高维统计研讨会

Random Matrices and Complex Data Analysis Workshop

Dates: December 10-12
(Arrival date: December 9; Departure date: December 13)

Venue: School of Statistics and Management,
Shanghai University of Finance and Economics

Organization and Scientific Committee: (alphabet sequence)

Zhidong Bai	Northeast Normal University
Xingdong Feng	Shanghai University of Finance and Economics
Weiming Li	Shanghai University of Finance and Economics
Jianfeng Yao	The University of Hong Kong

Acknowledgement: The workshop is sponsored by the School of Statistics and Management, Shanghai University of Finance and Economics. The organizers also thank support from Department of Statistics and Actuarial Science, The University of Hong Kong.

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Invited Speakers: (alphabet sequence)

Arup Bose	Indian Statistical Institute
Jinyuan Chang	Southwestern University of Finance and Economics
Songxi Chen	Peking University
Xiucai Ding	Duke University
Fang Han	University of Washington
Johannes Heiny	Ruhr University Bochum
David Hong	University of Pennsylvania
Jiang Hu	Northeast Normal University
Iain Johnstone	Stanford University
Olivier Ledoit	University of Zurich
Zhenyu Liao	CentraleSupelec
Miles Lopes	University of California, Davis
Matthew McKay	Hong Kong University of Science and Technology
Guangming Pan	Nanyang Technological University
Yanrong Yang	Australian National University
Wang Zhou	National University of Singapore
Xinghua Zheng	Hong Kong University of Science and Technology

Invited Posters: (alphabet sequence)

Daning Bi	Australian National University
Tieplova Daria	Universite Paris-Est Marne La Vallee
Dandan Jiang	Xi'an Jiaotong University
Zeng Li	Southern University of Science and Technology
Zhaoyuan Li	The Chinese University of Hong Kong, Shenzhen
Dangzheng Liu	University of Science and Technology of China
Wenya Luo	Northeast Normal University
Adam Nie	Australian National University
Zhenzhen Niu	Northest Normal University
Alexis Rosuel	Universite Paris-Est Marne La Vallee
Peng Tian	The Hong Kong University

Cheng Wang	Shanghai Jiao Tong University
Zhenggang Wang	The University of Hong Kong
Ruofan Xu	Australian National University
Bo Zhang	University of Science and Technology of China
Huiming Zhang	Peking University
Xiaozhuo Zhang	Northeast Normal University
Zhixiang Zhang	Nanyang Technological University
Changbo Zhu	University of Illinois at Urbana-Champaign

Doctoral Consortium: (alphabet sequence)

Chao Cheng	Shanghai University of Finance and Economics
Yeheng Ge	Shanghai University of Finance and Economics
Yan Liu	Northeast Normal University
Wenya Luo	Northeast Normal University
Yiming Tang	Shanghai University of Finance and Economics
Moming Wang	Shanghai University of Finance and Economics

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Agenda:

Dec 9 (Mon)		
14:00-19:00	Registration (Hall of the Howard Johnson Hotel)	
Dec 10 (Tue)		
08:45-09:00	Opening Ceremony	
Time	Speaker	Title
09:00-10:00	Iain Johnstone	Edgeworth and confidence interval correction in spiked PCA
10:00-10:30	Coffee Break	
10:30-11:30	Guangming Pan	High dimensional clustering: Covariance clustering for mixture data
11:30-12:30	Matthew McKay	Guiding rational vaccine design with random matrix theory
12:30-14:00	Lunch	
14:00-15:00	David Hong	Understanding parallel analysis methods for rank selection in PCA
15:00-16:00	Fang Han	Marginal and multivariate ranks, optimal transport theory, and Le Cam
16:00-16:30	Coffee Break	
16:30-17:30	Poster session	
17:30-18:30		

Dec 11 (Wed)		
Time	Speaker	Title
09:00-10:00	Songxi Chen	Multi-level thresholding test for high dimensional covariance matrices
10:00-10:30	Coffee Break	
10:30-11:30	Olivier Ledoit	Analytical nonlinear shrinkage of large-dimensional covariance matrices
11:30-12:30	Jinyuan Chang	Optimal covariance matrix estimation for high-dimensional noise in high-frequency data
12:30-14:00	Lunch	
14:00-15:00	Xinghua Zheng	Tests for principal eigenvalues and eigenvectors
15:00-16:00	Miles Lopes	Bootstrapping spectral statistics in high dimensions
16:00-16:30	Coffee Break	
16:30-17:30	Xiucai Ding	The landscape of separable covariance matrices
	Workshop Dinner	

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Dec 12 (Thu)		
Time	Speaker	Title
09:00-10:00	Arup Bose	Smallest singular value and limit eigenvalue distribution of a class of non-Hermitian random matrices with statistical application
10:00-10:30	Coffee Break	
10:30-11:30	Wang Zhou	Tracy-Widom limit for the largest eigenvalue of high-dimensional covariance matrices in elliptical distributions
11:30-12:30	Yanrong Yang	Nonparametric estimation for panel data models with heterogeneity and time-varyingness
12:30-14:00	Lunch	
14:00-15:00	Johannes Heiny	Spectral distributions of high-dimensional sample correlation matrices under infinite variance
15:00-16:00	Zhenyu Liao	Random matrix advances in large dimensional machine learning
16:00-16:30	Coffee Break	
16:30-17:30	Jiang Hu	Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression
17:30	Closing Remarks	

Dec 12 (Thu) Doctoral Consortium		
Time	Speaker	Title
13:30-14:00	Yiming Tang	Change point detection in dynamic networks using probit tensor factorization model
14:00-14:30	Wenya Luo	A modified BDS test
14:30-15:00	Chao Cheng	Robust subgroup analysis of high dimensional data
15:00-15:20	Coffee Break	
15:20-15:50	Yeheng Ge	Network-augmented feature screening for high dimensional censored data
15:50-16:20	Yan Liu	Community detection based on the L^∞ convergence of eigenvectors in DCBM
16:20-16:50	Moming Wang	On the estimation of high-dimensional integrated covariance matrices based on high-frequency data with multiple transactions

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Talks

Edgeworth and confidence interval correction in spiked PCA

Iain Johnstone Stanford University

The setting is principal components analysis with number of variables proportional to sample size, both large. The data are Gaussian with known spherical population covariance except for a fixed number of larger and distinct population eigenvalues, 'spikes'. If these spikes are large enough, i.e. 'supercritical', then to leading order the sample spike eigenvalues are known to be asymptotically independent Gaussian. We give the first order Edgeworth correction for this model (which is far from the usual smooth function of means setting) and note how repulsion of supercritical sample eigenvalues first becomes visible at this order. We outline implications for improved confidence intervals for the spike values, using a minimal conditioning strategy for post selection inference. This is joint work with Jeha Yang.

High dimensional clustering: Covariance clustering for mixture data

Guangming Pan Nanyang Technological University

This talk focuses on the clusters that are characterized by the different covariance matrices. We propose a new approach, covariance clustering method, to conduct clustering. Both theoretical and numerical properties of the covariance clustering method are discussed. Specifically, we propose one algorithm that is applicable to do the clustering in different settings. In addition, we prove that the misclustering error for this algorithm converges to zero with probability tends to one under mild conditions. Simulation studies also demonstrate that the covariance clustering method outperforms other methods under a variety of settings.

Guiding rational vaccine design with random matrix theory

Matthew McKay Hong Kong University of Science and Technology

This talk will describe how advances in random matrix theory (RMT) can aid the rational design of vaccines. With a focus on the hepatitis C virus (HCV) and the human immunodeficiency virus (HIV), high-dimensional RMT approaches will be introduced and applied to genetic sequence data measured from infected individuals. These approaches will be used to inform how the function/structure of viral proteins is mediated by correlated sets of genetic mutations, and to identify potential weaknesses that may be targeted by a vaccine. These results, used together with population-level immune system data, will be leveraged to specify novel vaccine candidates. The methodologies rely on spectral results of spiked correlation-coefficient matrices, which will also be discussed.

Understanding parallel analysis methods for rank selection in PCA

David Hong University of Pennsylvania

Principal Component Analysis (PCA) is a ubiquitous method for identifying latent factors that explain meaningful variation in data. An important question is how many factors to select in the analysis, and consequently many approaches have been developed for this task. We study the popular permutation-based parallel analysis method. This method randomly permutes each feature in the data, selecting all components whose singular values exceed the analogous singular values of the permuted data. Despite widespread use in leading textbooks and scientific publications, as well as empirical evidence for its accuracy, theoretical justification and understanding are current areas of active work. We show that this method consistently selects so-called perceptible components in certain high-dimensional factor models; small components that do not separate from the noise are imperceptible and are not selected. The key idea is that permutation "destroys" the low-rank structure from meaningful factors but not the noise, allowing the method to identify components "above the noise". We conclude with a discussion of ongoing work that replaces permutation with sign-flipping. This new signflip-based parallel analysis addresses some weaknesses of the permutation approach, and we illustrate how this modification improves performance for some multinomial models arising in genetics.

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Marginal and multivariate ranks, optimal transport theory, and Le Cam

Fang Han University of Washington

This talk aims to connect five keywords in statistics/probability -- ranks, (degenerate) U-statistics, combinatorial (non-) CLT, optimal transport theory, and Le Cam's contiguity lemma -- through one theme, nonparametric independence testing. The corresponding results show the existence of consistent rate-optimal distribution-free tests of two null hypotheses, mutual independence and independence of two random vectors, both for the first time. In technical terms, we give (1) the first Cramer-type moderate deviation theorem for degenerate U-statistics, (2) a new type of combinatorial non-central limit theorem for double- and multiple-indexed permutation statistics, and (3) a nontrivial use of Le Cam's third lemma with elements of non-normal limits.

Multi-level thresholding test for high dimensional covariance matrices

Songxi Chen Peking University

We consider testing the equality of two high-dimensional covariance matrices by carrying out a multi-level thresholding procedure, which is designed to detect sparse and faint differences between the covariances. A novel U-statistic composition is developed to establish the asymptotic distribution of the thresholding statistics in conjunction with the matrix blocking and the coupling techniques.
The multi-thresholding test is shown to be powerful in detecting sparse and weak differences between two covariance matrices. The test is shown to have attractive detection boundary and attain the optimal minimax rate in the signal strength under different regimes of high dimensionality and the sparsity of the signals.

Analytical nonlinear shrinkage of large-dimensional covariance matrices

Olivier Ledoit University of Zurich

This paper establishes the first analytical formula for optimal nonlinear shrinkage of large-dimensional covariance matrices. We achieve this by identifying and mathematically exploiting a deep connection between nonlinear shrinkage and nonparametric estimation of the Hilbert transform of the sample spectral density. Previous nonlinear shrinkage methods were numerical: QuEST requires numerical inversion of a complex equation from random matrix theory whereas NERCOME is based on a sample-splitting scheme. The new analytical approach is more elegant and also has more potential to accommodate future variations or extensions. Immediate benefits are that it is typically 1,000 times faster with the same accuracy, and accommodates covariance matrices of dimension up to 10,000. The difficult case where the matrix dimension exceeds the sample size is also covered.

Optimal covariance matrix estimation for high-dimensional noise in high-frequency data

Jinyuan Chang Southwestern University of Finance and Economics

In this paper, we consider efficiently learning the structural information from the high-dimensional noise in high-frequency data via estimating its covariance matrix with optimality. The problem is uniquely challenging due to the latency of the targeted high-dimensional vector containing the noises, and the practical reality that the observed data can be highly asynchronous – not all components of the high-dimensional vector are observed at the same time points. To meet the challenges, we propose a new covariance matrix estimator with appropriate localization and thresholding. In the setting with latency and asynchronous observations, we establish the minimax optimal convergence rates associated with two commonly used loss functions for the covariance matrix estimations. As a major theoretical development, we show that despite the latency of the signal in the high-frequency data, the optimal rates remain the same as if the targeted high-dimensional noises are directly observable. Our results indicate that the optimal rates reflect the impact due to the asynchronous observations, which are slower than that with synchronous observations. Furthermore, we demonstrate that the proposed localized estimator with thresholding achieves the minimax optimal convergence rates. We also illustrate the empirical performance of the proposed estimator with extensive simulation studies and a real data analysis.

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Tests for principal eigenvalues and eigenvectors

Xinghua Zheng Hong Kong University of Science and Technology

We establish Central Limit Theorems for principal eigenvalues and eigenvectors under a large factor model setting. As an application, we develop two-sample tests for both principal eigenvalues and principal eigenvectors. These tests can be used to detect structural breaks in large factor models. While there exist such tests, they can not distinguish between individual eigenvalues and/or eigenvectors. Our tests provide unique insights into the source of structural breaks.

Bootstrapping spectral statistics in high dimensions

Miles Lopes University of California, Davis

Statistics derived from the eigenvalues of sample covariance matrices are called spectral statistics, and they play a central role in multivariate testing. Although bootstrap methods are an established approach to approximating the laws of spectral statistics in low-dimensional problems, such methods are relatively unexplored in the high-dimensional setting. The aim of this work is to focus on linear spectral statistics as a class of prototypes for developing a new bootstrap in high dimensions, a method we refer to as the spectral bootstrap. In essence, the proposed method originates from the parametric bootstrap and is motivated by the fact that in high dimensions it is difficult to obtain a nonparametric approximation to the full data-generating distribution. In addition to proving the consistency of the proposed method, we present encouraging empirical results in a variety of settings. Lastly, and perhaps most interestingly, we show through simulations that the method can be applied successfully to statistics outside the class of linear spectral statistics, such as the largest sample eigenvalue and others.

The landscape of separable covariance matrices

Xiucai Ding Duke University

High-dimensional data obtained at space-time points has been increasingly employed in various scientific fields, such as geophysical and environmental sciences, wireless communications, medical imaging and financial economics. The structural assumption of separability is a popular assumption in the analysis of spatio-temporal data. In this talk, I will report the recent results on the eigenvalues and eigenvectors of separable sample covariance matrices. Specifically, we study a class of separable sample covariance matrices of the form $\widetilde{Q} = \widetilde{A}^{1/2} X \widetilde{B} X^\top \widetilde{A}^{1/2}$: Here $\widetilde{A} \in \mathbb{R}^{p\times p}$ and$\widetilde{B} \in \mathbb{R}^{n\times n}$ are positive definite matrices whose spectrums may consist of bulk spectrums plus several spikes, i.e. larger eigenvalues that are separated from the bulks and X is a rectangular matrix consisting of i.i.d. entries. We prove that when $p/n = O(1)$ and both $\widetilde{A}$ and $\widetilde{B}$ satisfy certain regularity conditions, the largest eigenvalues of $\widetilde{Q}$ will have Tracy-Widom distributions and its eigenvectors will be completely delocalized. Furthermore, when either $\widetilde{A}$ or $\widetilde{B}$ has spikes, we prove the convergence of the outlier eigenvalues and the generalized components of the outlier eigenvectors with optimal convergence rates. Moreover, we also prove the delocalization of the non-outlier eigenvectors. Some statistical applications will be discussed based on our results.

Smallest singular value and limit eigenvalue distribution of a class of non-Hermitian random matrices with statistical application

Arup Bose Indian Statistical Institute

Suppose X is an $N \times n$ complex matrix whose entries are centered, independent, and identically distributed random variables with variance $1/n$ and whose fourth moment is of order $\mathcal{O}(n^{-2})$. Suppose A is a deterministic matrix whose smallest and largest singular values are bounded below and above respectively, and $z \neq 0$is a complex number. First we consider the matrix $XAX^* - z$, and obtain asymptotic probability bounds for its smallest singular value when N and n diverge to infinity and $N/n \to \gamma,\ 0 \le \gamma \le \infty$. Then we consider the special case where $A = J = [1_{i-j=1 \mod n}]$ is a circulant matrix. Using the above result, we show that the limit spectral distribution of XJX^* exists when $N/n \to \gamma,\ 0 < \gamma < \infty$and describe the limit explicitly. Assuming that X represents a $\mathbb{C}^N$ -valued time series which is observed over a time window of length n , the matrix XJX^* represents the one-step sample autocovariance matrix of this time series. A whiteness test against an MA correlation model for this time series is introduced based on the above limit result. Numerical simulations show the excellent performance of this test.

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Tracy-Widom limit for the largest eigenvalue of high-dimensional covariance matrices in elliptical distributions

Wang Zhou National University of Singapore

Let X be an $M \times N$ random matrices consisting of independent M -variate elliptically distributed column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ with general population covariance matrix Σ . In the literature, the quantity XX^* is referred to as the sample covariance matrix, where X^* is the transpose of X . In this article, we show that the limiting behavior of the scaled largest eigenvalue of $XX^*$is universal for a wide class of elliptical distributions, namely, the scaled largest eigenvalue converges weakly to the same limit as $M, N \to \infty$ with $M/N \to \phi > 0$regardless of the distributions that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ follow. In particular, via comparing the Green function with that of the sample covariance matrix of multivariate normally distributed data, we conclude that the limiting distribution of the scaled largest eigenvalue is the celebrated Tracy-Widom law.

Nonparametric estimation for panel data models with heterogeneity and time-varyingness

Yanrong Yang Australian National University

Panel data subject to heterogeneity in both cross-sectional and time-serial directions are commonly encountered across social and scientific fields. To address this problem, we propose a class of time-varying panel data models with individual-specific regression coefficients and interactive common factors. This results in a model capable of describing heterogeneous panel data in terms of time--varyingness in the time-serial direction and individual-specific coefficients among cross-sections. Another striking generality of this proposed model relies on its compatibility with endogeneity in the sense of interactive common factors. Model estimation is achieved through a novel duple least--square (DLS) iteration algorithm, which implements two least--square estimation methods recursively and incorporates nonparametric kernel estimation for time--varying coefficients simultaneously. Its unified ability in estimation is nicely illustrated according to flexible applications on various cases with exogenous or endogenous common factors. Established asymptotic theory for the proposed DLS estimators benefits practitioners by demonstrating effectiveness of iteration in eliminating estimation biases gradually along with iterative steps. We further show that our model and estimation method perform well on simulated data in various scenarios as well as an OECD healthcare expenditure dataset. The time--varyingness and heterogeneity among cross-sections are confirmed by our empirical analysis.

Spectral distributions of high-dimensional sample correlation matrices under infinite variance

Johannes Heiny Ruhr University Bochum

In the first part of this talk, we consider the sample correlation matrix R associated to a $p \times n$ data matrix with iid entries. Assuming $p/n \to \gamma \in (0, \infty)$, the optimal condition on the entries for the convergence of the empirical spectral distributions to the Marchenko-Pastur law turns out to be slightly weaker than normal domain of attraction. In the case of entries with infinite $(2-\epsilon)$ -moments, we find a new class of Marchenko-Pastur type laws as limiting spectral distributions of R and compute their moments.
In the second part, we study point process convergence for sequences of iid random walks. The objective is to derive asymptotic theory for the extremes of these random walks. We show convergence of the maximum random walk to the Gumbel distribution under the existence of a $(2+\delta)$ th moment. We make heavily use of precise large deviation results for sums of iid random variables. As a consequence, we derive the joint convergence of the off-diagonal entries in sample covariance and correlation matrices of a high-dimensional sample whose dimension increases with the sample size. This generalizes known results on the asymptotic Gumbel property of the largest entry.

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Random matrix advances in large dimensional machine learning

Zhenyu Liao CentraleSupelec

The advent of the Big Data era has triggered a renewed interest in large dimensional machine learning problems. These methods, however, suffer from a double plague 1) as they involve nonlinear operators (e.g., kernel function in kernel methods or activation functions in a neural network context) and often arise from optimization problems that have only implicit solutions, they are difficult to fathom and rarely offer performance guarantees or hyper-parameter control and 2) they were often developed from small dimensional intuitions and tend to be highly suboptimal in handling real-world large dimensional problems. Recent advances in random matrix theory manage to simultaneously solve both problems: in assuming the dimension the size of datasets to be both large, concentration phenomena arise that allow for a renewed understanding and the possibility to assess, understand, and improve machine learning approaches, thereby opening the door to completely new paradigms.
In this talk, with the example of kernel-based learning, we highlight the counterintuitive "curse of dimensionality" phenomenon in large dimensional learning problems. With a simple Gaussian mixture modeling of the input data, we provide sharp performance prediction of various kernel-based methods and observe a surprisingly close match between theory and experiments on popular real-world datasets. We also briefly talk about the on-going research in large dimensional analyses of optimization-based methods, as well as a more realistic, concentration-based modeling of data that reflects some kind of universality in large dimensional machine learning.

Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression

Jiang Hu Northeast Normal University

Variable selection is essential for improving inference and interpretation in multivariate linear regression. Although a number of alternative regressor selection criteria have been suggested, the most prominent and widely used are the Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallow's C_p, and their modifications. However, for the data with high dimensionality in both responses and covariates, experience has shown that the performance of these classical criteria is not always satisfactory. In the present article, we begin by presenting the necessary and sufficient conditions (NSC) for the strong consistency of the high-dimensional AIC, BIC, and C_p, based on which we can identify some reasons for their poor performance. Specifically, we show that under certain mild high-dimensional conditions, if the BIC is strongly consistent, then the AIC is strongly consistent, but not vice versa. This result contradicts the classical understanding. In addition, we consider some NSC for the strong consistency of the high-dimensional knock-one-out (KOO) methods introduced by Nishii et al. (1988). Furthermore, we propose two general methods based on the KOO methods and prove their strong consistency. The proposed general methods remove the penalties while simultaneously reducing the conditions for the dimensions and sizes of the regressors. Simulation studies and real data analysis support our conclusions and show that the convergence rates of the two proposed general KOO methods are much faster than those of the original methods.

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Poster Session

Homogeneity and sub-homogeneity pursuit: An iterative complement clustering PCA

Daning Bi Australian National University

On the limit distribution of the canonical correlation coefficients between the past and the future of a high-dimensional white noise

Tieplova Daria Universite Paris-Est Marne La Vallee

Generalized fourth moment theorem and an application to CLT for spiked eigenvalues of high dimensional covariance matrices

Dandan Jiang Xi'an Jiaotong University

On testing high dimensional white noise

Zeng Li Southern University of Science and Technology

On John's test for sphericity in large fixed effects panel regression model

Zhaoyuan Li The Chinese University of Hong Kong, Shenzhen

Phase transitions of the largest singular value for products of complex Gaussian random matrices

Dangzheng Liu University of Science and Technology of China

A modified BDS test

Wenya Luo Northeast Normal University

Infinite dimensional Levy processes and stochastic time flow via weak subordination

Adam Nie Australian National University

On LR simultaneous test of high-dimensional mean vector and covariance matrix under non-normality

Zhenzhen Niu Northest Normal University

On the behavior of estimated spectral coherence matrix of uncorrelated high dimensional time series, and applications

Alexis Rosuel Universite Paris-Est Marne La Vallee

Study of separable multivariate long range dependent series using random matrix theory

Peng Tian The Hong Kong University

Penalized interaction estimation for ultrahigh dimensional quadratic regression

Cheng Wang Shanghai Jiao Tong University

Central limit theorem for linear spectral statistics of Wigner-type matrices with block structure

Zhenggang Wang The University of Hong Kong

Forecasting for spatial functional time series

Ruofan Xu Australian National University

A new estimator of near unit root for high dimensional nonstationary time series

Bo Zhang University of Science and Technology of China

High-dimensional negative binomial regression—consistency and weak signals detection

Huiming Zhang Peking University

Analysis of the limiting spectral distribution of noncentral F type matrices

Xiaozhuo Zhang Northeast Normal University

Asymptotic independence of spiked eigenvalues and linear spectral statistics for large sample covariance matrices

Zhixiang Zhang Nanyang Technological University

Interpoint distance based two sample tests in high dimension

Changbo Zhu University of Illinois at Urbana-Champaign

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Doctoral Consortium

Robust subgroup analysis of high dimensional data

Chao Cheng Shanghai University of Finance and Economics

It becomes more and more attractive to identify subgroup structures in data analysis as populations are probably heterogeneous. The pairwise fusion penalty approach is an adaptive and data-driven method. In this paper, we consider the penalized M-estimators, which can deal with high dimensional data coupled with outliers. The penalties are applied both on covariates and treatment effects. Hence the estimation is expected to achieve both variable selection and data clustering simultaneously. An algorithm is proposed that can process relatively large datasets based on parallel computation. The convergence analysis of the algorithm, the oracle property of the estimators, and the selection consistency of the modified BIC criterion are also established. The performance of the method is assessed with an intensive simulation study.

Network-augmented feature screening for high dimensional censored data

Yeheng Ge Shanghai University of Finance and Economics

Feature screening is playing a vital role in ultra-high dimensional data analysis. With unique characteristics of censoring, screening methods for survival data are more challenging and are still lacking compared to continuous and categorical responses. Despite considerable successes, the existing survival feature screening methods share a common limitation that the importance of dependence network structure among predictors is less considered. In this study, we propose a robust network-based feature screening method for survival data, where this network structure is well accommodated. The proposed method is based on the graph Laplacian regularization and partial absorbing random walk techniques, which has a solid statistical ground and efficient realized algorithm. Sure screening properties under ultra-high dimensional setting are rigorously established. Extensive simulation studies show that the proposed method outperforms the alternatives whenever the network information is given a priori or estimated from the observed data. Analysis of the Cancer Genome Atlas breast cancer data lead to biologically sensible findings with better predictive accuracy and selection stability.

Community detection based on the convergence of eigenvectors in DCBM

Yan Liu Northeast Normal University

Spectral clustering is one of the most popular algorithms for community detection in network analysis. Based on this rationale, in this paper we give the convergence rate of eigenvectors for the adjacency matrix in the l^{∞} norm, under the stochastic block model (BM) and degree corrected stochastic block model (DCBM), adding some mild and rational conditions. We also extend this result to a more general model, presented based on the DCBM such that the value of random variables in the adjacency matrix is not 0 or 1, but an arbitrary real number. During the process of proving the above conclusion, we obtain the relationship of the eigenvalues in the adjacency matrix and the corresponding `population' matrix, which vary in dimension from the community-wise edge probability matrix. Using that result, we can give an estimate of the number of the communities in a known set of network data. Meanwhile we proved the consistency of the estimator. Furthermore, according to the derivation of proof for the convergence of eigenvectors, we propose a new approach to community detection -- Spectral Clustering based on Difference of Ratios of Eigenvectors (SCDRE). Our simulation experiments demonstrate the superiority of our method in community detection.

A modified BDS test

Wenya Luo Northeast Normal University

BDS test is used to test whether a given sequence of random variables is i.i.d. (independent and identically distributed). The BDS test has been used in economics and finance to examine a fitted time series model is adequate by testing the residual sequence is nearly i.i.d. Though BDS test is widely used, it has a drawback of over-rejection even though the sample size T is moderately large, such as $T \in (100,1000)$. In this study, we propose a MBDS test (modified BDS test) by removing some items from the correlation integral which is the root of BDS test. Theoretical calculation and simulation results support that MBDS test correct the bias of BDS test effectively.

高维统计研讨会
Random Matrices and Complex
Data Analysis Workshop

Change point detection in dynamic networks using probit tensor factorization model

Yiming Tang Shanghai University of Finance and Economics

The probit tensor factorization (PTF) model is a newly proposed model which shows advantages in both prediction accuracy and interpretability in the area of statistical relational learning. In this paper, we propose a change point detection method for dynamic networks which penalizes the difference of the Frobenius norms of latent factor matrices at adjacent time in the PTF model. Some simulation results are presented to show both the estimation accuracy and the change point detection accuracy of our method. We also create and analyze the dynamic network of stocks of America's largest 62 public companies in 2006-2008 to detect the change points during the global financial crisis.

On the estimation of high-dimensional integrated covariance matrices based on high-frequency data with multiple transactions

Moming Wang Shanghai University of Finance and Economics

High-frequency data in financial markets often include multiple transactions at each recording time due to the mechanism of recording. Using random matrix theory, this paper considers the estimation of integrated covariance (ICV) matrices of high-dimensional diffusion processes based on multiple high-frequency observations. We start by studying the estimator, the time-variation adjusted realized covariance (TVA) matrix proposed by Zheng and Li (2011), without microstructure noise. We show that in the high-dimensional case, for a class C of diffusion processes, the limiting spectral distribution (LSD) of the averaged TVA depends not only on that of the ICV but also on the number of multiple transactions at each recording time. However, in practice, observed prices are always contaminated by market microstructure noise. Thus, we also study the limiting behavior of pre-averaging averaged TVA matrices based on noisy multiple observations. We show that for processes in class C, the pre-averaging averaged TVA has two desirable properties: it eliminates the effects of microstructure noise and multiple transactions, and its LSD depends solely on that of the ICV matrix. Further, three types of nonlinear shrinkage estimators of the ICV matrix are proposed based on high-frequency noisy multiple observations. Simulation studies support our theoretical results and demonstrate the finite sample performance of the proposed estimators. Finally, high-frequency portfolio strategies are evaluated under these estimators in real data analysis.