HKU 2024 Summer Workshop on Statistics and Data Science

PROGRAM

Time	Title	Speaker
08:30 - 09:00	Registration	
09:00 - 09:10	Opening	
09:10 - 10:00	Simultaneous Decorrelation of Matrix Time Series	Cun-Hui Zhang
10:05 - 10:55	Spectral Change Point Estimation for High Dimensional Time Series by Sparse Tensor Decomposition	Kung-Sik Chan
10:55 - 11:20	Coffee Break	
11:20 - 12:10	Chain Graph Models: Identifiability, Estimation and Asymptotics	Junhui Wang
12:10 - 14:30	Lunch Break	
14:30 - 15:20	Cross-Sectional Data, Disease Dynamics and Beyond	Mei-Cheng Wang
15:25 - 16:15	Robust Estimation for the Number of Factors in a High Dimensional Factor Model via Spearman Correlations	Jianfeng Yao
16:15 - 16:40	Coffee Break	
16:40 - 17:30	Transfer Learning via Sufficient Representation	Jian Huang
17:30 - 17:40	Closing	
18:30	Banquet (invitation only)	

HKU 2024 Summer Workshop on Statistics and Data Science

ABSTRACTS

Cun-Hui Zhang, Rutgers University

Simultaneous Decorrelation of Matrix Time Series

We propose a contemporaneous bilinear transformation for a $p \times q$ matrix time series to alleviate the difficulties in modeling and forecasting matrix time series when p and/or q are large. The resulting transformed matrix assumes a block structure consisting of several small matrices, and those small matrix series are uncorrelated across all times. Hence an overall parsimonious model is achieved by modelling each of those small matrix series separately without the loss of information on the linear dynamics. Such a parsimonious model often has better forecasting performance, even when the underlying true dynamics deviates from the assumed uncorrelated block structure after transformation. The uniform convergence rates of the estimated transformation are derived, which vindicate an important virtue of the proposed bilinear transformation, i.e. it is technically equivalent to the decorrelation of a vector time series of dimension $\max(p, q)$ instead of $p \times q$. The proposed method is illustrated numerically via both simulated and real data examples. This is joint work with Yuefeng Han, Rong Chen and Qiwei Yao.

Kung-Sik Chan, University of Iowa

Spectral Change Point Estimation for High Dimensional Time Series by Sparse Tensor Decomposition

Multivariate time series may be subject to partial structural changes over certain frequency band, for instance, in neuroscience. We study the change point detection problem with high dimensional time series, within the framework of frequency domain. The overarching goal is to locate all change points and delineate which series are activated by the change, over which frequencies. In practice, the number of activated series per change and frequency could span from a few to full participation. We solve the problem by first computing a CUSUM tensor based on spectra estimated from blocks of the time series. A frequency-specific projection approach is applied for dimension reduction. The projection direction is estimated by a proposed tensor decomposition algorithm that adjusts to the sparsity level of changes. Finally, the projected CUSUM vectors across frequencies are aggregated for change point detection. We provide theoretical guarantees on the number of estimated change points and the convergence rate of their locations. We derive error bounds for the estimated projection direction for identifying the frequency-specific series activated in a change. We provide data-driven rules for the choice of parameters. The efficacy of the proposed method is illustrated by simulation and a stock returns application. The talk is based on joint work with Xinyu Zhang.

Junhui Wang, Chinese University of Hong Kong

Chain Graph Models: Identifiability, Estimation and Asymptotics

In this talk, we consider a flexible chain graph (CG) model, which admits both undirected and directed edges in one graph and thus can encode much more diverse relations among objects. We first establish the identifiability conditions for the CG model through a low rank plus sparse matrix decomposition, where the sparse matrix implies the sparse undirected edges within each chain component and the low rank matrix implies the presence of hub nodes with multiple children or parents. On this ground, we develop an efficient estimation method for reconstructing the CG structure, which first identifies the chain components via estimated undirected edges, determines the causal ordering of the chain components, and eventually estimates the directed edges among the chain components. Its theoretical properties will be discussed in terms of both asymptotic and finite-sample probability bounds on model estimation and graph reconstruction. The advantage of the proposed method is also demonstrated through extensive numerical experiments on both synthetic data and the Standard & Poor’s 500 index data.

Mei-Cheng Wang, Johns Hopkins University

Cross-Sectional Data, Disease Dynamics and Beyond

A cross-sectional population is defined as a specific population of living individuals at the sampling or observational time. Cross-sectionally sampled data with binary disease outcome are commonly analyzed in observational studies, frequently as an initial attempt, for the purpose of identifying how covariates or risk factors correlate with disease occurrences. At Johns Hopkins University, cross-sectional data analyses using standard methods (testing statistics, logistic regression, et) are commonly conducted in doctoral dissertations by students with public health or medicine majors. Publications in medicine or public health journals involving such data analysis can also be easily found online by searching the key words such as `logistic regression' or `logistic model' and `cross-sectional data' or `cross-sectional study.' It is generally understood that cross-sectionally collected binary disease outcome is not as informative as longitudinally collected time-to-event data, but there is insufficient understanding as to whether bias can possibly exist in cross-sectional data and, if it exists, how the bias is related to the population risk of interest. In this talk we study bias of absolute risk, relative risk and odds ratio arising from cross-sectional data via the birth-illness-death process, and connect the so-called 'survival bias' to cross-sectional population. While the presence of bias may or may not be surprising, the bad news is that the bias is likely to change the interpretation toward the wrong direction. With auxiliary information on additional data or distributions, bias-correction methods are possible but the additionally required information may not be easy to obtain. Recommendations are discussed/invited at the end of this talk to find ways to provide advice to our project collaborators, students and statisticians.

Jianfeng Yao, Chinese University of Hong Kong, Shenzhen

Robust Estimation for the Number of Factors in a High Dimensional Factor Model via Spearman Correlations

Determining the number of factors in high-dimensional factor models is crucial yet challenging, particularly when dealing with heavy-tailed data. In this paper, we introduce a novel estimator based on the spectral properties of the Spearman sample correlation matrix in high-dimensional settings, where both the dimension and sample size increase comparably. Our estimator is robust against heavy tails present in either common factors or idiosyncratic errors. We establish the consistency of our estimator under mild conditions. Numerical experiments demonstrate that our estimator outperforms existing methods.

This is a joint work with Jiaxin Qiu (HKU) and Zeng Li (SUSTech).

Jian Huang, Hong Kong Polytechnic University

Transfer Learning via Sufficient Representation

Transfer learning is an important approach for addressing the challenges posed by limited data availability in applications. It achieves this by transferring knowledge from well-established source domains to a less familiar target domain. However, conventional transfer learning methods often encounter difficulties due to rigid model assumptions and the necessity for a high degree of similarity between the source and target domains. In this paper, we present a method for transfer learning via sufficient data representation, which we refer to as TransRep. The key idea of TransRep is to enable knowledge transfer through intrinsic data representations, rather than relying on the transfer of model parameters. This approach allows TransRep to facilitate knowledge transfer across diverse tasks and model types. We evaluate the effectiveness of TransRep based on the generalization capacity of the representations, rather than the precise similarities between the target functions. Our theoretical analysis reveals that the extensive knowledge embedded in the source domains, as evidenced by a multitude of latent representations, constitutes a significant advantage of transfer learning. We assess the performance of our TransRep framework on finite samples through detailed simulations and empirical analyses using real-world data sets. This is joint work with Yeheng Ge and Xueyu Zhou.