Network Analysis and Deep Learning Workshop 2019

Alibaba Auditorium, Guanghua Building No.2, Peking University

Schedule

Date	Time	Speaker	Title	Chair
Wednesday Dec 18	08:20-08:30	Opening		
	08:30-09:10	Ji Zhu, University of Michigan	Network I	Jiashun Jin, Carnegie Mellon University
	09:10-09:50	Tracy Ke, Harvard University		
	09:50-10:10	Break		
	10:10-10:50	Jun Liu, Havard University	Statistical learning and deep learning I	Tracy Ke, Harvard University
	10:50-11:30	Dacheng Xiu, University of Chicago		
	11:30-11:40	Break		
	11:40-12:20	Feifang Hu, George Washington University	Personalized medicine	Ruibin Xi, Peking University
	14:00-14:40	Ming Yuan, Columbia University	Statistical learning and deep learning II	Fang Yao, Peking University
	14:40-15:20	Weichen Wang, Princeton University		
	15:20-15:40	Break		
	15:40-16:20	Pengsheng Ji, University of Georgia	Network II	Wei Lin, Peking University
	16:20-17:00	Xiaodong Li, University of California at Davis		
	17:00-17:10	Break		
	17:10-17:50	Shurong Zheng, Northeast Normal University	Random Matrix Theory I	Yumou Qiu, Iowa State University
Thursday Dec 19	08:30-09:10	Cun-Hui Zhang, Rutgers University	High dimensional data	Songxi Chen, Peking University
	09:10-09:50	Cheng Zhang, Peking University		
	09:50-10:10	Break		
	10:10-10:50	Peter Song, University of Michigan	Network III	Cheng Yong Tang,Temple University
	10:50-11:30	Rui Song, NC State University		
	11:30-11:40	Break		
	11:40-12:20	Zhigang Bao, Hong Kong University of Science and Technology	Random Matrix Theory II	Wang Miao, Peking University
	14:00-14:40	Yumou Qiu, Iowa State University	Statistical learning and deep learning III	Jinzhu Jia, Peking University
	14:40-15:20	Hongtu Zhu, Didi Chuxing Technology Co.		
	15:20-15:30	Summary		

Network I

Ji Zhu（University of Michigan）

Title: Generative Link Prediction for Incomplete Networks with Node Features

Abstract: Link prediction is one of the fundamental problems in network analysis. Most existing methods require at least partial observation of connections for every node. In real-world networks, however, there often exist nodes that do not have any link information, and it is imperative to make link predictions for such nodes based on their node features. In this talk, we consider a general framework in which a network consists of three types of nodes: nodes having features only, nodes having link information only, and nodes having both. Our goal is to predict links between nodes having features only and all other nodes. Under this setting, we have proposed a family of generative models for incomplete networks and node features, and we have developed a variational auto-encoder algorithm for model estimation and link prediction and investigated different encoder structures. We have also designed a cross-validation scheme under the problem setting. The proposed method has been evaluated on an online social network and two citation networks and achieved superior performance comparing with existing methods. This talk is based on joint work with Boang Liu, Binghui Liu and Elizaveta Levina.

Tracy Ke (Harvard University)

Title: Quantitative Analysis of Statisticians' Publications

Abstract: We have collected and cleaned a data set for the publications of statisticians, consisting of titles, authors, abstracts, MSC numbers, keywords, and citation counts of 83,661 papers published in 36 journals in statistics, probability, and related fields, spanning 41 years.

The data set motivates an array of interesting problems.

In this talk, I will discuss two topics. (a) The community structure of authors. We construct a citee network from the data and conduct mixed membership estimation. The results reveal a "Statistics Triangle", where the three vertices represent the three primary areas in statistic--Bayes, Biostatistics, and Nonpametrics. The results also allow for characterization of research trajectories of individual authors over the years. We also construct a co-authorship network from the data and conduct community detection. It gives rise to a hierarchical tree of communities with 25 leaf clusters, which reveal the collaboration patterns among authors. (b) The topic structure in papers. We use the paper abstracts in our data set as the text documents and conduct topic modeling. It results in 11 "topics", where each "topic" corresponds to a research area in statistics, such as "Hypothesis Testing", "Experimental Design", "Machine Learning", etc. We use the estimated topic weights to study the hot topics in statistical research, the topic interests of individual authors, and citation exchanges between topics.

The above quantitative analysis uses an array of statistical methods we have developed for network data analysis and text mining, which I will also briefly overview.
(The work is collaborated with Pengsheng Ji, Jiashun Jin, and Wanshan Li)

Statistical learning and deep learning I

Jun Liu (Harvard University, COPSS Presidents' Award)

Title: Knockoffs or perturbations, that is a question

Abstract: Simultaneously finding multiple influential variables and controlling the false discovery rate (FDR) for linear regression models is a fundamental problem with a long history. Researchers recently have proposed and examined a few innovative approaches surrounding the idea of creating "knockoff" variables (like spike-ins in biological experiments) to control FDR. As opposed to creating knockoffs, a classical statistical idea is to introduce perturbations and examine the impacts. We introduce here a perturbation-based Gaussian Mirror (GM) method, which creates for each predictor variable a pair of perturbed "mirror variables" by adding and subtracting a randomly generated Gaussian random variable, and proceeds with a certain regression method, such as the ordinary least-square or the Lasso. The mirror variables naturally lead to a test statistic highly effective for controlling the FDR. The proposed GM method does not require strong conditions for the covariates, nor any knowledge of the noise level and relative magnitudes of dimension p and sample size n. We observe that the GM method is more powerful than many existing methods in selecting important variables, subject to the control of FDR especially under the case when high correlations among the covariates exist. Additionally, we further extend the method for determining important variables in general neural network models and potentially other complex models. If time permits, I will also discuss a simpler bootstrap-type perturbation method for estimating FDRs, which is also more powerful than knockoff methods when the predictors are reasonably correlated. The presentation is based on joint work with Xing Xin, Zhigen Zhao, Chenguang Dai, Buyu Lin, Gui Yu.

Dacheng Xiu (University of Chicago)

Title: Thousands of Alpha Tests

Abstract: Data snooping is a major concern in empirical asset pricing. By exploiting the "blessings of dimensionality" we develop a new framework to rigorously perform multiple hypothesis testing in linear asset pricing models, while limiting the occurrence of false positive results typically associated with data-snooping. We first develop alpha test statistics that are asymptotically valid, allow for weak dependence in the cross-section, and are robust to the possibility of omitted factors. We then combine them in a multiple-testing procedure that ensures that the rate of false discoveries is ex-ante bounded below a prespecified 5% level. We also show that this method can detect all positive alphas with reasonable strength. Our procedure is designed for high-dimensional settings and works even when the number of tests is large relative to the sample size, as in many finance applications. We illustrate the empirical relevance of our methodology in the context of hedge fund performance (alpha) evaluation. We find that our procedure is able to select – among more than 3,000 available funds – a subset of funds that displays superior in-sample and out-of-sample performance compared to the funds selected by standard methods.

Personalized medicine

Feifang Hu (George Washington University)

Title: Statistical Inference of Adaptive Randomized Clinical Trials for Precision Medicine

Abstract: Adaptive randomization is frequently used in clinical trials (especially for precision medicine). However, since the randomization inevitably uses the covariate information when forming balanced treatment assignments, the validity of classical statistical inference following such randomization is often unclear. In this talk, we derive the theoretical properties of statistical inference post general covariate-adjusted randomization under the linear model framework. More important, we explicitly unveil the relationship between covariate-adjusted and inference properties. We apply the proposed general theory to various randomization procedures including complete randomization (CR), re-randomization (RR), pairwise sequential randomization (PSR), and Atkinson's DA- optimality biased coin design (DA-BCD) and compare their performance analytically. We then proposed a new adjusted approach to obtain valid and more powerful tests. These results open a new door to understand and analyze comparative studies based on covariate-adjusted randomization. Simulation studies provide further evidence of the advantages of the proposed framework and theoretical results. This talk is based on joint research with Wei Ma, Yichen Qin, Yang Li and some others.

Statistical learning and deep learning II

Ming Yuan (Columbia University, Editor of Annals of Statistics)

Weichen Wang (Princeton University)

Title: Robust Large Covariance Estimation for Heavy-tailed Factor Models

Abstract: We propose a general Principal Orthogonal complement Thresholding (POET) framework for large-scale covariance matrix estimation based on the approximate factor model. A set of high-level sufficient conditions for the procedure to achieve optimal rates of convergence under different matrix norms is established to better understand how POET works. Such a framework allows us to recover existing results for sub-Gaussian data in a more transparent way that only depends on the concentration properties of the sample covariance matrix. As a new theoretical contribution, for the first time, such a framework allows us to exploit conditional sparsity covariance structure for the heavy-tailed data. In particular, for the elliptical distribution, we propose a robust estimator based on the marginal and spatial Kendall’s tau to satisfy these conditions. In addition, we study conditional graphical model under the same framework. The technical tools developed in this paper are of general interest to high-dimensional principal component analysis. Thorough numerical results are also provided to back up the developed theory.

Network II

Pengsheng Ji (University of Georgia)

Title: Statistics for Statisticians: Looking into the Past through Citations

Abstract: We have a new dataset covering about 80K statistical papers published in the last 40 years and use it to study a few aspects of the field of statistics through citations. First, we present the dynamic ranking of statistics journals using the Stigler model and PageRank. Second, we predict the highly cited papers using logistic model and G. boost and identify the most important features of these papers.

Xiaodong Li (University of California at Davis)

Title: Spectral methods in networks: hierarchical structures and risk estimation

Abstract: Hierarchical structures are common in real world data sets, and diverse clustering methods have been proposed to explore such structures. In the first part of this talk, we focus on the top-down hierarchical clustering based on the Fiedler vectors of graph-Laplacians. In particular, we show that Fiedler vector based hierarchical clustering is consistent under general tree structures and broad ranges of connectivity probabilities. Our analysis relies on careful exploiting the algebraic properties of graph Laplacian, as well as recently-developed probabilisitic tools in controlling the L-infinity norm perturbation of eigenvectors of random matrices.

The second part of this talk is motivated by a basic question in network analysis: How to determine the number of communities in a network dataset. The spectrum of the adjacency matrix is widely used in practice, but the cut-off is usually difficult to determine. We formulate this problem as evaluation of graphon estimation via spectral hard thresholding. Based on Efron's approximate GSURE framework for independent binary data, we derive GSURE formulas for spectral graphon estimation. The key of this derivation is to calculate the divergence of spectral functions. The performance of the proposed methods is illustrated by experiments on real world data.

Random Matrix Theory I

Shurong Zheng (Northeast Normal University)

Title: Estimating Number of Factors by Adjusted Eigenvalues Thresholding

Abstract: Determining the number of common factors is an important and practical topic in high dimensional factor models. The existing literatures are mainly based on the eigenvalues of the covariance matrix. Due to the incomparability of the eigenvalues of the covariance matrix caused by heterogeneous scales of observed variables, it is very difficult to give an accurate relationship between these eigenvalues and the number of common factors.

To overcome this limitation, we appeal to the correlation matrix and show surprisingly that the number of eigenvalues greater than $1$ of population correlation matrix is the same as the number of common factors under some mild conditions. To utilize such a relationship, we study the random matrix theory based on the sample correlation matrix in order to correct the biases in estimating the top eigenvalues and to take into account of estimation errors in eigenvalue estimation. This leads us to propose adjusted correlation thresholding (ACT) for determining the number of common factors in high dimensional factor models, taking into account the sampling variabilities and biases of top sample eigenvalues. We also establish the optimality of the proposed methods in terms of minimal signal strength and optimal threshold. Simulation studies lend further support to our proposed method and show t

High dimensional data

Cun-Hui Zhang (Rutgers University, Editor of Statistical Science)

Title: Second order Stein: SURE for SURE and other applications

Abstract: Stein's formula states that a random variable of the form z>f(z)□div f(z) is mean-zero for all functions f with integrable gradient. Here, div f is the divergence of the function f and z is a standard normal vector. We develop Second Order Stein formulas for statistical inference with high-dimensional data. In the simplest form, the Second Order Stein formula characterizes the variance of z>f(z)□div f(z). A first application of the Second Order Stein formula is an Unbiased Risk Estimate for Stein's Unbiased Risk Estimator (SURE for SURE): an unbiased estimate provides information about the squared distance between SURE and the prediction error in the Gaussian sequence model. SURE for SURE has a simple form and can be computed explicitly for almost differentiable estimators, for example the Lasso and the Elastic Net. Other applications of the Second Order Stein formula are provided in high-dimensional regression. This includes novel bounds on the variance of the size of the model selected by the Lasso, and a general semi-parametric scheme to de-bias an almost differentiable initial estimator in order to estimate a low-dimensional projection of the unknown regression coefficient vector. This is joint work with Pierre Bellec.

Cheng Zhang (Peking University)

Title: Variational Bayesian Phylogenetic Inference

Abstract: Bayesian phylogenetic inference is currently done via Markov chain Monte Carlo with simple mechanisms for proposing new states, which hinders exploration efficiency and often requires long runs to deliver accurate posterior estimates. In this paper we present an alternative approach: a variational framework for Bayesian phylogenetic analysis. We approximate the true posterior using an expressive graphical model for tree distributions, called a subsplit Bayesian network, together with appropriate branch length distributions. We train the variational approximation via stochastic gradient ascent and adopt multi-sample based gradient estimators for different latent variables separately to handle the composite latent space of phylogenetic models. We show that our structured variational approximations are flexible enough to provide comparable posterior estimation to MCMC, while requiring less computation due to a more efficient tree exploration mechanism enabled by variational inference. Moreover, the variational approximations can be readily used for further statistical analysis such as marginal likelihood estimation for model comparison via importance sampling. Experiments on both synthetic data and real data Bayesian phylogenetic inference problems demonstrate the effectiveness and efficiency of our methods.

Network III

Peter Song (University of Michigan)

Title: Regression analysis of networked data

Abstract: We develop a new regression analysis approach to evaluating associations of covariates with outcomes measured from networks. This development is motivated from a study of infant growth that collects outcomes of event related potentials (ERP, a type of neuroimaging) measured over electroencephalogram (EEG) electrodes on the scalp. We propose a new generalized method of moments (GMM) that incorporates both established and data-driven knowledge of network topology among nodes in the estimation and inference to achieve robustness and efficiency. The GMM approach is computationally fast and stable to handle the regression analysis of network data, and conceptually it is simple with desirable properties in both estimation and inference. Both simulation studies and real EEG data analysis will be presented for illustration. This is a joint work with Yan Zhou.

Rui Song (NC State University)

Title: GNN-GAN: Graph Neural Network with Generative Adversarial Networks

Abstract: In this talk we consider a novel framework for semi-supervised classification problem on graph neural network (GNN) with generative adversarial networks (GAN). We first investigate the state-of-the-art GNN model (classifier) such as graph convolutional network (GCN) and graph attention network (GAT) on the semi-supervised classification problem. We then construct a generator which plays a non-cooperative game with the classifier to improve the performance of the corresponding GNN model. Theoretically we can demonstrate how the classifier benefits from joint training with a generator. With a number of experiments on benchmark data, we show that our approach outperforms the original GNN model by a significant margin.

Random Matrix Theory II

Zhigang Bao (Hong Kong University of Science and Technology)

Title: On Cramer-von Mises statistic for the spectral distribution of random matrices

Abstract: Let F_n and F be the empirical and limiting spectral distributions of an n by n Wigner matrix. The Cramer-von Mises (CvM) statistic is a classical goodness-of-fit statistic that characterizes the distance between F_n and F in l^2-norm. In this talk, we will consider a mesoscopic approximation of the CvM statistic for Wigner matrices, and derive its limiting distribution. The distribution fits well the heuristic prediction given by the analogue of the log-correlated Gaussian field.
This is a joint work with Yukun He.

Statistical learning and deep learning III

Yumou Qiu (Iowa State University**)**

Title: Statistical learning for high-throughput plant phenotyping

Abstract: Studying the association between phenotypic traits and genetic variation is one of the central goals in biological science. Despite the exceptional progress in building high-throughput image-based plant phenotyping systems, extracting accurate measurements of plant traits from the raw data and the following statistical analysis is currently a bottleneck. In this talk, we will introduce the problems and challenges in analyzing plant phenotyping data. In particular, we will discuss: 1. supervised learning methods, and especially neural networks, for image segmentation and feature extraction; 2. statistical method to analyze organ-aware plant growth dynamics; 3. Inference for high-dimensional partial correlation matrices for studying the conditional dependence structure among the extracted plant features and gene expression levels.

Hongtu Zhu (Didi Chuxing Technology Co.)

Title: Challenges in Analyzing Two-sided Market and Its Application on Ridesourcing Platform

Abstract: In this talk, we will introduce a general analytical framework for large scale data obtained from two-sided markets, especially ride-sourcing platforms like DiDi. This framework integrates classical methods including Experiment Design, Causal Inference and Reinforcement Learning, with modern machine learning methods, such as Graph Convolutional Models, Deep Learning, Transfer Learning and Generative Adversarial Network. We aim to develop fast and efficient approaches to address five major challenges for ride-sharing platform, ranging from demand-supply forecasting, demand-supply diagnosis, MDP-based policy optimization, A-B testing, to business operation simulation. Each challenge requires substantial methodological developments and inspires many researchers from both industry and academia to participate in this endeavor. Based on our preliminary results for the policy optimization challenge, we receive the Daniel Wagner Prize for Excellent in Operations Research Practice in 2019. All the research accomplishments presented in this talk are joint work by a group of researchers at Didi Chuxing and our international collaborators.