observable factors and latent factors exist in the model simultaneously. Such factor structure can make sure both interpretability and goodness-of-fit at the same time. Necessary estimation methodologies for this partially observable factor model are developed in this paper for the high frequency data. The proposed estimation methodology is robust to jumps, microstructure noise and asynchronous observation times simultaneously. When the observable factors are exogenous, we provide the estimation theory for the integrated eigenvalues of the residual covariance matrix, which including the bias-corrected estimator, central limit theorem and asymptotic variance estimator. As a result, the asymptotic normality of the bias-corrected estimator can be applied to test the existence of the latent factors. When the observable factors are endogenous, we propose a novel framework of high frequency unsupervised instrumental learning (HF-UIL), which can help people quantify the contributions of the observable factors into the latent factors. This is the first work on high frequency instrumental variables, and it can be regard as an necessary extension of the Projected-PCA (or Instrumented PCA) in the world of continuous-time model. Statistical inferences have been established for the loadings of the observable factors onto the latent factors. Monte Carlo simulation demonstrates the validity of our estimation methodologies. Empirical study demonstrates that (i) in the exogenous setting, the latent factors significantly exist in the residual process of the high frequency regression; (ii) in the endogenous setting, the correlations between the observable factors and latent factors do exist significantly. This is the joint work with Wenqi Lu (NKU) and Siyu Xie (Northwestern).

Do Equity and Options Markets Agree about Volatility?

Casten Chong, The Hong Kong University of Science and Technology

We address this question by deriving tight pricing kernel restrictions from zero-date options, which are options that expire on the same day they are traded. These restrictions concern the volatility of small and frequent asset price moves that the equity and options markets must agree on in a frictionless economy where the two markets are integrated. We show that violations of such restrictions lead to local arbitrage opportunities that can be exploited using a static portfolio of zero-date options and a dynamic position in the underlying asset. These local arbitrage opportunities are characterized by arbitrarily high reward-to-risk ratios and cause local explosion of conditional moments of the aggregate pricing kernel. Empirically, we find no evidence of such local arbitrage opportunities. Thus, in spite of the nontrivial risk premium embedded in zero-date options, their prices correctly reflect the time-varying volatility of the underlying asset.

Measuring Intraday Liquidity

Merrick Li, The Chinese University of Hong Kong

Intraday prices exhibit a transitory pricing error component that signals market inefficiency in liquidity provision. We propose an "oracle" estimator for the spot scale of pricing errors, relying solely on transaction prices, thereby serving as a natural measure of intraday liquidity. Our estimator effectively disentangles shocks to fundamental values from shocks to liquidity demand. Empirically, we demonstrate that the conventional U-shaped pattern of intraday liquidity may vanish once these two types of shocks are separated. Additionally, we observe a surge in economic uncertainty, as indicated *by increased volatilities of fundamental values, yet only a mild deterioration in liquidity before the FOMC announcements.*

The Explicative Market Microstructure Noise

Wenhao Cui, Beihang University

High-frequency financial data are often contaminated by the market microstructure effect. In this study, we consider a scenario where a fraction of the market microstructure noise could be explained by observable trading information, referred to as the explicative noise component. To formally analyze this component, we first propose a variable importance measure that allows us to assess the price impact of a subset of trading information. After identifying the significant trading information, we then propose a nonparametric estimator of the explicative market microstructure noise and establish its corresponding asymptotic properties. We rely on Monte Carlo simulations calibrated with real data to examine the finite sample performance of the proposed method. Lastly, through an empirical application based on real data, we find that the explicative noise component plays a crucial role in explaining the return process. The signed bid-ask spread serves as the primary source of the explicative noise component, and its price impact is largely nonlinear.

Session 24	**Learning in Finance** *Organizer/Chair: Yi Ding, University of Macau*	E22- 2009	11 Dec, 14:30 - 16:00

Online Inference for Robust Policy Evaluation in Reinforcement Learning

Yichen Zhang, Purdue University

Recently, reinforcement learning has gained prominence in modern statistics, with policy evaluation being a key component. Unlike traditional machine learning literature on this topic, our work places emphasis on statistical inference for the parameter estimates computed using reinforcement learning algorithms. While most existing analyses assume random rewards to follow standard distributions, limiting their applicability, we embrace the concept of robust statistics in reinforcement learning by simultaneously addressing issues of outlier contamination and heavy-tailed rewards within a unified framework. In this paper, we develop an online robust policy evaluation procedure, and establish the limiting distribution of our estimator, based on its Bahadur representation. Furthermore, we develop a fully-online procedure to efficiently conduct statistical inference based on the asymptotic distribution. This paper bridges the gap between robust statistics and statistical inference in reinforcement learning, offering a more versatile and reliable approach to policy evaluation.

Trading Volume Alpha

Chao Zhang, The Hong Kong University of Science and Technology, Guangzhou

Portfolio optimization chiefly focuses on risk and return prediction, yet implementation costs also play a critical role. Predicting trading costs is challenging, however, since costs depend endogenously on trade size and trader identity, thus impeding a generic solution. We focus on a key, yet general, component of trading costs that abstracts from these challenges -- trading volume. Individual stock trading volume is highly predictable, especially with machine learning. We model the economic benefits of predicting stock volume through a portfolio framework that trades off portfolio tracking error versus net-of-cost performance -- translating volume prediction into net-of-cost portfolio alpha. We find the benefits of predicting individual stock volume to be substantial, and potentially as large as those from stock return prediction.

Liquidity jump networks

Yumin Xu, Peking University

We propose a network model to study liquidity dry-ups of a large panel of stocks. Our liquidity jump network captures liquidity co-jump dependence through a community structure. We develop consistent estimators of the model parameters and prove the strong consistency of spectral clustering for community detection in a setting where both serial and cross-sectional dependence are allowed. Using high-frequency limit order book data of Chinese stocks, bothserial and cross-sectional dependence are allowed. Using high-frequency limit order book data of Chinese stocks, we discover stock liquidity co-jump communities that cannot be solely explained by stock characteristics or industries. Community information is shown to be helpful in liquidity prediction and portfolio construction.

Machine Learning for Nonstationary Data

Jin Xi, Chinese Academy of Science

Machine learning offers a promising set of tools for forecasting. However, some of the well-known properties do not apply to nonstationary data. This paper uses a simple procedure to extend machine learning methods to nonstationary data that does not require the researcher to have prior knowledge of which variables are nonstationary or the nature of the nonstationarity. I illustrate theoretically that using this procedure with LASSO or adaptive LASSO generates consistent variable selection on a mix *of stationary and nonstationary explanatory variables. In an empirical exercise, I examine the success of this approach at forecasting U.S. inflation rates and the industrial production index using a number of different machine learning methods. I find that the proposed method either significantly improves prediction accuracy over traditional practices or delivers comparable performance, making it a reliable choice for obtaining stationary components of high-dimensional data.*

Session 28	**Recent Advances in High-dimensional and Nonparametric Statistical Inference** *Organizer/Chair:* *Xinghao Qiao, The University of Hong Kong;* *Long Feng, The University of Hong Kong*	E22-2010	11 Dec, 14:30 - 16:00

Robust bias-corrected empirical likelihood for nonparametric functions with an application to regression discontinuity designs

Qin Fang, University of Sydney

In this article, we investigate local empirical likelihood-based inference for nonparametric functions. We introduce novel robust bias-corrected and residual-adjusted empirical likelihood ratios and demonstrate that these ratios exhibit standard chi-squared limits without the need for undersmoothing. Notably, the ratio is self-scale invariant, eliminating the need for a plug-in estimate of the limiting variance. We then propose the new theory-based, robust confidence interval estimators for causal effects identified from sharp and fuzzy regression-discontinuity designs. Our simulation study shows that these confidence intervals offer nearly correct empirical coverage and favourable interval lengths on average, significantly outperforming existing alternatives in the literature.

Error estimation in high-dimensional linear regression with corrected errors

Shaojun Guo, Renmin University of China

Estimating variance and error correlation is a fundamental problem in statistical modeling. In high-dimensional linear regression, where the dimensionality is comparable to the sample size and the error term are correlated, traditional estimation techniques often suffer from significant bias and are therefore not applicable. To address this issue, we propose a new procedure based on spectral density that consistently estimate variance and error correlation. Our asymptotic results demonstrate that the resulting estimators are asymptotically both unbiased and normally distributed. The simulation studies lend further support to our theoretical claims.

Spectral ranking inferences based on general multiway comparisons

Weichen Wang, The University of Hong Kong

Spectral Ranking Inferences based on General Multiway Comparisons

Abstract: This paper studies the performance of the spectral method in the estimation and uncertainty quantification of the unobserved preference scores of compared entities in a general and more realistic setup. Specifically, the comparison graph consists of hyper-edges of possible heterogeneous sizes, and the number of comparisons can be as low as one for a given hyper-edge. Such a setting is pervasive in real applications, circumventing the need to specify the graph randomness and the restrictive homogeneous sampling assumption imposed in the commonly used Bradley-Terry-Luce (BTL) or Plackett-Luce (PL) models. Furthermore, in scenarios where the BTL or PL models are appropriate, we unravel the relationship between the spectral estimator and the Maximum Likelihood Estimator (MLE). We discover that a two-step spectral method, where we apply the optimal weighting estimated from the equal weighting vanilla spectral method, can achieve the same asymptotic efficiency as the MLE. Given the asymptotic distributions of the estimated preference scores, we also introduce a comprehensive framework to carry out both one-sample and two-sample ranking inferences, applicable to both fixed and random graph settings. It is noteworthy that this is the first time effective two-sample rank testing methods have been proposed. Finally, we substantiate our findings via comprehensive numerical simulations and subsequently apply our developed methodologies to perform statistical inferences for statistical journals and movie rankings.

Forecasting Global Economy via Matrix Autoregressive Model with Trade Network

Long Feng, The University of Hong Kong

Forecasting financial and economic variables across multiple countries has long been a significant challenge. Two primary approaches have been utilized to address this issue: the vector autoregressive model with exogenous variables (VARX) and the matrix autoregression (MAR). The VARX model captures domestic dependencies while using exogenous variables to represent international factors driven by global trade. In contrast, the MAR model simultaneously considers variables from multiple countries but ignores the trade network. In this paper, we propose an extension of the MAR model that incorporates the trade network, enabling the study of both international dependencies and the impact of trade on the global economy. Additionally, we introduce a sparse component to the model to differentiate between systemic and nonsystemic factors. To estimate the parameters, we propose both a likelihood estimation method and a bias-corrected version. We provide theoretical and empirical analyses of the model's properties, alongside presenting intriguing economic insights derived from our findings. Joint work with Sanyou Wu, Yan Xu, and Dan Yang.

Session 34	**Model-agnostic Statistical Inference** *Organizer/Chair: Changliang Zou, Nankai University; Guanghui Wang, East China Normal University*	E22-2011	11 Dec, 14:30 - 16:00

Model-agnostic statistical inference

Yajie Bao, Nankai University

We study the problem of post-selection predictive inference in an online fashion. To avoid devoting resources to unimportant units, a preliminary selection of the current individual before reporting its prediction interval is common and meaningful in online predictive tasks. Since the online selection causes a temporal multiplicity in the selected prediction intervals, it is important to control the real-time false coverage-statement rate (FCR) which measures the overall miscoverage level. We develop a general framework named CAP (Calibration after Adaptive Pick) that performs an adaptive pick rule on historical data to construct a calibration set if the current individual is selected and then outputs a conformal prediction interval for the unobserved label. We provide tractable procedures for constructing the calibration set for popular online selection rules. We proved that CAP can achieve an exact selection-conditional coverage guarantee in the finite-sample and distribution-free regimes. To account for the distribution shift in online data, we also embed CAP into some recent dynamic conformal prediction algorithms and show that the proposed method can deliver long-run FCR control. Numerical results on both synthetic and real data corroborate that CAP can effectively control FCR around the target level and yield more narrowed prediction intervals over existing baselines across various settings.

A Model-agnostic Approach for Variable Selection with FDR Control

Mengtao Wen, Nankai University

Selecting covariates that are related to the response is an important statistical problem. In this talk, we propose a novel selection procedure based on the model-agnostic variable importance measure, while guaranteeing the false discovery rate control. The proposed method integrates efficient computation of statistics, handling the strong dependence between statistics, and power enhancement. The false discovery rate can be controlled both theoretically and practically. Extensive numerical experiments and real-data analysis demonstrate the satisfactory performance of the proposed method.

Testing-oriented model selection in conformalized multiple testing

Haojie Ren, Shanghai Jiao Tong University

Conformalized multiple testing is a popular area in statistics and machine learning, focusing on controlling uncertainty and risks for decision-making problems in a model-free manner. The model selection problem in conformal prediction intervals has recently garnered increasing attention, but the problem in conformalized multiple testing has not been given much attention yet. The latter one aims to identify the model with the best testing performance with both null and alterntive data. In this work, we introduce a unified framework that encompasses existing model selection strategies by decoupling the dependence among non-conformity scores and constructing valid p-values using swapping techniques. Further, we present the Testing-Oriented Model Selection (TOMS) approaches within the proposed framework and demonstrate their validity and advantages in theory. Numerical studies confirm the efficiency and adaptability of TOMS across diverse scenarios.

Changepoint Detection in Complex Models: Cross-Fitting Is Needed

Guanghui Wang, East China Normal University

Changepoint detection is commonly approached by minimizing the sum of in-sample losses to quantify the model's overall fit across distinct data segments. However, we observe that flexible modeling *techniques, particularly those involving hyperparameter tuning or model selection, often lead to inaccurate changepoint estimation due to biases that distort the target of in-sample loss minimization. To mitigate this issue, we propose a novel cross-fitting methodology that incorporates out-of-sample loss evaluations using independent samples separate from those used for model fitting. This approach ensures consistent changepoint estimation, contingent solely upon the models' predictive accuracy across nearly homogeneous data segments. Extensive numerical experiments demonstrate that our proposed cross-fitting strategy significantly enhances the reliability and adaptability of changepoint detection in complex scenarios.*

Session 38	**Recent Advances in Statistical and Probabilistic Learning** *Organizer/Chair: Jialiang Li, National University of SingaporeJialiang LiNational University of Singapore*	E22-2013	11 Dec, 14:30 - 16:00

Multiple Imputation for Analysis of Interval-censored Data with Missing and Censored Covariates

Liming Xiang, Nanyang Technological University

We propose a novel multiple imputation approach under a class of semiparametric transformation models for interval censored data when covariates are missing or suffer censoring. Our proposal utilizes the information from the fully observed covariate values and the failure time outcomes to impute the missing and censored covariates iteratively with the use of rejection sampling, making the imputation model compatible to the substantive model and resulting in more efficient estimation than the existing methods. An extensive simulation study is conducted and demonstrates that the proposed approach works well in practice. Finally, we apply the proposed approach to a set of data on Alzheimer's disease.

Tree Shape Indices in Phylogeny: Recurrence, Asymptotics and Applications

Kwok Pui Choi, National University of Singapore

In evolutionary biology, hypotheses about micro-evolutionary and macro-evolutionary processes are commonly tested by comparing the tree shape indices of the empirical evolutionary tree with those predicted by neutral models. Key to this approach is to compute the joint distribution of these tree shape indices under random models of interest.

In this talk, we study the joint distribution of two such tree shape indices: number of cherries and number of pitchforks, for random phylogenetic trees generated by an α-random tree growth model, $\alpha \in[0,1]$. Based on a non-uniform version of an extended Pólya urn models in which negative entries are permitted for their replacement matrices, we obtain the strong law of large numbers and the central limit theorem for the joint distribution of these two indices. Furthermore, we derive a recursive formula for computing the exact joint distribution of these two indices, leading to exact formulas for their means and higher order asymptotic expansions of their second moments. Application to analyse the evolution of a variety of sugarcane will be described.

Multiply robust estimation for general multivalued treatment effects with missing outcomes

Xiaorui Wang, Southern University of Science and Technology

Interventions with multivalued treatments are common in medical and health research, leading to a growing interest in developing estimators for multivalued treatment effects using observational data. In practice, missing outcome data is a common occurrence, which poses significant challenges to the estimation of treatment effects. In this paper, we propose two multiply robust estimators for estimating the general multivalued treatment effects with outcome missing at random, including the average treatment effect (ATE), quantile treatment effect (QTE), and expectile treatment effect (ETE). The resulting estimators are root-n consistent and asymptotically normal, provided that the candidate models for the propensity score contain the correct model, and so do the candidate models for either the probability of being observed or outcome regression. Extensive simulation studies are conducted to investigate the finite-sample performance of the proposed estimators. The proposed methods are also applied to a real-world dataset of CHARLS with about 21% outcome missing, estimating the ATE, QTE and ETE of three types of social activities on the cognitive function of middle-aged and elderly people in China.

Geometry and factorization of multivariate Markov chains

Michael Choi, Yale-NUS College

We introduce a framework rooted in a rate distortion problem for Markov chains, and show how a suite of commonly used Markov Chain Monte Carlo (MCMC) algorithms are specific instances within it, where the target stationary distribution is controlled by the distortion function. Our approach offers a unified variational view on the optimality of algorithms such as Metropolis-Hastings, Glauber dynamics, the swapping algorithm and Feynman-Kac path models. Along the way, we analyze factorizability and geometry of multivariate Markov chains. Specifically, we demonstrate that induced chains on factors of a product space can be regarded as information projections with respect to a particular divergence. This perspective yields Han–Shearer type inequalities for Markov chains as well as applications in the context of large deviations and mixing time comparison. Finally, to demonstrate the significance of our framework, we propose a new projection sampler based on the swapping algorithm that provably accelerates the mixing time by multiplicative factors related to the number of temperatures and the dimension of the underlying state space.
This is based on joint work with Geoffrey Wolfer (Waseda University) and Youjia Wang (National University of Singapore).

Session 41	**High Dimensional Regression and Testing in Business** *Organizer/Chair: Lilun Du, City University of Hong Kong*	E22-2014	11 Dec, 14:30 - 16:00

Quantile regression and consistent variable selection across multiple quantiles, a case study of equity premium predictability

Shaobo Li, University of Kansas

This paper considers model selection and estimation for quantile regression when multiple quantiles are of interest. In practice, it is often a desired property that the selected variables are consistent across each quantile. To achieve this, we adopt the group lasso penalty where the multiple quantiles are grouped for each covariate. Approaches that do not guarantee this property often suffer from the crossing quantile issue, e.g., prediction for a higher quantile is smaller than that of a lower quantile. The proposed approach greatly mitigates this issue due to consistent model selection across all quantiles. Consistency results are provided that allow the number of predictors to increase with the sample size. A Huberized quantile loss function and an augmented data approach are implemented for computational efficiency. Extensive simulation studies are conducted to show the effectiveness of the proposed approach. As a case study, we re-examine the problem of equity premium predictability under the US stock market. We find that the proposed approach generally outperforms various benchmarks including Lasso, quantile-Lasso, and stepwise selection, for both in-sample model fitting and out-of-sample forecast. The proposed approach is particularly useful when a full spectrum distribution of the return prediction is of interest rather than the mean value.

Robust Quantile Factor Models

Junlong Feng, The Hong Kong University of Science and Technology

We propose a factor model and an estimator of the factors and loadings that are robust to weak factors. The factors can have arbitrarily weak influence on the mean or quantile of the outcome variable at most quantile levels; each factor only needs to have a strong impact on the outcome's quantile near one unknown quantile level. The estimator is asymptotically normal at the same rate of the principal component analysis under strong factors. It does not require the knowledge of whether or at which quantile level the factors are weak and how weak they are. We also develop a weak-factor-robust estimator of the number of factors and a consistent selector of quantile or mean factors of any desired strength of influence. Monte Carlo simulations demonstrate the effectiveness of our methods.

Employing Synthetic Statistics for Information Aggregation

Bowen Gang, Fudan University

Learning from the collective wisdom of crowds parallels the statistical concept of fusion learning from multiple data sources or studies. However, integrating inferences from diverse sources poses significant challenges due to cross-source heterogeneity and data-sharing limitations. Studies often rely on varied designs and modeling techniques, and stringent data privacy norms can prohibit even the sharing of summary statistics. In this talk, I will discuss the construction of "synthetic statistics" that mimic the summary statistics used for inference, enabling the fusion of inference results from multiple sources.

How does Multiple Testing help in Large-Scale Recommender Systems?

Lilun Du, City University of Hong Kong

Many important tasks of large-scale recommender systems can be naturally cast as testing multiple linear forms for noisy matrix completion. These problems, however, present unique challenges because of the subtle bias-and-variance tradeoff of and an intricate dependence among the estimated entries

induced by the low-rank structure. In this paper, we develop a general approach to overcome these difficulties by introducing new statistics for individual tests with sharp asymptotics both marginally and jointly, and utilizing them to control the false discovery rate (FDR) via a data splitting and symmetric aggregation scheme. We show that valid FDR control can be achieved with guaranteed power under nearly optimal sample size requirements using the proposed methodology. Extensive numerical simulations and real data examples are also presented to further illustrate its practical merits.

Session 13	**Recent Advances in Statistical Machine Learning** *Organizer/Chair: Yichao Wu, The University of Illinois at Chicago*	E22-2015	11 Dec, 14:30 - 16:00

Sparse Sufficient Dimension Reduction via Least Squares Support Vector Machine and its Extensions

Seung Jun Shin, Korea University

In this work, we explored sparse sufficient dimension reduction (SDR) through the framework of the penalized principal machine (PM). We proposed the penalized principal least square support vector machine (P^{2}LSM) as a primary example of this approach. The P^{2}LSM employs a squared loss function, facilitating efficient computation using the group coordinate descent algorithm. Our method enhances computational efficiency compared to conventional sparse SDR methods, particularly for large-scale datasets. We also extend our approach to include penalized principal asymmetric least squares regression (P^{2}AR), penalized principal L$_2$-SVM (P^{2}L2M), and penalized principal (weighted) logistic regression (P^2(W)LR), demonstrating its versatility. Additionally, a computational advantage, oracle property of the proposed methods are investigated. Extensive simulations and real data analyses illustrate the efficacy of the proposed methods in yielding sparse, interpretable solutions without compromising predictive accuracy.

Non-Euclidean object filtering for large-scale discriminant analysis

Xin Chen, Southern University of Science and Technology

The classification of random objects within metric spaces lacking a vector structure has attracted increasing attention. However, the inherent complexity of such non-Euclidean data often restricts existing models to handling only a few features, leaving a gap in real-world applications. To address this, we propose a data-adaptive filtering procedure to identify informative features from a large-scale set of random objects, leveraging a novel Kolmogorov-Smirnov type statistic defined on the metric space. Our method, applicable to data in general metric spaces with binary labels, exhibits remarkable flexibility. It is model-free, as its implementation does not rely on any specific classifier class. Theoretically, it guarantees strong selective consistency while controlling the false discovery rate. Empirically, equipped with a Wasserstein metric, it demonstrates superior performance compared to Euclidean competitors across various settings. We conduct a complete study on autism data, which identifies significant brain regions associated with the disease. Interesting findings reveal distinct interaction patterns among brain regions in individuals with and without autism, achieved by filtering hundreds of thousands of covariance matrices representing various brain l connectivities.

On Linear Convergence of ADMM for Decentralized Quantile Regression

Heng Lian, City University of Hong Kong

The alternating direction method of multipliers (ADMM) is a natural method of choice for distributed parameter learning. For smooth and strongly convex consensus optimization problems, it has been shown that ADMM and some of its variants enjoy linear convergence in the distributed setting, much like in the traditional non-distributed setting. The optimization problem associated with parameter estimation in quantile regression is neither smooth nor strongly convex (although is convex) and thus it can only have sublinear convergence at best. Although this insinuates slow convergence, we show that, if the local sample size is sufficiently large compared to parameter dimension and network size, distributed estimation in quantile regression actually exhibits linear convergence up to the statistical precision.

Regularized Estimation for Endogenous Threshold Regression via Cross-Validation

Wei Zhong, Xiamen University

Heterogeneity and endogeneity become increasingly common in econometric practice. Threshold regression provides a simple yet flexible modeling strategy to account for heterogeneity by allowing for increased threshold effects in functional form. This paper studies estimation and inference for multiple threshold regression with endogenous regressors. We exploit a novel estimation framework based on the control function to handle endogeneity, which first identifies a set of possible threshold estimates using the group LASSO estimation, and then refines the candidate set by a predetermined information criterion. However, the performance of information criterion is sensitive to the choice of penalization factor and the optimal penalization magnitude usually varies from different contexts. We further develop a sophisticated cross-validation criterion to determine the number of thresholds in a data-adaptive manner. In cases where regressors and threshold variable are both endogenous, the proposed approaches remain applicable with slight adjustments using the control function framework. Numerical examples demonstrate the favorable performance of our proposals, including an application to the threshold effect of 401(k) plans on the wealth.

Session 50	**Recent Progress on the Statistics for High Frequency Data** *Organizer/Chair: Zhi Liu, University of Macau*	E22-2017	11 Dec, 14:30 - 16:00

Cryptocrashes

Aleksey Kolokolov, University of Manchester

This paper proposes a new nonparametric test for detecting short-lived locally explosive trends (drift bursts) in pure-jump processes. The new test is designed specifically to detect intraday flash crashes and gradual jumps in cryptocurrency prices recorded at a high frequency. Empirical analysis shows that drift bursts in bitcoin price occur, on average, every second day. Their economic importance is highlighted by showing that hedge funds holding cryptocurrency in their portfolios are exposed to a risk factor associated with the intensity of bitcoin crashes. On average, hedge funds do not profit from intraday bitcoin crashes and do not hedge against the associated risk.

Mutually exciting point processes with latency

Yoann Potiron, Keio University

A novel statistical approach to estimating latency, defined as the time it takes to learn about an event and generate response to this event, is proposed. Our approach only requires a multidimensional point process describing event times, which circumvents the use of more detailed datasets which may not even be available. We consider the class of parametric Hawkes models capturing clustering effects and define latency as a known function of kernel parameters, typically the mode of kernel function. Since latency is not well-defined when the kernel is exponential, we consider maximum likelihood estimation in the mixture of generalized gamma kernels case and derive the feasible central limit theory with in-fill asymptotics. As a byproduct, central limit theory for a latency estimator and related tests are provided. Our numerical study corroborates the theory. An empirical application on high frequency data transactions from the New York Stock Exchange and Toronto Stock Exchange shows that latency estimates for the US and Canadian stock exchanges vary between 2 and 5 milliseconds from 2020 to 2021.

Functional Volatility Forecasting

Zhiyuan Zhang, Shanghai University of Finance and Economics

Widely used volatility forecasting methods are usually based on low frequency time series models.Although some of them employ high frequency observations, these intraday data are often summarized into low frequency point statistics, e.g., daily realized measures, before being incorporated into a forecasting model. This paper contributes to the volatility forecasting literature by instead predicting the next-period intraday volatility curve via a functional time series forecasting approach. Asymptotic theory related to the estimation of latent volatility curves via functional principal analysis is formally established, laying a solid theoretical foundation of the proposed forecasting method. In contrast with non-functional methods, the proposed functional approach fully exploits the rich intraday information and hence leads to more accurate volatility forecasts. This is confirmed by extensive comparisons between the proposed method and those widely used non-functional methods in both Monte Carlo simulations and an empirical study on a number of stocks and equity indices from the Chinese market.

When asynchronicity meets price staleness: Robust estimation of high-frequency covariance

Haibin Zhu, Jinan University

Existing literature has shown that both asynchronicity and price staleness yield a downward bias in covariance estimation, a puzzle known as the Epps effect. We propose a novel estimator of high-

frequency covariance, that is resilient to the concurrent presence of asynchronicity and price staleness. We establish the asymptotic properties of this new estimator and present a feasible central limit theorem that accounts for time-varying staleness probabilities. The proposed robust covariance estimator also yields consistent estimators for beta and correlation in the presence of these anomalies. Additionally, we propose a pre-averaging method to address microstructure noise. Through theoretical derivation and Monte Carlo simulations, we demonstrate that our new estimator outperforms existing related estimators.

Session 55	**High Dimensional Statistical Inference** *Organizer/Chair: Hongyuan Cao, Florida State University*	E22-2018	11 Dec, 14:30 - 16:00

Penalized Deep Partially Linear Cox Models

Yi Li, University of Michigan

Partially linear Cox models have gained popularity for analysis of censored data by dissecting the hazard function into parametric and nonparametric components, allowing for the effective incorporation of both well-established risk factors (such as age and clinical variables) and emerging risk factors (e.g., image features) within a unified framework. However, when the dimension of parametric components exceeds the sample size, the task of model fitting becomes formidable, while nonparametric modeling grapples with the curse of dimensionality. We propose a novel Penalized Deep Partially Linear Cox Model (Penalized DPLC), which incorporates the SCAD penalty to select important texture features and employs a deep neural network to estimate the nonparametric component of the model. We prove the convergence and asymptotic properties of the estimator and compare it to other methods through extensive simulation studies, evaluating its performance in risk prediction and feature selection.

High-dimensional mediation analysis for longitudinal mediators and survival outcomes

Lei Liu, Washington University in St Louis

Mediation analysis with high-dimensional mediators is important in identifying epigenetic pathways between environmental exposure to health outcomes. However, high-dimensional mediation analysis methods for longitudinal mediators and a survival outcome remain underdeveloped. This study addresses this gap by proposing a methodology that accommodates time-varying mediators, characterized through multivariate longitudinal observable variables, to explore mediation effects over time. Our approach uses a longitudinal mixed effects model to examine the relationship between exposure and the mediating process. We connect the mediating process to the survival outcome using a Cox proportional hazards model with time-varying mediators. To handle high-dimensional data, we first employ a mediation-based sure independence screening method for dimension reduction. A Lasso inference procedure is utilized to identify significant time-varying mediators. We adopt a joint significance test to accurately control the family wise error rate in testing high-dimensional mediation hypotheses. Simulation studies and an analysis of the Coronary Artery Risk Development in Young Adults (CARDIA) Study demonstrate the utility and validity of our method.

Demographic Parity-aware Individualized Treatment Rules

Wen Su, City University of Hong Kong

Combining dependent p-values presents a longstanding challenge in statistical inference, particularly when aggregating results from diverse methods to boost signal detection. P-value combination tests using heavy-tailed distribution-based transformations, such as the Cauchy combination test and the harmonic mean p-value, have recently garnered significant interest in genetic applications for their potential to efficiently handle arbitrary p-value dependencies. In the talk, I will present our new results providing a deeper understanding of the heavy-tailed combination tests. Specifically, though researchers have shown that these combination tests are asymptotically valid for pairwise quasi-asymptotically independent test statistics, such as bivariate normal variables, we find out that they are also asymptotically equivalent to the Bonferroni test under the same conditions, making them uninteresting. On the other hand, we show when quasi-asymptotic independence is violated, such as when the test statistics follow multivariate t distributions, these tests are still asymptotically valid, and can be asymptotically much more powerful than the Bonferroni test. Our new results provide a broader *justification of the heavy-tailed combination tests and indicate their practical utility when some p-values are highly dependent.*

A robust and powerful replicability analysis for high dimensional data

Hongyuan Cao, Florida State University

Identifying replicable signals across different studies provides stronger scientific evidence and more powerful inference. Existing literature on high dimensional replicability analysis either imposes strong modeling assumptions or has low power. We develop a powerful and robust empirical Bayes approach for high dimensional replicability analysis. Our method effectively borrows information from different features and studies while accounting for heterogeneity. We show that the proposed method has better power than competing methods while controlling the false discovery rate, both empirically and theoretically. Analyzing datasets from the genome-wide association studies reveals new biological insights that otherwise cannot be obtained by using existing methods.