SRCC Workshop on Advanced Statistical Methods and Their Applications

Date : 06 July 2023 (Thursday)

Time : 10:00–17:30

Venue : FSC1217, Fong Shu Chuen Library
Ho Sin Hang Campus
Hong Kong Baptist University

Lijian YANG (Center for Statistical Science, Tsinghua University), **10:00-10:30**

Title: Hypotheses Testing of Functional Principal Components

Abstract: We propose a test for the hypothesis that the standardized functional principal components (FPCs) of a functional data equal a given set of orthonormal basis (e.g., the Fourier basis). Using estimates of individual trajectories that satisfy certain approximation conditions, a chi-square type statistic is constructed and shown to be oracally efficient under the null hypothesis in the sense that its limiting distribution is the same as an infeasible statistic using all trajectories, known by "oracle". The null limiting distribution is an infinite Gaussian quadratic form, and a consistent estimator of its quantile is obtained. A test statistic based on the chi-square type statistic and approximate quantile of the Gaussian quadratic form is shown to be both of the nominal asymptotic significance level and asymptotically correct. It is further shown that B-spline trajectory estimates meet the required approximation conditions. Simulation studies illustrate superior finite sample performance of the proposed testing procedure. For the EEG (ElectroEncephalogram) data, the proposed procedure has confirmed an interesting discovery that the centered EEG data is generated from a small number of elements of the standard Fourier basis. This is a joint work with Zening Song, Nankai University and Yuanyuan Zhang, Soochow University.

Yingcun XIA (Department of Statistics and Applied Probability, National University of Singapore), **10:30-11:00**

Title: Ensemble Projection Pursuit for General Nonparametric Regression

Abstract: The projection pursuit regression (PPR) has played an important role in the development of statistics and machine learning. However, when compared to other established methods like random forests (RF) and support vector machines (SVM), PPR has yet to showcase a similar level of accuracy as a statistical learning technique. In this paper, we revisit the estimation of PPR and propose an \textit{optimal} greedy algorithm and an ensemble approach via "feature bagging", hereafter referred to as ePPR, aiming to improve the efficacy. Compared to RF, ePPR has two main advantages. Firstly, its theoretical consistency can be proved for more general regression functions as long as they are L^2 integrable, and higher consistency rates can be achieved. Secondly, ePPR does not split the samples, and thus each term of PPR is estimated using the whole data, making the minimization more efficient and guaranteeing the smoothness of the estimator. Extensive comparisons based on real data sets show that ePPR is more efficient in regression and classification than RF and other competitors. The efficacy of ePPR, as a variant of Artificial Neural Networks (ANN), demonstrates that with suitable

statistical tuning, ANN can equal or even exceed RF in dealing with small to medium-sized datasets. This finding challenges the widespread belief that ANN's superiority over RF is limited to processing big data.

Break Time: 11:00-11:20

Tracy KE (Department of Statistics, Harvard University), **11:20-11:50**

Title: Mixed Membership Estimation for Large Social Networks

Abstract: Given a large social network, mixed membership estimation (MME) aims to obtain a K-dimensional weight vector pi_i for each node i, where K is the number of perceivable communities in the network and pi_i(k) is the fractional weight that node i puts on community k. MME can be viewed as a "soft" community detection problem. We propose a spectral algorithm, Mixed-SCORE, and show that this algorithm is both computationally fast and minimax optimal. We apply Mixed-SCORE to a co-citation network of statisticians constructed from the MADStat data set (http://zke.fas.harvard.edu/MADStat.html). We find that the estimated communities represent the primary research areas in statistics and the estimated pi_i's describe the research interests/impacts of individual authors. Our results also produce a "statistics triangle", reminiscent of the "statistical philosophy" triangle in Efron (1998).

Xin TONG (Department of Data Sciences and Operations, University of Southern California), **11:50-12:20**

Title: Individual-centered Partial Information

Abstract: In statistical network analysis, we often assume either the full network is available or multiple subgraphs can be sampled to estimate various global properties of the network. However, in a real social network, people frequently make decisions based on their local view of the network alone. Here, we consider a partial information framework that characterizes the local network centered at a given individual by path length L and gives rise to a partial adjacency matrix.

Lunch Time: 12:20-14:00

Jiancheng JIANG (Department of Mathematics and Statistics, University of North Carolina at Charlotte), **14:00-14:30**

Title: Inference for Ultra High-dimensional Quasi-likelihood Models based on Data Splitting

Abstract: In this talk, we develop a valid framework for inference of ultra-high dimensional quasi-likelihood models, based on a novel weighted estimation approach. The weighted estimator is obtained by minimizing the variance function. We split the full data into two subsets and perform model selection on one subset while computing the maximum quasi-likelihood estimator on the other. The two estimators are then aggregated using optimal weighted matrices. Using the weighted estimator, we construct confidence regions for a group of components of the regression vector and perform the Wald test for a linear structure of the group components. Theoretically, we establish the asymptotic normality of the weighted estimator, and the asymptotic distributions of the Wald test statistic under the null and alternative, without assuming model selection consistency. We highlight the advantages of the proposed tests through theoretical and empirical comparisons with some competitive tests, which guarantees that

our proposed inference framework is locally optimal. Furthermore, we prove that when selection consistency is achieved, the proposed Wald test is asymptotically identical in distribution to the oracle test which knows the support of the regression vector. We also demonstrate the superior finite sample performance of our proposed tests through extensive simulations. Finally, we illustrate the application of our methodology to a breast cancer dataset.

Yichuan ZHAO (Department of Mathematics and Statistics, Georgia State University), **14:30-15:00**

Title: Novel Empirical Likelihood Inference for the Mean Difference with Right-Censored Data

Abstract: This paper focuses on comparing two means and finding a confidence interval for the difference of two means with right-censored data using the empirical likelihood method combined with the i.i.d. random functions representation. Some early researchers proposed empirical likelihood-based confidence intervals for the mean difference based on right-censored data using the synthetic data approach. However, their empirical log-likelihood ratio statistic has a scaled chi-squared distribution. To avoid the estimation of the scale parameter in constructing confidence intervals, we propose an empirical likelihood method based on the i.i.d. representation of Kaplan–Meier weights involved in the empirical likelihood ratio. We obtain the standard chi-squared distribution. We also apply the adjusted empirical likelihood to improve coverage accuracy for small samples. We investigate a new empirical likelihood method, the mean empirical likelihood, within the framework of our study. Via extensive simulations, the proposed empirical likelihood confidence interval has better coverage accuracy than those from existing methods. Finally, our findings are illustrated with a real data set.

Nan LIN (Department of Mathematics and Statistics, Washington University in St. Louis), **15:00-15:30**

Title: Distributed Quantile Regression for Longitudinal Dig Data

Abstract: Weighted quantile regression (WQR) is an effective tool for analyzing longitudinal data with heterogeneity, especially for its mild distributional requirement on the data. For small or moderate data, the WQR estimation problem is traditionally solved by linear programming algorithms, such as the interior point (IP) method. However, when applied to big data, especially high-dimensional big data, the IP method is often computationally too expensive or infeasible due to its full matrix factorization in every iteration. We propose a distributed algorithm, WQR-ADMM, for WQR in distributed longitudinal big data based on the multi-block alternating direction method of multipliers and establish its convergence property. Simulation studies demonstrate that WQR-ADMM is faster than IP in big data, particularly for the cases where the dimension p is large, and has favorable estimation accuracy in both non-distributed and distributed environments. We further illustrate the practical performance of WQR-ADMM by analyzing a Beijing air quality data set.

Break Time: 15:30-15:50

Lucy XIA (Department of Information Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology), **15:50-16:20**

Title: Fairness-adjusted Neyman-Pearson Classifiers

Abstract: Automated algorithmic decision-making is an essential process for many organizations, and developing efficient statistical methods for this purpose is a top priority. However, achieving organizational efficiency is a complex task since multiple aspects need to be optimized simultaneously. These aspects include algorithmic fairness, i.e., minimizing systemic bias against certain disadvantaged

social groups, and economic efficiency, i.e., minimizing the cost induced by incorrect decisions. To address these targets, we utilize a dual-focused Neyman-Pearson (NP) classification paradigm that seeks minimal type II error under simultaneous control over both the type I error and fairness bias. Leveraging an LDA model, we develop a new oracle framework for dual-focused NP classification, which is a first of its kind. Our proposed finite-sample-based classifiers satisfy both the fairness constraint and type I error constraint with high probability at the population level. We also derive oracle bounds on the excess type II error. Notably, our new classifier does not require sample splitting, which was necessary for most existing NP methods, leading to further increased data efficiency. Numerical and real data analyses demonstrate its superior performance.

Weichen WANG (HKU Business School, University of Hong Kong), **16:20-16:50**

Title: Ranking Inferences Based on the Top Choice of Multiway Comparisons

Abstract: This paper considers ranking inference of n items based on the observed data on the top choice among M randomly selected items at each trial. This is a useful modification of the Plackett-Luce model for M-way ranking with only the top choice observed and is an extension of the celebrated Bradley-Terry-Luce model that corresponds to M=2. Under a uniform sampling scheme in which any M distinguished items are selected for comparisons with probability p and the selected M items are compared L times with multinomial outcomes, we establish the statistical rates of convergence for underlying n preference scores using both ℓ_2-norm and ℓ_∞-norm, with the minimum sampling complexity. In addition, we establish the asymptotic normality of the maximum likelihood estimator that allows us to construct confidence intervals for the underlying scores. Furthermore, we propose a novel inference framework for ranking items through a sophisticated maximum pairwise difference statistic whose distribution is estimated via a valid Gaussian multiplier bootstrap. The estimated distribution is then used to construct simultaneous confidence intervals for the differences in the preference scores and the ranks of individual items. They also enable us to address various inference questions on the ranks of these items. Extensive simulation studies lend further support to our theoretical results. A real data application illustrates the usefulness of the proposed methods convincingly.

Jie LI (School of Statistics, Renmin University of China), **16:50-17:20**

Title: Time-varying Treatment Effects of Functional Data with Latent Confounders: Application to Sleep Heart Health Studies

Abstract: Exploring the causal effect between variables is an important issue in lots of scientific research. Existing literature on causal inference mainly studies one-dimensional or multi-dimensional data, but functional data with repeated observations per individual frequently appears in a wide variety of applications. In functional data, treatment design may change across time, and its treatment effect is a time-varying function. Besides, most methods for treatment effect estimation based on observational data rely on the ignorability assumption that treatment assignment is independent of the potential outcomes given the observable covariates. This assumption can be violated when unobserved latent covariates are involved. We propose a novel method for unbiased treatment effect estimation with unobserved latent covariates for functional data. We propose to solve this challenging problem using a joint likelihood method with a Monte Carlo EM algorithm. Moreover, our proposed method is flexible to estimate both heterogeneous treatment effect of individuals and average treatment effect, providing a reliable inferential tool in making treatment decisions. It can also be applied to the irregular and sparse data. The method leads to meaningful discoveries when applied to investigate the dynamic effect of sleep quality on heart rate variability.

-- All interested are most welcome! --