Workshop Program

Morning sessions:

Venue: CPD-3.28, Central Podium Levels 3 (CPD-3, Jockey Club Tower), Centennial Campus, HKU.

Time	People	Titles /Activities
08:30–09:00	Registration	
09:00–09:10	Opening: Haipeng Shen	
Session 1	**Chair: Weichen Wang**	
09:10–09:50	Iain Johnstone	Expectation Propagation and Maximum Likelihood in Generalized Linear Mixed Models
09:50–10:00	Photo	
10:00–10:30	Coffee break	
Session 2	**Chair: Dan Yang**	
10:30–11:10	Min-ge Xie	Repro Samples Method and Principled Random Forests
11:10–11:50	Qiman Shao	High Dimensional Gaussian Approximation
11:50-14:30	Lunch Break (Please transit to HKU iCube for afternoon sessions)	

Afternoon Sessions:

Venue: HKU-iCube, Room 4005-07, Two Exchange Square, 8 Connaught Place, Central, Hong Kong.

Time	People	Titles /Activities
Session 3	**Chair: Xinghao Qiao**	
14:30–15:10	Qiwei Yao	Autoregressive Networks with Dependent Edges
15:10–15:50	Yongtao Guan	Bias-Correction and Test for Mark-Point Dependence with Replicated Marked Point Processes
15:50–16:20	Coffee break	
Session 4	**Chair: Zhanrui Cai**	
16:20–17:00	Runze Li	Model-Free Statistical Inference on High-Dimensional Data
17:00–17:40	Yingying Li	Learning the Stochastic Discount Factor
17:40–17:50	Closing: Dan Yang	
17:50–20:30	Banquet (Invitation only)	

For detailed direction to the venues, please visit https://hkubs-stat.github.io/HKU-2024-Summer-Workshop.

Abstract

1. **Expectation Propagation and Maximum Likelihood in Generalized Linear Mixed Models**

   Speaker: Iain Johnstone, Stanford University

   Abstract: We consider a class of generalized linear mixed models in which both the number of groups and the number of observations within each group are large, and in which usual likelihood analysis encounters both computational and technical challenges. Matt Wand and colleagues have adapted the machine learning technique of expectation propagation (EP) to yield state-of-the-art estimation of parameters in such models. Here we ask: are the EP estimators asymptotically efficient? A main challenge is to define an appropriate objective function that captures the EP iteration and approximates maximum likelihood well enough to inherit its efficiency. A second issue is to show that maximum likelihood actually is efficient, due to integrals over random effects in the likelihood. For this we propose a novel method based on classical complex analysis. This is joint work with a group including the late Peter Hall, Matt Wand, Song Mei and Apratim Dey.

2. **Repro Samples Method and Principled Random Forests**

   Speaker: Min-ge Xie, Rutgers University

   Abstract: Repro Samples method introduces a fundamentally new inferential framework that can be used to effectively address frequently encountered, yet highly non-trivial and complex inference problems involving discrete or non-numerical unknown parameters and/or non-numerical data. In this talk, we present a set of key developments in the repro samples method and use them to develop a novel machine learning ensemble tree model, termed principled random forests. Specifically, repro samples are artificial samples that are reproduced by mimicking the genesis of observed data. Using the repro samples and inversion techniques stemmed from fiducial inference, we can establish a confidence set for the underlying ('true') tree model that generated, or approximately generated, the observed data.

   We then obtain a tree ensemble model using the confidence set, from which we derive our inference. Our development is principled and interpretable since, firstly, it is fully theoretically supported and provides frequentist performance guarantees on both inference and predictions; and secondly, the approach only assembles a small set of trees in the confidence set and thereby the model used is interpretable. The development is further extended to handle a causal inference setting of heterogeneous treatment effects. Numerical results have demonstrated superior performance of our proposed approach than several existing poste selection, random forest, bagging, and causal trees ensemble methods.

   The repro samples method provides a new toolset for developing interpretable AI and for helping address the blackbox issues in complex machine learning models. The development of the principle random forest is our first attempt on this direction.

3. **High Dimensional Gaussian Approximation**

Speaker: Qiman Shao, Southern University of Science and Technology

Abstract: Berry-Esseen type bounds for Gaussian approximation of standardized sums have been extensively studied under finite moment conditions for lower dimensional data and under sub-exponential moment conditions for high dimensional data. However, since the standardized coefficients such as the population standard deviations are typically unknown, it is essential for statistical inference to study the high dimensional Gaussian approximation of self-normalized sums. In this talk, we shall give a brief review on self-normalized limit theory and establish a Cramer type moderate deviation theorem for self-normalized Gaussian approximation under finite moment conditions.

4. **Autoregressive Networks with Dependent Edges**

Speaker: Qiwei Yao, The London School of Economics and Political Science

Abstract: We propose an autoregressive framework for modelling dynamic networks with dependent edges. It encompasses the models which accommodate, for example, transitivity, density-dependent and other stylized features often observed in real network data. By assuming the edges of network at each time are independent conditionally on their lagged values, the models, which exhibit a close connection with temporal ERGMs, facilitate both simulation and the maximum likelihood estimation in the straightforward manner. Due to the possible large number of parameters in the models, the initial MLEs may suffer from slow convergence rates. An improved estimator for each component parameter is proposed based on an iteration based on the projection which mitigates the impact of the other parameters. Based on a martingale difference structure, the asymptotic distribution of the improved estimator is derived without the stationarity assumption. The limiting distribution is not normal in general, and it reduces to normal when the underlying process satisfies some mixing conditions. Illustration with a transitivity model was carried out in both simulation and two real network data sets.

5. **Bias-Correction and Test for Mark-Point Dependence with Replicated Marked Point Processes**

Speaker: Yongtao Guan, The Chinese University of Hong Kong, Shenzhen

Abstract: Mark-point dependence plays a critical role in research problems that can be fitted into the general framework of marked point processes. In this work, we focus on adjusting for mark-point dependence when estimating the mean and covariance functions of the mark process, given independent replicates of the marked point process. We assume that the mark process is a Gaussian process and the point process is a log-Gaussian Cox process, where the mark-point dependence is generated through the dependence between two latent Gaussian processes. Under this framework, naive local linear estimators ignoring the mark-point dependence can be severely biased. We show that this bias can be corrected using a local linear estimator of the cross-covariance function and establish uniform convergence rates of the bias-corrected estimators. Furthermore, we propose a test statistic based on local linear estimators for mark-point independence, which is shown to converge to an asymptotic normal distribution in a parametric root n convergence rate. Model diagnostics tools are developed for key model assumptions and a robust functional permutation test is proposed for a more general class of mark-point processes. The

effectiveness of the proposed methods is demonstrated using extensive simulations and applications to some real data examples.

6. **Model-Free Statistical Inference on High-Dimensional Data**

Speaker: Runze Li, Pennsylvania State University

Abstract: This paper aims to develop an effective model-free inference procedure for high-dimensional data. We first reformulate the hypothesis testing problem via sufficient dimension reduction framework. With the aid of new reformulation, we propose a new test statistic and show that its asymptotic distribution is χ2 distribution whose degree of freedom does not depend on the unknown population distribution. We further conduct power analysis under local alternative hypotheses. In addition, we study how to control the false discovery rate of the proposed χ2 tests, which are correlated, to identify important predictors under a model-free framework. To this end, we propose a multiple testing procedure and establish its theoretical guarantees. Monte Carlo simulation studies are conducted to assess the performance of the proposed tests and an empirical analysis of a real-world data set is used to illustrate the proposed methodology.

7. **Learning the Stochastic Discount Factor**

Speaker: Yingying Li, Hong Kong University of Science and Technology

Abstract: We develop a statistical framework to learn the high-dimensional stochastic discount factor (SDF) from a large set of characteristic-based portfolios. Specifically, we build on the maximum-Sharpe ratio estimated and sparse regression method proposed in Ao, Li and Zheng (RFS,2019) to construct the SDF portfolio, and develop a statistical inference theory to test the SDF loadings. Applying our approach to 194 characteristic-based portfolios, we find that the SDF constructed by about 20 of them performs well in achieving a high Sharpe ratio and explaining the cross-section of expected returns of various portfolios. Joint work with Zhanhui Chen, Yi Ding and Xinghua Zheng.