HKU 2023 Summer Workshop on Statistics and Data Analytics

PROGRAMME

Time	Title	Speaker
08:30 - 09:00	Registration	
09:00 - 09:10	Opening	
09:10 - 10:00	Statistical Inference on a Four-Regime Segmented Regression Model	Songxi Chen
10:05 - 10:55	A Model-Agnostic Graph Neural Network for Integrating Local and Global Information	Annie Qu
10:55 - 11:20	Coffee Break	
11:20 - 12:10	The Statistical Triangle	Jiashun Jin
12:10 - 14:30	Lunch Break	
14:30 - 15:20	Network Community Detection Using Higher-Order Structures	Ji Zhu
15:25 - 16:15	Optimality of Cp in the Scaled Lasso Path	Cunhui Zhang
16:15 - 16:40	Coffee Break	
16:40 - 17:30	Prediction with Confidence – General Framework for Predictive Inference	Regina Y. Liu
17:30 - 17:40	Closing	
18:30	Banquet (invitation only)	

HKU 2023 Summer Workshop on Statistics and Data Analytics

ABSTRSCTS

Songxi Chen, Peking University

Statistical Inference on a Four-Regime Segmented Regression Model

Segmented regression models are attractive for their flexibility and interpretability as compared to the global parametric and the nonparametric models, and yet are challenging in both estimation and inference. We consider a four-regime segmented model for temporally dependent data with two segmenting boundaries depending on multivariate covariates with non-diminishing boundary effects. A mixed integer quadratic programming algorithm is formulated to facilitate the least square estimation to both the regression and the boundary coefficients. The rates of convergence and the asymptotic distributions of the least square estimators are obtained, which shows differential convergence rates and limiting distributions between the regression and the boundary coefficients. Estimation and testing for degenerated segmented models with less than four segments are also considered with a testing procedure to decide if neighboring segments can be merged. Numerical simulations and a case study on air pollution in Beijing are conducted to demonstrate the proposed model and the inference results. In particular, it shows that the segmented models with three or four regimes are suitable for the modeling of the meteorological effects on the PM2.5 concentration. A joint work with Han Yan Guanghua School of Management, Peking University

Annie Qu, University of California, Irvine

A Model-Agnostic Graph Neural Network for Integrating Local and Global Information

Graph neural networks (GNNs) have achieved promising performance in a variety of graph focused tasks. Despite their success, the two major limitations of existing GNNs are the capability of learning various-order representations and providing interpretability of such deep learning-based black-box models. To tackle these issues, we propose a novel Model-agnostic Graph Neural Network (MaGNet) framework. The proposed framework is able to extract knowledge from high-order neighbors, sequentially integrates information of various orders, and offers explanations for the learned model by identifying influential compact graph structures. In particular, MaGNet consists of two components: an estimation model for the latent representation of complex relationships under graph topology, and an interpretation model that identifies influential nodes, edges, and important node features. Theoretically, we establish the generalization error bound for MaGNet via empirical Rademacher complexity and showcase its power to represent the layer-wise neighborhood mixing. We conduct comprehensive numerical studies using both simulated data and a real-world case study on investigating the neural mechanisms of the Rat Hippocampus, demonstrating that the performance of MaGNet is competitive with state-of-the-art methods.

Jiashun Jin, Carnegie Mellon University

The Statistical Triangle

In his Fisher's Lecture in 1996, Efron suggested that there is a philosophical triangle in statistics with "Bayesian", "Fisherian", and "Frequentist" being the three vertices, and most of the statistical methods can be viewed as a convex linear combination of the three philosophies. We collected and cleaned a data set consisting of the citation and bibtex (e.g., title, abstract, author information) data of 83,331 papers published in 36 journals in statistics and related fields, spanning 41 years. Using the data set, we constructed 21 co-citation networks, each for a time window between 1990 and 2015. We propose a dy- namic Degree-Corrected Mixed- Membership (dynamic-DCMM) model, where we model the research interests of an author by a low-dimensional weight vector (called the network memberships) that evolves slowly over time. We propose dynamic-SCORE as a new ap- proach to estimating the memberships. We discover a triangle in the spectral domain which we call the Statistical Triangle, and use it to visualize the research trajectories of individual authors. We interpret the three vertices of the triangle as the three primary research ar- eas in statistics: "Bayes", "Biostatistics" and "Nonparametrics". The Statistical Triangle further splits into 15 sub-regions, which we interpret as the 15 representative sub-areas in statistics. These results provide useful insights over the research trend and behavior of statisticians.

Ji Zhu, University of Michigan, Ann Arbor

Network Community Detection Using Higher-Order Structures

Many real-world networks commonly exhibit an abundance of subgraphs or higher-order structures, such as triangles and by-fans, surpassing what is typically observed in randomly generated networks. However, statistical models accounting for this phenomenon are limited, especially when community structure is of interest. This limitation is coupled with a lack of community detection methods that leverage subgraphs or higher-order structures. In this paper, we propose a novel community detection method that effectively incorporates these higher-order structures within a network. We also develop a finite-sample error bound for community detection accuracy under an edge-dependent network model, which includes both community and triangle structures. This error bound is characterized by the expected triangle degree, leading to the proposed method's consistency. To our knowledge, this is the first statistical error bound and consistency result considering a single network's community detection under a network model with dependent edges. Through simulations and a real-world data example, we demonstrate that our method reveals network communities otherwise obscured by methods that disregard higher-order structures. The talk is based on joint work with Xianshi Yu.

Cunhui Zhang, Rutgers University

Optimality of Cp in the Scaled Lasso Path

In sparse linear regression, the Lasso estimator requires a proper penalty to achieve the optimal rate in prediction error. Theory suggests that the proper penalty is proportional to the noise level of the regression model. The noise level is usually unknown and often treated as a nuisance parameter in theoretical studies. The scaled Lasso eliminates the dependence on the unknown noise level via an iterative minimization scheme. It essentially reduces the tuning parameter selection to a constant factor within a narrow band in a scaled Lasso path. Stein's unbiased risk estimator, or equivalently Cp in linear regression, is a commonly used criterion to select an estimator with minimal prediction error among a collection of candidates. We propose to use Cp to choose the penalty level in the scaled Lasso path to fine-tune the constant factor. Using second order Stein's methods, we prove a theoretical guarantee in the form of an oracle inequality that up to a higher-order term, Cp achieves the minimal prediction error within the proper band of penalty levels. Simulation studies under broad settings demonstrate the superior performance of the proposed method in supporting our theory. This talk is based on joint works with Pierre Bellec and Chong Wu.

Regina Y. Liu, Rutgers University

Prediction with Confidence – General Framework for Predictive Inference

We present a general framework for prediction in which a prediction is in the form of a distribution function, called ‘predictive distribution function’. This predictive distribution function is well suited for prescribing the notion of confidence under the frequentist interpretation and providing meaningful answers for prediction-related questions. Its very form of a distribution function also lends itself as a useful tool for quantifying uncertainty in prediction. A general approach under this framework is formulated and illustrated using the so-called confidence distributions (CDs). This CD-based prediction approach inherits many desirable properties of CD, including its capacity to serve as a common platform for directly connecting the existing procedures of predictive inference in Bayesian, fiducial and frequentist paradigms. We discuss the theory underlying the CD-based predictive distribution and related efficiency and optimality. We also propose a simple yet broadly applicable Monte-Carlo algorithm for implementing the proposed approach. This concrete algorithm together with the proposed definition and associated theoretical development provide a comprehensive statistical inference framework for prediction. Finally, the approach is illustrated by simulation studies and a real project on predicting the application submissions to a government agency. The latter shows the applicability of the proposed approach to even dependent data settings.

This is joint work with Jieli Shen, Goldman Sachs, and Minge Xie, Rutgers University.