LARGE-SCALE NONPARAMETRIC AND SEMIPARAMETRIC INFERENCE FOR LARGE, COMPLEX, AND NOISY DATASETS

by

Fang Han

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

July, 2015

Abstract

Massive Data bring new opportunities and challenges to data scientists and statisticians. On one hand, Massive Data hold great promises for discovering subtle population patterns and heterogeneities that are not possible with small-scale data. On the other hand, the size and dimensionality of Massive Data introduce unique statistical challenges and consequences for model misspecification. Some important factors are as follows.

Complexity: Since Massive Data are often aggregated from multiple sources, they often exhibit heavy-tailedness behavior with nontrivial tail dependence.

Noise: Massive Data usually contain various types of measurement error, outliers, and missing values.

Dependence: In many data types, such as financial time series, functional magnetic resonance image (fMRI), and time course microarray data, the samples are dependent with relatively weak signals.

These challenges are difficult to address and require new computational and statistical tools. More specifically, to handle these challenges, it is necessary to develop statistical methods

ABSTRACT

that are robust to data complexity, noise, and dependence. Our work aims to make headway in resolving these issues. Notably, we give a unified framework for analyzing high dimensional, complex, noisy datasets having temporal/spatial dependence. The proposed methods enjoy good theoretical properties. Their empirical usefulness is also verified in large-scale neuroimage and financial data analysis.

Advisors:

Han Liu, PhD and Brian Caffo, PhD

Committee:

Terri Beaty, PhD (chair); Suchi Saria, PhD

Alternates:

Hongkai Ji, PhD; Fernando Pineda, PhD

Acknowledgments

I would love to thank my advisors, Han and Brian, for their consistent support and guidance throughout my PhD study. Their ceaseless care, inspiration, and encouragement are invaluable to me, and lead me to the broad area of statistics and its applications to neuroscience. I would also love to thank Mei-Cheng Wang, another of my mentors, for her years of support and friendship.

I also wish to thank my fellow students, Yang Ning, Shanshan Li, Tuo Zhao, Haochang Shou, Yingying Wei, Zhenke Wu, Juemin Yang, and Li Chen for their friendship. Special thanks go to Huitong Qiu and Sheng Xu, two very close friends at the Hopkins.

Finally, I am thankful to my family. I am indebted to my parents and my wife, Ning. Their support throughout my life is incredible. It would be impossible for me to finish this work without their encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We are entering the era of Massive Data — a term that refers to the explosion of available information. This Massive Data movement is driven by the fact that large amounts of high dimensional data are routinely produced and stored in large volume cheaply. In genomics, in only a few years, we have seen a dramatic drop in price for whole genome sequencing, going from millions of dollars and man hours to sequence one human genome, to now routine whole genome sequence for all subjects in observational studies. Similar revolutions in measurement have occurred in other areas, like social media analysis, biomedical imaging, and high frequency finance. The existing trend that data can be produced and stored more massively and cheaply is likely to be maintained, or even accelerated, in the future. This trend will have a deep impact on science, engineering, and business. For example, scientific advances are becoming more and more data-driven, and researchers increasingly think of themselves as consumers of data. The massive amounts of high dimensional data bring both opportunities and new challenges to data analysis. As such, statistical analysis for such data is becoming increasingly important.

We distinguish between Massive Data and Big Data. Whereas Big Data usually refers to a large number of records, Massive Data are characterized by both high dimensionality and large numbers of records. In addition, Massive Data are often aggregated from multiple sources at different time points. This creates the issues of heterogeneity and experimental variation. Such heterogeneity requires us to develop more adaptive and robust procedures, which is precisely the aim of this work.

In detail, there exist several Massive Data qualities worthy of attention. These include:

(1) complexity: Massive Data are often an aggregation of multiple subpopulations, and often exhibit heavy tails and tail dependency. Accordingly, we must deal with heterogeneity and non-Gaussian models. (2) Noise: Because the data are usually aggregated from numerous sources, they frequently contain various types of measurement error, endogenous covariates, outliers, and missing values. (3) Dependence: In many modern data types, such as equity, fMRI, and time course microarray data, the samples are not independent with relatively weak signals. To handle these challenges, it is necessary to develop statistical methods that are robust to data complexity, noise, and dependence.

We begin by briefly summarizing the key ideas in this thesis. For addressing the challenge of data complexity, our main focus is on the heavy-tailed data generated in many different areas (e.g., finance, social media, imaging). In conventional statistics, heavy-tailedness is often not an issue and can usually be addressed via parametrically modeling. However, such approaches are questionable in high dimensions, and especially for Massive Data. This is due to the data complexity phenomena: It is commonly too difficult, or even impossible, to have a parametric model that could fully capture the characteristics in Massive Data.

In the high dimensional statistics literature, there has not been a good balance between heavy-tailedness and flexible modeling. In high dimensions, for addressing the flexible modeling issue, it is most common to assume a nonparametric subGaussian model, i.e., assuming data are marginally or multivariately Gaussian distributed. Moment-based estimators are then encouraged, and different kinds of minimax optimal procedures are pro-

posed for different problems. However, there are two fundamental problems in this line of research. First, the nonparametric subGaussian models fail to capture heavy-tailedness. Second, moment-based estimators induced by subGaussian models perform poorly when data are heavy-tailed.

In light of these facts, this thesis discusses a unified framework. We propose semiparametric modeling approaches coupled with nonparametric statistical methods. There are two main insights. First, the semiparametric modeling approaches take a good balance between heavy-tailedness and flexible modeling. Second, induced nonparametric methods are actually optimal in parameter estimation, of performance comparable to Gaussian-based approaches under the Gaussian assumption.

For understanding the first insight, note a semiparametric model is one that has both finite- and infinite-dimensional parameters, and (usually) the parameters of interest are finite-dimensional. This thesis consists of two important semiparametric models: the elliptical and transelliptical. The elliptical model is constructed via randomly scaling the Gaussian, while the transelliptical modeling combines the ideas of random scaling and copulas for constructing flexible statistical models. In both models, there exist finite-dimensional parameters (mean and covariance of the latent Gaussian), as well as infinite-dimensional parameters (scaling random variables and marginal transformations). In addition, the elliptical distribution can have arbitrarily symmetric margins, while the transellitpical can have arbitrary margins. Hence, both the elliptical and transelliptical could be arbitrarily heavy-tailed. They are suitable candidates for flexibly modeling the heavy-tailed data.

In detail, the elliptical is a one-layer extension, while the transelliptical is a two-layer extension of the Gaussian. In the first layer, the Gaussian random vector, $\boldsymbol{Y}$, is stochastically scaled by a random variable, η, producing an elliptical random vector, $\boldsymbol{Z}$. In the second layer, unspecified univariate strictly increasing functions, $g_1, \ldots, g_d$, are applied to the margins of the elliptical random vector, $\boldsymbol{Z}$, producing the transelliptical random vector, $\boldsymbol{X}$. Figure 1.1 illustrates the generating scheme of the transelliptical distribution.

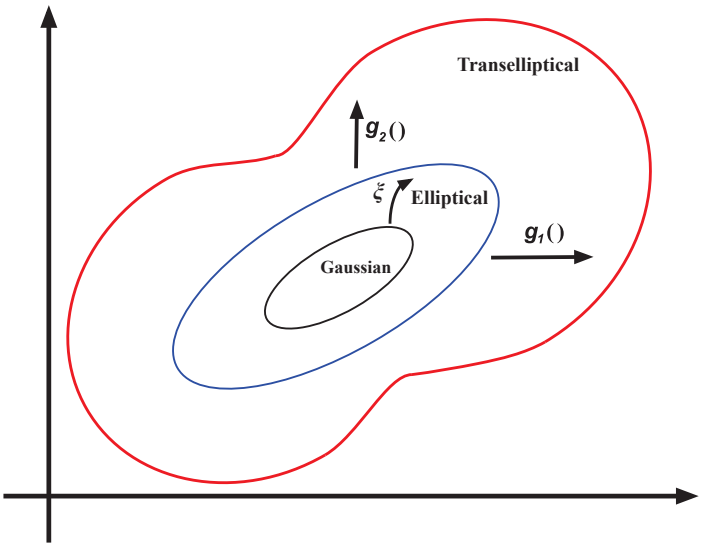


Figure 1.1: The graph illustrating the data generating schemes of the elliptical and transelliptical distributions. The black, blue, and red curves illustrate the contours of Gaussian, elliptical, and transelliptical distributions. Here the Gaussian is first scaled by a positive random variable ξ to the elliptical, then marginally transformed by two strictly increasing functions g_1, g_2 to the transelliptical.

For understanding the second insight, we note both the elliptical and transelliptical assume certain geometric constraints on the data. For the elliptical, it is symmetry. For the transelliptical, it is a copula structure coupled with symmetry in the latent layer. In comparison, the nonparametric nonGaussian models require data light-tailedness, and hence

require certain moment constraints on the data. Via replacing moment constraints by geometric constraints, we are able to design statistically efficient methods which fully exploit such model structures, and we prove to be actually optimal.

After obtaining these insights, we proceed to give an overview on the following chapters. This thesis consists of five journal articles, each of which is put in a chapter and designed in the spirit of addressing the issues raised from Massive Data analysis using the main ideas discussed above. For this, Chapters 2 and 3 are focused on the transelliptical model, while Chapter 4 is focused on the elliptical model. Chapter 5 discusses an even more flexible distribution-free model coupled with nonparametric methods. Chapter 6 then illustrates an application to the brain imaging data.

In detail, in Chapter 2, we present a work addressing how to efficiently conduct scale-invariant PCA on possibly heavy-tailed data. Specifically, Chapter 2 introduces a semiparametric model: the meta-elliptical (also called the transelliptical) model. Building on this model, we introduce a method for conducting scale-invariant sparse principal component analysis (PCA) on high dimensional non-Gaussian data, called Transelliptical Component Analysis (TCA). Compared to sparse PCA, TCA has weaker modeling assumptions and is more robust to possible data contamination. Theoretically, TCA achieves a parametric rate of convergence in estimating the parameter of interests under a flexible semiparametric distribution family. Computationally, TCA exploits a rank-based procedure and is as efficient as sparse PCA. Empirically, TCA outperforms most competing methods on both synthetic and real-world economics datasets. Chapter 3 further shows TCA is minimax optimal in

conducting estimation of eigenvectors.

Although TCA proves to be a statistically efficient approach for conducting PCA on possibly heavy-tailed data, its strength is constrained because it can only conduct scale-invariant PCA. For further relaxing this, Chapter 4 introduces a robust alternative to principal component analysis (PCA) — named elliptical component analysis (ECA). ECA works for possibly heavy-tailed but symmetric data. ECA aims at estimating the eigenspace of the covariance matrix of elliptical data. To cope with the heavy-tailed elliptical distributions, a multivariate rank statistic is exploited. At the model-level, we consider two settings where the leading eigenvectors of the covariance matrix are either non-sparse or sparse. Methodologically, we propose ECA procedures corresponding to both non-sparse and sparse settings. Theoretically, we provide both non-asymptotic and asymptotic analysis in quantifying the theoretical performances of ECA. Under the non-sparse setting, we show ECA's performance is highly related to the effective rank of the covariance matrix. Under the sparse setting, the results are twofold. First, we show that the sparse ECA estimator based on a combinatoric program attains the optimal rate of convergence. Second, building upon some recent developments in estimating sparse leading eigenvectors, we show a computationally efficient sparse ECA estimator can attain the optimal rate of convergence under a suboptimal scaling. We also apply ECA to study a brain imaging data extracted from Autism Brain Imaging Data Exchange (ABIDE) project, and show ECA has the potential to deliver better results for inference based on these estimated principal components.

With all the above chapters focusing on high dimensional robust estimation procedures,

Chapter 5 proceeds to address high dimensional robust testing problems. In particular, Chapter 5 considers the problem of testing mutual independence of all entries in a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ based on n independent observations. For this, we consider two families of distribution-free test statistics that converge weakly to an extreme value type I distribution. We further study the power of the corresponding tests against alternatives. In particular, we show the power approaches one when the maximum magnitude of the pairwise Pearson's correlation coefficients is larger than $C\sqrt{\log d/n}$ for some absolute constant C. This result is rate optimal. As important examples, we show the tests based on Kendall's tau and Spearman's rho are rate optimal tests of independence. For further generalization, we consider accelerating the rate of convergence by approximating the exact distributions of these test statistics. This section also studies the tests of two more structural hypotheses: m-dependence and data homogeneity. For these, we propose two rank-based tests and show their optimality against certain alternatives (More details will be provided in Chapter 5.).

Built on the robust procedures proposed in the previous chapters, the last chapter considers an extensive data analysis of brain imaging data. In detail, Chapter 6 presents a unified framework for conducting inference on complex aggregated data in high dimensional settings, which are strongly motivated by the intrinsic structure of the brain imaging data. We assume these data are a collection of multiple non-Gaussian realizations with underlying undirected graphical structures. Utilizing the concept of median graphs in summarizing the commonality across these graphical structures, we provide a novel semipara-

metric approach to modeling such complex aggregated data, along with robust estimation of the median graph itself, which is assumed to be sparse. We prove that the estimator is consistent in graph recovery and give an upper bound on its rate of convergence. We further provide experiments on both synthetic and real datasets to illustrate the empirical usefulness of the proposed models and methods. In particular, an extensive study on the ADHD-200 brain imaging dataset, of subjects with and without attention deficit hyperactive disorder (ADHD), was conducted.

Chapter 2

Scale-Invariant Sparse PCA on High Dimensional Meta-Elliptical Data

2.1 Introduction

Principal component analysis (PCA) is a powerful tool for reducing the dimensions of large data sets and helping identify key features in large datasets. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be n observations of a d-dimensional random vector $\boldsymbol{X}$ with covariance matrix $\boldsymbol{\Sigma}$. PCA aims at estimating the leading eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of $\boldsymbol{\Sigma}$ that best explain patterns of clustering in a dataset.

When the dimension d is small compared with the sample size n, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ can be consistently estimated by the leading eigenvectors $\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_m$ of the sample covariance matrix (Anderson, 1958). However, when d increases at the same order or even faster than n, this approach can lead to poor estimates. In particular, Johnstone and Lu (2009) showed the angle between $\widehat{\boldsymbol{u}}_1$ and $\boldsymbol{u}_1$ may not converge to 0 if $d/n \to c$ for some constant $c > 0$. To handle this challenge, one popular assumption is to impose sparsity constraint on the leading eigenvectors. For example, when estimating the leading eigenvector $\boldsymbol{u}_1 := (u_{11}, \ldots, u_{1d})^T$, we may assume that $s := \text{card}(\{j : u_{1j} \neq 0\}) < n$. Under this assumption, different variants of sparse PCA have been developed, more details can be found in d'Aspremont et al. (2007), Zou et al. (2006), Shen and Huang (2008), Witten et al. (2009), Journée et al. (2010), and Zhang and El Ghaoui (2011). The theoretical properties of sparse PCA in feature selection and parameter estimation have been investigated by Amini and Wainwright (2009), Ma (2013), Paul and Johnstone (2012), Vu and Lei (2012), and Berthet and Rigollet (2012).

There are several drawbacks of the classical PCA and sparse PCA approaches: (i) Nei-

ther approach is scale-invariant, i.e., changing the measurement scale of variables makes the estimates different (Chatfield and Collins, 1980); (ii) It is not robust to possible data contamination or outliers (Puri and Sen, 1971); and (iii) The theory of sparse PCA relies heavily on the Gaussian or sub-Gaussian assumption, which may not be realistic for many real-world applications.

In the low dimensional settings, remedies for drawbacks (ii) and (iii) include generalizing the Gaussian distribution to elliptical distribution (Fang et al., 1990), and considering some robust estimators (Huber and Ronchetti, 2009). One research line is to develop various PCA estimators for the elliptical data (Möttönen and Oja, 1995; Choi and Marden, 1998; Marden, 1999; Visuri et al., 2000; Croux et al., 2002; Jackson and Chen, 2004). Theoretical properties of these elliptical distribution based PCA estimators have been established under the classical asymptotic framework (where the dimension d is fixed) by Hallin et al. (2010), Oja (2010), and Croux and Dehon (2010). Along another research line, multiple robust PCA estimators have been proposed to address the outlier and heavy tailed issues via replacing the sample covariance matrix by a robust scatter matrix. Such robust scatter matrix estimators include M-estimator (Maronna, 1976), S-estimator (Davies, 1987), median absolute deviation (MAD) proposed by Hampel (1974), and S_n and Q_n estimators (Rousseeuw and Croux, 1993). These robust scatter matrix estimators have been exploited to conduct robust (sparse) principal component analysis (Gnanadesikan and Kettenring, 1972; Maronna and Zamar, 2002; Hubert et al., 2002; Croux and Ruiz-Gazen, 2005; Croux et al., 2013). The theoretical performances of PCA based on these robust

estimators in low dimensions were further analyzed in Croux and Haesbroeck (2000).

Here we propose a new method for conducting sparse principal component analysis on non-Gaussian data. Our method can be viewed as a scale-invariant version of sparse PCA, but is applicable to a wide range of distributions belonging to the meta-elliptical family (Fang et al., 2002). The meta-elliptical (also called the transelliptical) family extends the elliptical family. In particular, a continuous random vector $\boldsymbol{X} := (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ follows a meta-elliptical distribution if there exists a set of univariate strictly increasing functions $f := \{f_j\}_{j=1}^d$ such that $f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T$ follows an elliptical distribution with location parameter $\mathbf{0}$ and scale parameter $\boldsymbol{\Sigma}^0$, whose diagonal values are all 1. We call $\boldsymbol{\Sigma}^0$ the *latent generalized correlation matrix*. By treating $\{f_j\}_{j=1}^d$ as nuisance parameters, our method estimates the leading eigenvector $\boldsymbol{\theta}_1$ of $\boldsymbol{\Sigma}^0$ by exploiting a rank-based estimating procedure and can be viewed as a scale-invariant PCA conducted on $f(\boldsymbol{X})$. Theoretically we show when s is fixed, it achieves a parametric rate of convergence in estimating the leading eigenvector. Computationally, it is as efficient as sparse PCA. Empirically, we show the proposed method outperforms the classical sparse PCA and two robust alternatives on both synthetic and real-world datasets.

The rest of this chapter is organized as follows. In the next section, we review the elliptical distribution family and introduce the meta-elliptical distribution. In Section 2.3, we present the statistical model, introduce the rank-based estimators, and provide computational algorithm for parameter estimation. In Section 2.4, we provide theoretical analysis. In Section 2.5, we provide empirical studies on both synthetic and real-world datasets.

More discussion and comparison with related methods are in the last section.

2.2 Elliptical and Meta-Elliptical Distributions

In this section, we briefly review the elliptical distribution and introduce the meta-elliptical distribution family. We start by first introducing the notation: Let $\mathbf{M} = [\mathbf{M}_{jk}] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$ be a d-dimensional matrix and a d-dimensional vector. We denote $\boldsymbol{v}_I$ to be the subvector of $\boldsymbol{v}$ whose entries are indexed by a set I. We also denote $\mathbf{M}_{I,J}$ to be the submatrix of $\mathbf{M}$ whose rows are indexed by I and columns are indexed by J. Let $\mathbf{M}_{I*}$ and $\mathbf{M}_{*J}$ be the submatrix of $\mathbf{M}$ with rows in I, and the submatrix of $\mathbf{M}$ with columns in J. Let $\text{supp}(\boldsymbol{v}) := \{j : v_j \neq 0\}$. For $0 < q < \infty$, we define the ℓ_0, ℓ_q and ℓ_∞ vector norms as $\|\boldsymbol{v}\|_0 := \text{card}(\text{supp}(\boldsymbol{v}))$, $\|\boldsymbol{v}\|_q := (\sum_{i=1}^d |v_i|^q)^{1/q}$ and $\|\boldsymbol{v}\|_\infty := \max_{1\leq i\leq d} |v_i|$. We define the matrix $\ell_{\max}$ norm as the elementwise maximum value: $\|\mathbf{M}\|_{\max} := \max\{|\mathbf{M}_{ij}|\}$. Let $\Lambda_j(\mathbf{M})$ be the j-th largest eigenvalue of $\mathbf{M}$. In particular, we denote $\Lambda_{\min}(\mathbf{M}) := \Lambda_d(\mathbf{M})$ and $\Lambda_{\max}(\mathbf{M}) := \Lambda_1(\mathbf{M})$ to be the smallest and largest eigenvalues of $\mathbf{M}$. Let $\|\mathbf{M}\|_2$ be the spectral norm of $\mathbf{M}$. We define $\text{vec}(\mathbf{M}) := (\mathbf{M}_{*1}^T, \ldots, \mathbf{M}_{*d}^T)^T$ and $\mathbb{S}^{d-1} := \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$ be the d-dimensional unit sphere. For any two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ and any two squared matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, we denote the inner product of $\boldsymbol{a}$ and $\boldsymbol{b}$, $\mathbf{A}$ and $\mathbf{B}$ by $\langle \boldsymbol{a}, \boldsymbol{b}\rangle := \boldsymbol{a}^T\boldsymbol{b}$ and $\langle \mathbf{A}, \mathbf{B}\rangle := \text{Tr}(\mathbf{A}^T\mathbf{B})$. For any matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we denote $\text{diag}(\mathbf{M})$ to be the diagonal matrix with the same diagonal entries as $\mathbf{M}$. For any univariate function f, we denote $f(\mathbf{M}) = [f(\mathbf{M}_{jk})]$ to be a $d \times d$ matrix with f applied

on each entry of $\mathbf{M}$. Let $\mathbf{I}_d$ be the identity matrix in $\mathbb{R}^{d\times d}$. For two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, we denote $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$ if they are identically distributed.

2.2.1 Elliptical Distribution

We briefly overview the elliptical distribution. In the sequel, we say a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is *continuous* if the marginal distributions are all continuous. $\boldsymbol{X}$ possesses density if it is absolutely continuous with respect to the Lebesgue measure.

Definition 2.2.1 (Elliptical distribution)**.** *A random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T$ follows an elliptical distribution if and only if $\boldsymbol{Z}$ has a stochastic representation: $\boldsymbol{Z} \overset{d}{=} \boldsymbol{\mu} + \xi\mathbf{A}\boldsymbol{U}$. Here $\boldsymbol{\mu} \in \mathbb{R}^d$, $q := \operatorname{rank}(\mathbf{A})$, $\mathbf{A} \in \mathbb{R}^{d\times q}$, $\xi \geq 0$ is a random variable independent of $\boldsymbol{U}$, $\boldsymbol{U} \in \mathbb{S}^{q-1}$ is uniformly distributed on the unit sphere in $\mathbb{R}^q$. Letting $\boldsymbol{\Sigma} := \mathbf{A}\mathbf{A}^T$, we denote $\boldsymbol{Z} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. We call $\boldsymbol{\Sigma}$ the scatter matrix.*

In Definition 2.2.1, there can be multiple $\mathbf{A}$'s corresponding to the same $\boldsymbol{\Sigma}$, i.e., there exist $\mathbf{A}_1 \neq \mathbf{A}_2 \in \mathbb{R}^{d\times q}$ such that $\mathbf{A}_1\mathbf{A}_1^T = \mathbf{A}_2\mathbf{A}_2^T = \boldsymbol{\Sigma}$. To make the representation unique, we always parameterize an elliptical distribution by the scatter matrix $\boldsymbol{\Sigma}$ instead of $\mathbf{A}$.

The model family in Definition 2.2.1 is not identifiable. For example, $\boldsymbol{\Sigma}$ is unique only up to a constant scaling, i.e., for some constant $c > 0$, if we define $\xi^* = \xi/c$ and $\mathbf{A}^* = c\mathbf{A}$, then $\xi\mathbf{A}\boldsymbol{U} \overset{d}{=} \xi^*\mathbf{A}^*\boldsymbol{U}$. To make the model identifiable, we require the additional condition that $\max_{1\leq i\leq d} \boldsymbol{\Sigma}_{ii} = 1$. We define $\boldsymbol{\Sigma}^0 := \operatorname{diag}(\boldsymbol{\Sigma})^{-1/2} \cdot \boldsymbol{\Sigma} \cdot \operatorname{diag}(\boldsymbol{\Sigma})^{-1/2}$ to be

the *generalized correlation matrix*.

Table 2.1: Normality test for the stock daily log-return data. This table illustrates the number of 452 stocks rejecting the null hypothesis of normality at the significance level 0.05.

Significance level	**Kolmogorov-Smirnov**	**Shapiro-Wilk**	**Lilliefors**
0.05	428	449	449
0.05/452	269	448	426

2.2.2 Meta-Elliptical Distribution

Real world data are usually nonGaussian and asymmetric. To illustrate the nonGaussianity and asymmetry issues, we consider the stock log return data in S&P 500 index, collected from Yahoo! Finance (`finance.yahoo.com`) from January 1, 2003 to January 1, 2008, including 452 stocks and 1,257 data points. Table 2.1 illustrates the nonGaussian distribution of the stock daily log-return data throughout five years[1]. Here we conduct the three marginal normality tests as in Table 2.1 at the significant level of 0.05. It is clear that at most 24 out of 452 stocks would pass any of three normality tests. Even with Bonferroni correction there are still over half stocks that fail to pass any of these normality tests. Figure 2.1 plots the histograms of three typical stocks, "eBay Inc.", "Macy's Inc.", and "Wells Fargo", in the sectors of information technology, consumer discretionary, and financials respectively. The log-return values are skewed to the left.

[1]For daily closing prices $S_1, \ldots, S_T$, the daily log returns are $\{\log(S_t/S_{t-1}), t = 2, \ldots, T\}$.

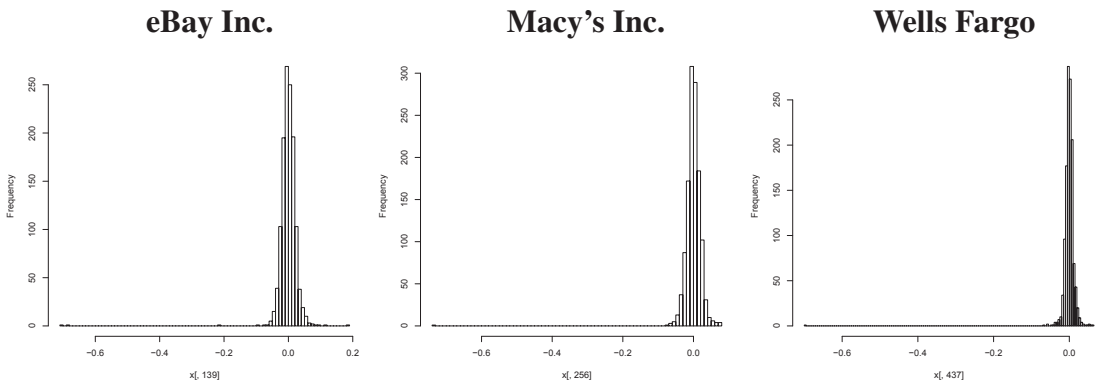


Figure 2.1: Illustration of the asymmetry issue of the log-return stock data.

Though the elliptical distribution family has been widely used to model heavy-tail data (Oja, 2010), it assumes the distribution contours to exhibit ellipsoidal structure. To relax this assumption, Fang et al. (2002) introduced the concept of meta-elliptical distribution under a copula framework. In this section, we introduce the concept of meta-elliptical using a different approach, which extends the family defined in Fang et al. (2002).

First, we define two sets of symmetric matrices:

$$\mathcal{R}_d^+ = \{\mathbf{\Sigma} \in \mathbb{R}^{d\times d} : \mathbf{\Sigma}^T = \mathbf{\Sigma}, \mathrm{diag}(\mathbf{\Sigma}) = \mathbf{I}_d, \mathbf{\Sigma} \succ 0\},$$

$$\mathcal{R}_d = \{\mathbf{\Sigma} \in \mathbb{R}^{d\times d} : \mathbf{\Sigma}^T = \mathbf{\Sigma}, \mathrm{diag}(\mathbf{\Sigma}) = \mathbf{I}_d, \mathbf{\Sigma} \succeq 0\}.$$

The meta-elliptical distribution family is defined as follows:

Definition 2.2.2 (Meta-elliptical distribution)**.** *A continuous random vector* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ *follows a meta-elliptical distribution, denoted by* $\boldsymbol{X} \sim ME_d(\mathbf{\Sigma}^0, \xi; f_1, \ldots, f_d)$*, if there ex-*

ist univariate strictly increasing functions $f_1, \ldots, f_d$ *such that*

$$(f_1(X_1), \ldots, f_d(X_d))^T \sim EC_d(\mathbf{0}, \mathbf{\Sigma}^0, \xi), \quad \text{where } \mathbf{\Sigma}^0 \in \mathcal{R}_d. \tag{2.2.1}$$

Here, $\mathbf{\Sigma}^0$ *is called the latent generalized correlation matrix. When*

$$(f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(\mathbf{0}, \mathbf{\Sigma}^0),$$

$\boldsymbol{X}$ *follows a nonparanormal distribution, denoted by* $\boldsymbol{X} \sim NPN_d(\mathbf{\Sigma}^0; f_1, \ldots, f_d)$.

The meta-elliptical is a strict extension to the nonparanormal defined in Liu et al. (2012a). They both assume after unspecified marginal transformations the data follow certain distributions. However, the nonparanormal exploits a Gaussian base distribution, while the meta-elliptical exploits an elliptical base distribution.

On the other hand, we would like to point out Definition 2.2.2 extends the family originally defined in Fang et al. (2002) in three aspects: (i) The generating variable ξ does not have to be absolutely continuous; (ii) The parameter $\mathbf{\Sigma}^0$ is strictly enlarged from $\mathcal{R}_d^+$ to $\mathcal{R}_d$; and (iii) $\boldsymbol{X}$ does not necessarily possess density. Moreover, even if these two definitions are the same confined in the distribution set with density existing, we can define the meta-elliptical in fundamentally different ways by characterizing the transformation functions instead of characterizing their density functions. By exploiting this new definition, we find several results provided in the later sections can be easier to understand. Hence we also call the meta-elliptical defined in this chapter the transelliptical (transit-elliptical).

The meta-elliptical family is rich and contains many useful distributions, including multivariate Gaussian, rank-deficient Gaussian, multivariate t, logistic, Kotz, symmetric Pearson type-II and type-VII, the nonparanormal, and various other asymmetric distributions such as multivariate asymmetric t distribution (Fang et al., 2002). To illustrate the modeling flexibility of the meta-elliptical family, Figure 2.2 visualizes the density functions of two meta-elliptical distributions.

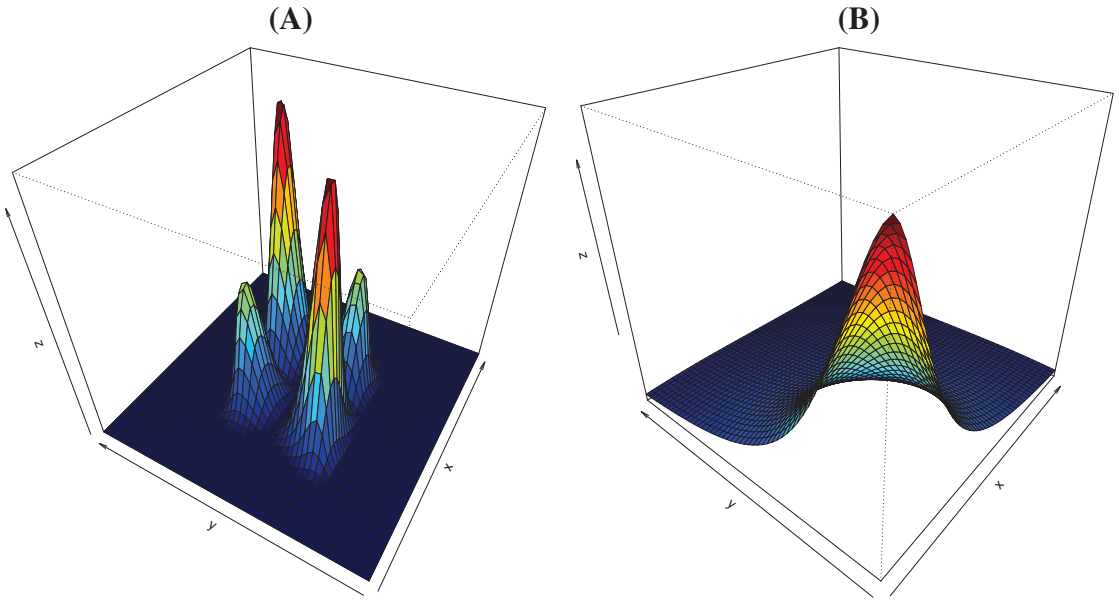


Figure 2.2: Densities of two 2-dimensional meta-elliptical distributions. (A) The component functions have the form $f_1(x) = \text{sign}(x)|x|^2$ and $f_2(x) = x^3$, and after transformation follows a Gaussian distribution. (B) The component functions have the form $f_1(x) = f_2(x) = \log(x)$, and after transformation follows a Cauchy distribution. In both cases the latent generalized correlation matrix has all off-diagonal values to be 0.5.

2.3 Methodology

We propose a new scale-invariant sparse PCA method based on the meta-elliptical distribution family. More specifically, under a meta-elliptical model $\boldsymbol{X} \sim ME_d(\boldsymbol{\Sigma}^0, \xi; f_1, \ldots, f_d)$, the proposed method aims at estimating the leading eigenvector of $\boldsymbol{\Sigma}^0$. Since the diagonal entries of $\boldsymbol{\Sigma}^0$ are all 1, the proposed method is scale-invariant. From Definition 2.2.2, the proposed method is equivalent to conducting scale-invariant sparse PCA on the transformed data $(f_1(X_1), \ldots, f_d(X_d))^T$ which follow an elliptical distribution.

2.3.1 Statistical Model

The statistical model of our proposed method is defined as follows:

Definition 2.3.1. *We consider the following model, denoted by $\mathcal{M}_d(\boldsymbol{\Sigma}^0, \xi, f; \boldsymbol{\theta}_1, s)$, which is defined to be the set of distributions:*

$$\mathcal{M}_d(\boldsymbol{\Sigma}^0, \xi, f; \boldsymbol{\theta}_1, s) := \big\{ \boldsymbol{X} : \boldsymbol{X} \sim ME_d(\boldsymbol{\Sigma}^0, \xi; f_1, \ldots, f_d) \text{ such that } \boldsymbol{\theta}_1, \text{the leading eigenvector of } \boldsymbol{\Sigma}^0, \text{ satisfies } \|\boldsymbol{\theta}_1\|_0 = s. \big\}. \quad (2.3.1)$$

This model allows asymmetric and heavy tail distributions with nontrivial tail dependency. It can be used as a powerful tool for modeling real-world data.

2.3.2 Method

We now provide the proposed method to exploit the model (2.3.1). One of the key components of the proposed rank based method is the Kendall's tau correlation matrix estimator, which will be explained in the next section.

2.3.2.1 Kendall's tau based Correlation Matrix Estimator

The Kendall's tau statistic was introduced by Kendall (1948) for estimating pairwise correlation and has been used for principal component analysis in low dimensions (Croux et al., 2002; Gibbons and Chakraborti, 2003). More specifically, let $\boldsymbol{X} := (X_1, \ldots, X_d)^T$ be a d-dimensional random vector and let $\widetilde{\boldsymbol{X}} := (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ be an independent copy of $\boldsymbol{X}$. The Kendall's tau correlation coefficient between X_j and X_k is defined as

$$\tau(X_j, X_k) := \mathbb{P}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0) - \mathbb{P}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) < 0).$$

The next proposition shows for the meta-elliptical distribution family, we have a one-to-one map between $\boldsymbol{\Sigma}^0_{jk}$ and $\tau(X_j, X_k)$.

Theorem 2.3.2. *Given $\boldsymbol{X} \sim ME_d(\boldsymbol{\Sigma}^0, \xi; f_1, \ldots, f_d)$ meta-elliptically distributed, we have*

$$\boldsymbol{\Sigma}^0_{jk} = \sin\left(\frac{\pi}{2}\tau(X_j, X_k)\right). \tag{2.3.2}$$

Proof. It is obvious that the Kendall's tau statistic is invariant under strictly increasing

transformations to the marginal variables. Moreover, Lindskog et al. (2003) show that the Kendall's tau statistic is invariant to different generating variables ξ's. Combining these two results and Equation (6.6) of Kruskal (1958), we obtain the desired result. □

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ with $\boldsymbol{X}_i := (X_{i1}, \ldots, X_{id})^T$ be n data points of $\boldsymbol{X}$. The sample version Kendall's tau statistic is defined as:

$$\widehat{\tau}_{jk} := \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \operatorname{sign}\left(x_{ij} - x_{i'j}\right) \operatorname{sign}\left(x_{ik} - x_{i'k}\right).$$

It is easy to see that $\widehat{\tau}_{jk}$ is an unbiased estimator of $\tau(X_j, X_k)$. Using $\widehat{\tau}_{jk}$, we define the Kendall's tau correlation matrix as follows:

Definition 2.3.3 (Kendall's tau correlation matrix)**.** *We define the Kendall's tau correlation matrix* $\widehat{\mathbf{R}} = [\widehat{\mathbf{R}}_{jk}]$ *to be a* d *by* d *matrix with element entry to be*

$$\widehat{\mathbf{R}}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right). \tag{2.3.3}$$

2.3.2.2 Rank-based Estimators

Given the model $\mathcal{M}_d(\boldsymbol{\Sigma}^0, \xi, f; \boldsymbol{\theta}_1, s)$, Theorem 2.3.2 provides a natural way to estimate $\boldsymbol{\theta}_1$. In particular, we solve the following optimization problem:

$$\widehat{\boldsymbol{\theta}}^*_{1,k} := \underset{\boldsymbol{v} \in \mathbb{R}^d}{\arg\max}\ \boldsymbol{v}^T \widehat{\mathbf{R}} \boldsymbol{v}, \ \text{ subject to } \ \boldsymbol{v} \in \mathbb{S}^{d-1} \cap \mathbb{B}_0(k), \tag{2.3.4}$$

where $\mathbb{B}_0(k) := \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_0 \leq k\}$, k is a sufficiently large tuning parameter, and $\widehat{\mathbf{R}}$ is the Kendall's tau correlation matrix. Equation (2.3.4) is a combinatorial optimization problem and hard to compute. The corresponding global optimum is denoted by $\widehat{\boldsymbol{\theta}}^*_{1,k}$.

Because the estimator $\widehat{\boldsymbol{\theta}}^*_{1,k}$ is very hard to compute, we consider an alternative way to estimate $\boldsymbol{\theta}_1$ using the truncated power algorithm proposed by Yuan and Zhang (2013). This algorithm yields an estimator $\widetilde{\boldsymbol{\theta}}_{1,k}$. Here $k := \|\widetilde{\boldsymbol{\theta}}_{1,k}\|_0$ is a hypothesized value for s (the number of nonzero elements of $\boldsymbol{\theta}_1$) and can be treated as a tuning parameter.

More specifically, we apply the classical power method, but within each iteration t we project the intermediate vector $\boldsymbol{x}_t$ to the intersection of the d-dimension sphere $\mathbb{S}^{d-1}$ and the ℓ_0 ball with radius $k > 0$. Specifically, we sort the absolute values of the elements of $\boldsymbol{x}_t$ from the highest to the lowest, find the highest k absolute values, truncate all the others to zero, and then normalize the truncated vector such that it lies in $\mathbb{S}^{d-1} \cap \mathbb{B}_0(k)$. To provide the detailed algorithm, we first introduce some additional notation. For any vector $\boldsymbol{v} \in \mathbb{R}^d$ and an index set $J \subset \{1, \ldots, d\}$, we define the truncation function $\text{TRC}(\cdot, \cdot)$ to be

$$\text{TRC}(\boldsymbol{v}, J) := \big(v_1 \cdot I(1 \in J),\ \ldots,\ v_d \cdot I(d \in J)\big)^T, \tag{2.3.5}$$

where $I(\cdot)$ is the indicator function. The truncated power algorithm is presented in Algorithm 1.

The formulation of the truncated power algorithm is nonconvex and the performance of the estimator relies on the selection of the initial vector $\boldsymbol{v}^{(0)}$. In practice, we use the

Algorithm 1 Truncated Power Method

Require: : Kendall's tau matrix $\widehat{\mathbf{R}}$, initial vector $\boldsymbol{v}^{(0)} \in \mathbb{S}^{d-1}$, and k as the tuning parameter.
Ensure: : $\widetilde{\boldsymbol{\theta}}_{1,k} := \boldsymbol{v}^{(\infty)}$
  Set $t = 1$.
  repeat
    Compute $\boldsymbol{x}_t = \widehat{\mathbf{R}}\boldsymbol{v}^{(t-1)}$
    if $\|\boldsymbol{x}_t\|_0 \leq k$ **then**
      $\boldsymbol{v}^{(t)} = \boldsymbol{x}_t / \|\boldsymbol{x}_t\|_2$
    else
      Let A_t be the indices of the elements in $\boldsymbol{x}_t$ with the largest k absolute values
      $\boldsymbol{v}^{(t)} = \mathrm{TRC}(\boldsymbol{x}_t, A_t) / \|\mathrm{TRC}(\boldsymbol{x}_t, A_t)\|_2$
    end if
    $t \leftarrow t + 1$
  until Convergence

estimate obtained from the SPCA algorithm (Zou et al., 2006) as the initial vector. We set the termination criteria to be $\|\boldsymbol{v}^{(t)} - \boldsymbol{v}^{(t-1)}\|_2 \leq 10^{-4}$.

In Section 2.4, we show that, by appropriately setting the initial vector $\boldsymbol{v}^{(0)}$, the algorithm converges and the corresponding estimator $\widetilde{\boldsymbol{\theta}}_{1,k}$ is a consistent estimator of $\boldsymbol{\theta}_1$. In practice, we have found this algorithm always converges on all the synthetic and real-world data.

2.3.3 Estimating the Top m Leading Eigenvectors

We exploit the iterative deflation method to estimate the top m leading eigenvectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ of $\boldsymbol{\Sigma}^0$. This method is proposed by Mackey (2009) and its empirical performance is further evaluated in Yuan and Zhang (2013). In detail, for any positive semidefi-

nite matrix $\mathbf{\Gamma} \in \mathbb{R}^{d \times d}$, its deflation with respect to the vector $\boldsymbol{v} \in \mathbb{R}^d$ is defined as:

$$\mathbf{D}(\mathbf{\Gamma}, \boldsymbol{v}) := (\mathbf{I}_d - \boldsymbol{v}\boldsymbol{v}^T)\mathbf{\Gamma}(\mathbf{I}_d - \boldsymbol{v}\boldsymbol{v}^T).$$

In this way, $\mathbf{D}(\mathbf{\Gamma}, \boldsymbol{v})$ is positive semidefinite, left and right orthogonal to $\boldsymbol{v}$, and symmetric. To estimate $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$, we exploit the following approach: (i) The estimate $\widehat{\boldsymbol{\theta}}_1$ (can be either $\widehat{\boldsymbol{\theta}}^*_{1,k}$ or $\widetilde{\boldsymbol{\theta}}_{1,k}$) of $\boldsymbol{\theta}_1$ is calculated using Equation (2.3.4) or the truncated power method; (ii) Given $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_j$, we estimate $\widehat{\boldsymbol{\theta}}_{j+1}$ by plugging $\mathbf{\Gamma}^{(j+1)} := \mathbf{D}(\mathbf{\Gamma}^{(j)}, \widehat{\boldsymbol{\theta}}_j)$ into Equation (2.3.4) or the truncated power method ($\mathbf{\Gamma}^{(1)} := \mathbf{\Sigma}^0$).

2.4 Theoretical Properties

In this section we provide the theoretical properties of the estimators $\widehat{\boldsymbol{\theta}}^*_{1,k}$ and $\widetilde{\boldsymbol{\theta}}_{1,k}$. In the analysis, we adopt the double asymptotic framework in which the dimension d increases with the sample size n. This framework more realistically reflects the challenges of many high dimensional applications (Bühlmann and van de Geer, 2011).

2.4.1 Latent Generalized Correlation Matrix Estimation

In this section, we focus on estimating the latent generalized correlation matrix $\mathbf{\Sigma}^0$. In the next theorem we prove the rate of convergence $O_P(\sqrt{\log d/n})$ for $|\widehat{\mathbf{R}}_{jk} - \mathbf{\Sigma}^0_{jk}|$ uniformly over all indices j, k. This is an important result, which indicates the Gaussian

parametric rate in estimating the correlation matrix obtained by Bickel and Levina (2008a) can be extended to the meta-elliptical distribution family using the Kendall's tau statistic.

Theorem 2.4.1. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim ME_d(\boldsymbol{\Sigma}^0, \xi; f_1, \ldots, f_d)$ and let $\widehat{\mathbf{R}}$ be defined as in Equation* (2.3.3)*. We have, with probability at least $1 - d^{-5/2}$,*

$$\|\widehat{\mathbf{R}} - \boldsymbol{\Sigma}^0\|_{\max} \leq 3\pi\sqrt{\frac{\log d}{n}}. \tag{2.4.1}$$

Proof. The result follows from Theorem 4.2 in Liu et al. (2012a) but with a slightly different probability bound. □

2.4.2 Leading Eigenvector Estimation

We analyze the estimation errors of the global optimum $\widehat{\boldsymbol{\theta}}_{1,k}^*$ and the estimator $\widetilde{\boldsymbol{\theta}}_{1,k}$ obtained from the truncated power algorithm. We say the model $\mathcal{M}_d(\boldsymbol{\Sigma}^0, \xi, f; \boldsymbol{\theta}_1, s)$ holds if the data are drawn from one probability distribution in $\mathcal{M}_d(\boldsymbol{\Sigma}^0, \xi, f; \boldsymbol{\theta}_1, s)$. The next theorem provides an upper bound on the angle between $\widehat{\boldsymbol{\theta}}_{1,k}^*$ and $\boldsymbol{\theta}_1$.

Theorem 2.4.2. *Let $\widehat{\boldsymbol{\theta}}_{1,k}^*$ be the global optimum to* (2.3.4)*, the model $\mathcal{M}_d(\boldsymbol{\Sigma}^0, \xi, f; \boldsymbol{\theta}_1, s)$ hold, and $k \geq s$. For any two vectors $\boldsymbol{v}_1 \in \mathbb{S}^{d-1}$ and $\boldsymbol{v}_2 \in \mathbb{S}^{d-1}$, let $|\sin\angle(\boldsymbol{v}_1, \boldsymbol{v}_2)| := \sqrt{1 - (\boldsymbol{v}_1^T\boldsymbol{v}_2)^2}$. Then we have, with probability at least $1 - d^{-5/2}$,*

$$|\sin\angle(\widehat{\boldsymbol{\theta}}_{1,k}^*, \boldsymbol{\theta}_1)| \leq \frac{6\pi}{\lambda_1 - \lambda_2} \cdot k\sqrt{\frac{\log d}{n}}, \tag{2.4.2}$$

where $\lambda_j := \Lambda_j(\mathbf{\Sigma}^0)$ *for* $j = 1, 2$.

Remark 2.4.3. *When* s, λ_1, λ_2 *do not scale with* (n, d) *and* $k \geq s$ *is a fixed constant, the rate of convergence in parameter estimation is* $O_P(\sqrt{\log d/n})$*, which is the minimax optimal parametric rate shown in Vu and Lei (2012) under certain model class.*

In the next corollary, we provide a feature selection result for the proposed method. When the selected tuning parameter k is large enough, we show that the support set of $\boldsymbol{\theta}_1$ can be consistently recovered in a fast rate by imposing a constraint on the minimum absolute value of the signal part of $\boldsymbol{\theta}_1$.

Corollary 2.4.4 (Feature selection)**.** *Let* $\widehat{\boldsymbol{\theta}}^*_{1,k}$ *be the global optimum to Equation* (2.3.4)*, the model* $\mathcal{M}_d(\Sigma^0, \xi, f; \theta_1, s)$ *hold, and* $k \geq s$*. Let* $\Theta :=$ *supp*$(\boldsymbol{\theta}_1)$*, and* $\widehat{\Theta}^*_k :=$ *supp*$(\widehat{\boldsymbol{\theta}}^*_{1,k})$*. If we further have* $\min_{j\in\Theta} |\theta_{1j}| \geq \frac{6\sqrt{2}\pi}{\lambda_1 - \lambda_2} \cdot k\sqrt{\frac{\log d}{n}}$*, then* $\mathbb{P}(\Theta \subset \widehat{\Theta}^*_k) \geq 1 - d^{-5/2}$.

In the next theorem, we provide a result on the convergence rate of the estimator $\widetilde{\boldsymbol{\theta}}_{1,k}$ obtained by exploiting the truncated power algorithm. This theorem, coming from Yuan and Zhang (2013), indicates under sufficient conditions $\widetilde{\boldsymbol{\theta}}_{1,k}$ converges to $\boldsymbol{\theta}_1$ in a $s\sqrt{\log d/n}$ rate.

Theorem 2.4.5. *If the model* $\mathcal{M}_d(\mathbf{\Sigma}^0, \xi, f; \boldsymbol{\theta}_1, s)$ *holds, the conditions in Theorem 1 in Yuan and Zhang (2013) hold, and* $k \geq s$*, we have, with probability at least* $1 - d^{-5/2}$*,*

$$|\sin\angle(\widetilde{\boldsymbol{\theta}}_{1,k}, \boldsymbol{\theta}_1)| \leq C \cdot (s + 2k)\sqrt{\frac{\log d}{n}},$$

for some generic constant C *not scaling with* (n, d, s).

The result in Theorem 2.4.5 is a direct consequence of Theorem 1 in Yuan and Zhang (2013) and therefore the proof is omitted. Here we note that, similar as Corollary 2.4.4, it can be shown that under certain conditions, $\text{supp}(\boldsymbol{\theta}_1) \subset \text{supp}(\widetilde{\boldsymbol{\theta}}_{1,k})$ with high probability.

2.4.3 Principal Component Estimation

In this section, we consider estimating the latent principal components of the meta-elliptically distributed data. To estimate the latent principal components instead of the eigenvectors of the latent generalized correlation matrix, one needs to obtain good estimates of the unknown transformation functions $f_1, \ldots, f_d$.

Let $\boldsymbol{X} \sim ME_d(\boldsymbol{\Sigma}^0, \xi; f_1, \ldots, f_d)$ follow a meta-elliptical distribution and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X}$ with $\boldsymbol{X}_i := (X_{i1}, \ldots, X_{id})^T$. Let $\boldsymbol{Z} := (f_1(X_1), \ldots, f_d(X_d))^T$ be the transformed random vector. By definition, $\boldsymbol{Z} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}^0, \xi)$ is elliptically distributed. Let Q_g be the marginal distribution function of $\boldsymbol{Z}$ (We know all the elements of $\boldsymbol{Z}$ share the same marginal distribution functions). If Q_g is known, we can estimate $f_1, \ldots, f_d$ as follows. For $j = 1, \ldots, d$, let $\widehat{F}_j(t; \delta_n)$ be defined as

$$\widehat{F}_j(t;\delta_n) := \begin{cases} \delta_n, & \text{if} \quad t < \delta_n \\ \frac{1}{n}\sum_{i=1}^n I(x_{ij} \le t), & \text{if} \quad \delta_n \le t \le 1-\delta_n \\ 1-\delta_n, & \text{if} \quad x > 1-\delta_n \end{cases}.$$

We define

$$\widehat{f}_j(t;\delta_n) := Q_g^{-1}(\widehat{F}_j(t;\delta_n)) \tag{2.4.3}$$

to be an estimator of f_j. When $Q_g(\cdot) = \Phi(\cdot)$, where $\Phi(\cdot)$ is the distribution function of the standard Gaussian, we have the following theorem, showing $\widehat{f}_j(\cdot;\delta_n)$ converges to $f_j(\cdot)$ uniformly over an expanding interval with high probability.

Theorem 2.4.6 (Han et al. (2013))**.** *Suppose that* $\boldsymbol{X} \sim NPN_d(\boldsymbol{\Sigma}^0; f_1, \ldots, f_d)$ *and for* $j = 1, \ldots, d$*, let* $g_j := f_j^{-1}$ *be the inverse function of* f_j*. For any* $0 < \gamma < 1$*, we define*

$$I_n := \left[g_j\left(-\sqrt{2(1-\gamma)\log n}\right), g_j\left(\sqrt{2(1-\gamma)\log n}\right)\right],$$

then $\sup_{t\in I_n} |\widehat{f}_j(t;(2n)^{-1}) - f_j(t)| = O_P\left(\sqrt{\frac{\log\log n}{n^\gamma}}\right)$*. Here* $\widehat{f}_j(t;\delta_n) := \Phi^{-1}(\widehat{F}_j(t;\delta_n))$.

Using Theorem 2.4.6, we have the following theorem, which shows, under appropriate conditions, we can recover the first principal component of any data point $\boldsymbol{X}$.

Theorem 2.4.7. *For any observation* $\boldsymbol{X} \sim NPN_d(\boldsymbol{\Sigma}; f_1, \ldots, f_d)$*, under the conditions of Theorem 2.4.2, letting*

$$\widehat{f}(\boldsymbol{X}) := \left(\widehat{f}_1(X_1;(2n)^{-1}), \ldots, \widehat{f}_d(X_d;(2n)^{-1})\right)^T \text{ and } f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T,$$

and b be any positive constant such that $(s+k)n^{-b/2} = o(1)$, we have

$$|\widehat{f}(\boldsymbol{X})^T\widehat{\boldsymbol{\theta}}^*_{1,k} - f(\boldsymbol{X})^T\boldsymbol{\theta}^*_1| = O_P\left(\sqrt{(s+k)\cdot\frac{\log\log n}{n^{1-b/2}}} + \frac{k}{\lambda_1-\lambda_2}\sqrt{\frac{(s+k)\log d\log n}{n}}\right),$$

where $\boldsymbol{\theta}^*_1 := \mathrm{sign}(\boldsymbol{\theta}_1^T\widehat{\boldsymbol{\theta}}^*_{1,k})\cdot\boldsymbol{\theta}_1$.

2.5 Experiments

In this section we evaluate the empirical performance of the proposed method on both synthetic and real-world datasets, and compare its performance with the classical sparse PCA and two additional robust sparse PCA procedures. We use the truncated power method proposed by Yuan and Zhang (2013) for parameter estimation. The following four methods are considered:

- Pearson: the classical high dimensional scale-invariant PCA using the Pearson's sample correlation matrix as the input;
- S_n: The sparse PCA using the robust S_n correlation matrix estimator (Rousseeuw and Croux, 1993; Maronna and Zamar, 2002) as the input;
- Q_n: The sparse PCA using the robust Q_n correlation matrix estimator (Rousseeuw and Croux, 1993; Maronna and Zamar, 2002) as the input;
- Kendall: The proposed method using the Kendall's tau correlation matrix as the input.

Here the robust Q_n and S_n correlation matrix estimates are calculated by the R package robustbase (Rousseeuw et al., 2009). We also tried the sparse robust PCA procedure proposed in Croux, Filzmoser, and Fritz. (2013), implemented in the R package pcaPP. However, we found the grid algorithm, which is used in their paper to estimate sparse eigenvectors, has convergence problem when the dimension is high, which makes the obtained estimator perform very bad. Therefore, we did not include this procedure in the draft for comparison.

2.5.1 Numerical Simulations

In the simulation study we sample n data points from a given meta-elliptical distribution. Here we set $d = 100$. We first construct $\mathbf{\Sigma}^0$ using a similar idea as in Yuan and Zhang (2013): First a covariance matrix $\mathbf{\Sigma}$ is synthesized through the eigenvalue decomposition, where the first two eigenvalues are given and the corresponding eigenvectors are pre-specified to be sparse. More specifically, let

$$\mathbf{\Sigma} := \sum_{j=1}^{2} (\omega_j - 1)\boldsymbol{u}_j \boldsymbol{u}_j^T + \mathbf{I}_d, \quad \text{where } \omega_1 = 6, \omega_2 = 3.$$

We set $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ as follows:

$$u_{1j} = \begin{cases} \frac{1}{\sqrt{10}} & 1 \le j \le 10 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad u_{2j} = \begin{cases} \frac{1}{\sqrt{10}} & 11 \le j \le 20 \\ 0 & \text{otherwise} \end{cases}.$$

The latent generalized correlation matrix $\mathbf{\Sigma}^0$ is $\mathbf{\Sigma}^0 = \mathrm{diag}(\mathbf{\Sigma})^{-1/2} \cdot \mathbf{\Sigma} \cdot \mathrm{diag}(\mathbf{\Sigma})^{-1/2}$. We then consider six different schemes to generate the data matrix $\mathbf{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n \times d}$:

Scheme 1: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim N_d(\mathbf{0}, \mathbf{\Sigma}^0)$.

Scheme 2: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim N_d(\mathbf{0}, \mathbf{\Sigma}^0)$, but with 5% entries in each $\boldsymbol{X}_i$ randomly picked up and replaced by -5 or 5.

Scheme 3: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim NPN_d(\mathbf{\Sigma}^0; f_1, \ldots, f_1)$ with $f_1(x) = x^3$.

Scheme 4: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim ME_d(\mathbf{\Sigma}^0, \xi_1; f_0, \ldots, f_0)$ with $f_0(x) = x$ and $\xi_1 \stackrel{d}{=} \sqrt{\kappa}\xi_1^*/\xi_2^*$. Here $\xi_1^* \stackrel{d}{=} \chi_d$ and $\xi_2^* \stackrel{d}{=} \chi_\kappa$ with $\kappa \in \mathbb{Z}^+$. In this setting, $\boldsymbol{X}$ follows a multivariate t distribution with degree of freedom κ (Fang et al., 1990). Here we set $\kappa = 3$.

Scheme 5: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim ME_d(\mathbf{\Sigma}^0, \xi_2; f_0, \ldots, f_0)$ with $\xi_2 \sim F(d, 1)$, i.e., ξ_2 follows an F-distribution with degree of freedom d and 1.

Scheme 6: Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim ME_d(\mathbf{\Sigma}^0, \xi_3; f_0, \ldots, f_0)$ with ξ_3 follows an exponential distribution with the rate parameter 1.

Here Schemes 1 to 3 represent three different versions of Gaussian data: (i) The perfect Gaussian data; (ii) The Gaussian data contaminated by outliers; (iii) The Gaussian data contaminated by marginal transformations. Schemes 4-6 represent three different elliptical distributions, which are all heavy-tailed and belong to the meta-elliptical family.

For $n = 50, 100, 200$, we repeatedly generate the data matrix $\mathbf{X}$ according to Schemes

1 to 6 for 1,000 times. To show the feature selection results for estimating the support set of the leading eigenvector $\boldsymbol{\theta}_1$, Figure 2.3 plots the false positive rates against the true positive rates for the four different estimators under different schemes.

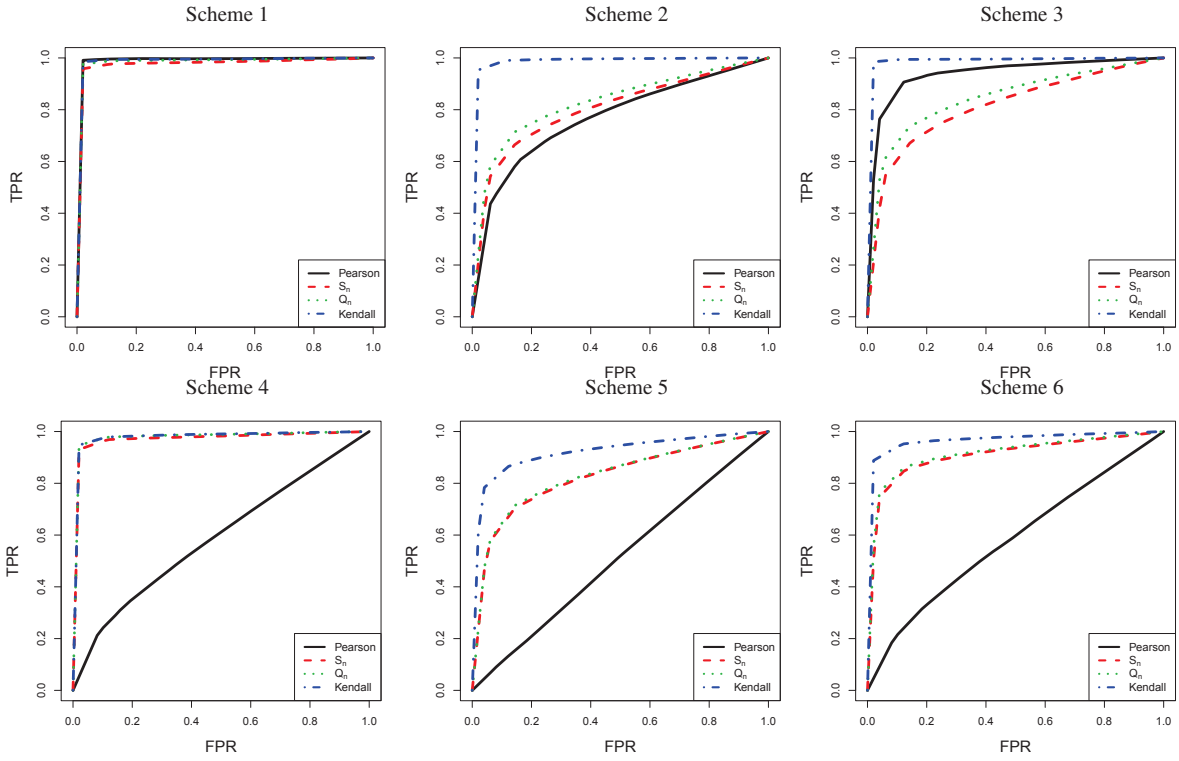


Figure 2.3: ROC curves under Scheme 1 to Scheme 6. Here $n = 100$ and $d = 100$.

To illustrate the parameter estimation performance, we conduct a quantitative comparison of the estimation accuracy of the four competing method. For all methods, we fix the tuning parameter (i.e., the cardinality of the estimate's support set) to be 10. Table 2.2 shows the averaged distances between the estimated leading eigenvector and $\boldsymbol{\theta}_1$, with standard deviations presented in the parentheses. Here the distance between two vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{S}^{d-1}$ is defined as $|\sin \angle(\boldsymbol{v}_1, \boldsymbol{v}_2)|$.

Both Figure 2.3 and Table 2.2 show that when the data are non-Gaussian but follow

Table 2.2: Quantitative comparison on the datasets under the six generating schemes. The averaged distances with standard deviations in parentheses are presented. Here n is changing from 50 to 200 and d is fixed to be 100.

Scheme	n	Pearson	S_n	Q_n	Kendall
Scheme 1	50	0.422(0.555)	0.607(0.473)	0.555(0.259)	0.473(0.266)
	100	0.121(0.158)	0.188(0.140)	0.158(0.110)	0.140(0.201)
	200	0.068(0.071)	0.072(0.072)	0.071(0.018)	0.072(0.024)
Scheme 2	50	0.911(0.878)	0.882(0.631)	0.878(0.105)	0.631(0.131)
	100	0.806(0.715)	0.737(0.264)	0.715(0.169)	0.264(0.213)
	200	0.484(0.354)	0.381(0.093)	0.354(0.222)	0.093(0.246)
Scheme 3	50	0.822(0.907)	0.921(0.473)	0.907(0.154)	0.473(0.101)
	100	0.562(0.700)	0.737(0.140)	0.700(0.214)	0.140(0.202)
	200	0.228(0.356)	0.410(0.072)	0.356(0.156)	0.072(0.255)
Scheme 4	50	0.947(0.679)	0.704(0.678)	0.679(0.095)	0.668(0.227)
	100	0.910(0.247)	0.269(0.248)	0.247(0.157)	0.238(0.239)
	200	0.873(0.079)	0.084(0.084)	0.079(0.232)	0.074(0.063)
Scheme 5	50	0.977(0.911)	0.910(0.854)	0.911(0.028)	0.854(0.102)
	100	0.976(0.718)	0.722(0.532)	0.718(0.028)	0.532(0.214)
	200	0.978(0.297)	0.305(0.147)	0.297(0.029)	0.147(0.244)
Scheme 6	50	0.959(0.848)	0.862(0.771)	0.848(0.060)	0.771(0.143)
	100	0.931(0.548)	0.569(0.373)	0.548(0.108)	0.373(0.250)
	200	0.840(0.156)	0.165(0.103)	0.156(0.223)	0.103(0.170)

a meta-elliptical distribution, Kendall constantly outperforms Pearson in terms of feature selection and parameter estimation. Moreover, when the data are indeed Gaussian distributed, there is no obvious difference between Kendall and Pearson, indicating our proposed rank-based method is a good alternative to the classical scale-invariant sparse PCA under the meta-elliptical model.

We then compare Kendall with S_n and Q_n. In Scheme 1, for the Gaussian data, Kendall slightly outperforms S_n and Q_n. For the data with outliers, S_n and Q_n performs better

than the classical sparse PCA estimates, but are not as robust as Kendall. For different elliptical distributions explored in Schemes 4 to 6, Kendall has the best overall performance compared to S_n and Q_n. The results for the non-elliptically distributed data, as explored in Scheme 3, shows a significant difference between our proposed method and the other two robust sparse PCA approaches. In this case we are interested in, instead of the correlation matrix of the meta-elliptically distributed data, the latent generalized correlation matrix, which S_n and Q_n fail to recover.

2.5.2 Equity Data Analysis

In this section, we investigate the performance of the four competing methods on the equity data explored in Section 2.2.2. The data come from Yahoo! Finance (`finance.yahoo.com`). We collect the daily closing prices for $J = 452$ stocks that are consistently in the S&P 500 index from January 1, 2003 to January 1, 2008. This gives us altogether $T = 1,257$ data points, each data point corresponds to the vector of closing prices on a trading day. Let $St = [St_{t,j}]$ denote the closing price of stock j on day t. We are interested in the log-return data $\mathbf{X} = [\mathbf{X}_{tj}]$ with $\mathbf{X}_{tj} := \log(St_{t,j}/St_{t-1,j})$.

We evaluate the ability of using only a small number of stocks to represent the trend of the whole stock market. To this end, we run the four competing methods on the log-return data $\mathbf{X}$ and obtain the top four leading eigenvectors. Here the iterative deflation method discussed in Section 2.3.3 is exploited with the same tuning parameter k in each deflation step. Let A_k be the support set of the estimated leading eigenvectors by one of the four

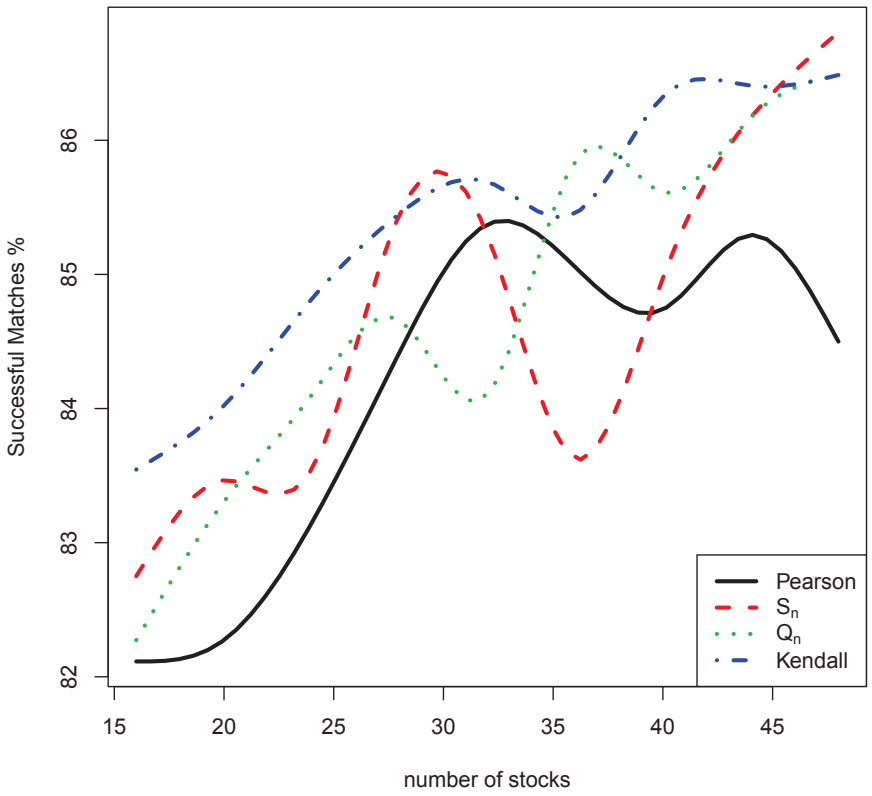


Figure 2.4: Successful matches of the market trend proportions only using the stocks in the support sets of the estimated loading vectors. The horizontal-axis represents the cardinalities of the estimates' support sets; the vertical-axis represents the percentage of successful matches.

methods. We define T_t^W and $T_t^{A_k}$ as

$$T_t^W := I\left(\sum_j St_{t,j} - \sum_j St_{t-1,j} > 0\right), \quad T_t^{A_k} := I\left(\sum_{j\in A_k} St_{t,j} - \sum_{j\in A_k} St_{t-1,j} > 0\right),$$

where $I(\cdot)$ is the indicator function. In this way, we can calculate the proportion of suc-

cessful matches of the market trend using the stocks in A_k as:

$$\rho_{A_k} := \frac{1}{T-1} \sum_{t=2}^{T} I(T_t^W = T_t^{A_k}).$$

We visualize the result by plotting $(\text{card}(A_k), \rho_{A_k})$ in Figure 2.4, which shows that Kendall summarizes the trend of the whole stock market better than the other three methods.

Table 2.3: The ten categories of the stocks with their numbers and abbreviations provided.

Name	Number	Abbreviation
Consumer Discretionary	70	CD
Consumer Staples	35	CS
Energy	37	E
Financials	74	F
Health Care	46	HC
Industrial	59	I
Information Technology	64	IT
Telecommunications Services	6	TS
Materials	29	M
Utilities	32	U

Moreover, we examine the stocks selected by the four competing methods. The 452 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors, including "Consumer Discretionary" (70 stocks), "Consumer Staples" (35 stocks), "Energy" (37 stocks), "Financials" (74 stocks), "Health Care" (46 stocks), "Industrials" (59 stocks), "Information Technology" (64 stocks), "Materials" (29 stocks), "Telecommunications Services" (6 stocks), and "Utilities" (32 stocks). Table 2.3 provides a more detailed description of these ten categories with their numbers and abbreviations provided.

Table 2.4: The categories of the nonzero terms in the top four leading eigenvectors calculated by the four competing methods. The abbreviations are listed in Table 2.3. (Note: 30F means 30 stocks are from the Financials category.)

Method	PC1	PC2	PC3	PC4
Pearson	29F,1I	6CD,5F,8I,1IT,10M	8F,2E,3M,17U	8CD,1F,1I,20IT
S_n	29F,1I	2CD,2F,12I,14M	3I,27IT	3F,27U
Q_n	29F,1I	2CD,2F,12I,1IT,13M	2I,28IT	3F,27U
Kendall	30F	15I, 15M	10CD, 10F,10I	3I, 27IT

We estimate the top four leading eigenvectors using the four competing methods with the same $k = 30$ in each deflation step. The obtained non-zero features' categories are presented in Table 2.4. In general, Kendall has the best ability in grouping the stocks of the same category together. Therefore, Kendall provides a more interpretable result.

2.6 Discussion

We propose a new scale-invariant sparse principal component analysis method for high dimensional meta-elliptical data. Our estimator is semiparametric but achieves a fast rate of convergence in parameter estimation, and is robust to both modeling assumption and data contamination. Therefore, the new estimator can be a good alternative to the classical sparse PCA method.

Although the rank-based Kendall's tau statistic has been exploited for principal component analysis in low dimensions (see, for example, Croux et al. (2002)), our work is funda-

mentally different from the existing literature. The main difference can be elaborated in the following three aspects: (i) We generalize the Kendall's tau statistic to high dimensions, while the current literature only focuses on the low dimension settings; (ii) Our theoretical analysis are fundamentally different from the previous low dimensional analysis, which exploits classical semiparametric theory under which the dimension d is usually fixed; (iii) Most existing methods and theories are built upon the Gaussian or elliptical model, while we consider the meta-elliptical model.

There is another trend in exploiting robust (sparse) PCA (see, for example, Maronna and Zamar (2002) and Croux et al. (2013)). The empirical comparisons conducted in this chapter indicate that, confined in the meta-elliptical family, the proposed rank-based method can be more efficient in parameter estimation and feature selection than these additional robust procedures. Moreover, our proposed method achieves the nearly parametric rate of convergence in parameter estimation, while to the best of our knowledge the performance of these robust sparse PCA procedures in high dimensions is mostly unknown.

Vu and Lei (2012) and Ma (2013) considered sparse principal component analysis and studied the rates of convergence under various modeling and sparsity assumptions. Our method is different from theirs in two aspects: (i) Their analysis relies heavily on the Gaussian or sub-Gaussian assumption, which no longer holds under the meta-elliptical model; (ii) They exploit the Pearson's sample covariance or correlation matrix as the algorithm input, while we advocate the usage of the Kendall's tau correlation matrix in the meta-elliptical model.

Liu et al. (2012a) and Xue and Zou (2012) proposed a procedure called the nonparanormal SKEPTIC, which exploits the nonparanormal family for graph estimation. The nonparanormal SKEPTIC also adopts rank-based methods in high dimensions. Our method is different from theirs in three aspects: (i) We advocate the use of meta-elliptical family, of which the nonparanormal is a subset; (ii) We advocate the use of the Kendall's tau, which is adaptive over the whole meta-elliptical family but instead of the Spearman's rho statistic; (iii) Their focus is on graph estimation, in contrast, this work focuses on principal component analysis. In a preliminary version of this work (Han and Liu, 2014b), they mainly focused on estimating the first leading eigenvector of the latent generalized correlation matrix by directly solving Equation (2.3.4), which is practically intractable. In contrast, we exploit a computationally feasible procedure (truncated power method) for scale-invariant sparse PCA, and provide theoretical guarantee of convergence for this algorithm. Moreover, our method estimates the latent principal components, which are crucial in practical applications, and we provide the theoretical analysis of convergence for the corresponding estimators.

For the principal component estimation algorithm in Section 2.4.3, when Q_g is unknown, we could estimate $f_1, \ldots, f_d$ using the following method:

1. Test whether the original data is elliptically distributed by using some existing techniques (Li et al., 1997; Huffer and Park, 2007; Sakhanenko, 2008). If yes, we set $\widehat{f}_j(t) = (t - \widehat{\mu}_j)/\widehat{\sigma}_j$. Here $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ are the marginal sample mean and standard deviation for the j-th entry.

2. If not, we construct a set of marginal distribution functions:

$$\Pi := \{Q_g : Q_g \text{ is a well defined marginal distribution function}\}.$$

3. For any $Q_g \in \Pi$, we calculate $\widehat{f} = \{\widehat{f}_1, \ldots, \widehat{f}_d\}$ using Equation (2.4.3).

4. We transform the data using $\widehat{f}$.

5. We test whether the transformed data is elliptically distributed by using the techniques exploited in step 1.

We iterate steps 3-5 until we cannot reject the null hypothesis in step 5 for some Q_g. This is a heuristic method whose theoretical justification is left for future investigation. Other future directions include analyzing the robustness property of the method to more noisy and dependent data.

Lastly, we note this chapter follows one main idea throughout the thesis: A semiparametric model coupled with a nonparametric robust method could be an appealing approach in tackling high dimensional complex data. Here we exploit the semiparametric transelliptical (meta-elliptical) model coupled with a nonparametric rank-based method.

Chapter 3

Statistical Analysis of Latent Generalized Correlation Matrix Estimation in Transelliptical Distribution

3.1 Introduction

Covariance and correlation matrices play a central role in multivariate analysis. An efficient estimation of covariance/correlation matrix is a major step in conducting many methods, including principal component analysis (PCA), scale-invariant PCA, graphical model estimation, discriminant analysis, and factor analysis. Large covariance/correlation matrix estimation receives a lot of attention in high dimensional statistics. This is partially because the sample covariance/correlation matrix is an inconsistent estimator where $d/n \not\to 0$ (Here d and n represent the dimensionality and sample size.).

Given n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of a d dimensional random vector $\boldsymbol{X} \in \mathbb{R}^d$ with the population covariance matrix $\boldsymbol{\Omega}$, let $\widehat{\mathbf{S}}$ be the Pearson's sample covariance matrix calculated based on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. For theoretical analysis, we adopt a similar double asymptotic framework as in Bickel and Levina (2008a), where we write d to be the abbreviation of d_n, which changes with n. Under this double asymptotic framework, where both the dimension d and sample size n can increase to infinity, Johnstone (2001), Baik and Silverstein (2006), and Jung and Marron (2009) pointed out settings where, even when $\boldsymbol{X}$ follows a Gaussian distribution with identity covariance matrix, $\widehat{\mathbf{S}}$ is an inconsistent estimator of $\boldsymbol{\Sigma}$ under spectral norm. In other words, letting $\|\cdot\|_2$ denote the spectral norm of a matrix, typically for $(n, d) \to \infty$, we have

$$\|\widehat{\mathbf{S}} - \boldsymbol{\Omega}\|_2 \not\to 0.$$

This observation motivates different versions of sparse covariance/correlation matrix estimation methods. See, for example, banding method (Bickel and Levina, 2008a), tapering method (Cai et al., 2010; Cai and Zhou, 2012), and thresholding method (Bickel and Levina, 2008b). However, although the regularization methods exploited are different, they all use the Pearson's sample covariance/correlation matrix as a pilot estimator, and accordingly the performance of the estimators relies on existence of higher order moments of the data. For example, letting $\|\cdot\|_{\max}$ and $\|\cdot\|_{2,s}$ denote the element-wise supremum norm and restricted spectral norm (detailed definitions provided later), in proving

$$\|\widehat{\mathbf{S}}-\mathbf{\Omega}\|_{\max}=O_P\left(\sqrt{\frac{\log d}{n}}\right)\quad\text{or}\quad\|\widehat{\mathbf{S}}-\mathbf{\Omega}\|_{2,s}=O_P\left(\sqrt{\frac{s\log(d/s)}{n}}\right)\tag{3.1.1}$$

(Here d and s are the abbreviation of d_n and s_n and $O_P(\cdot)$ is defined to represent the stochastic order with regard to n), it is commonly assumed that, for $d=1,2,\ldots$, $\boldsymbol{X}=(X_1,\ldots,X_d)^T$ satisfies the following subgaussian condition:

$$\begin{aligned}&(\text{marginal subgaussian})\quad\mathbb{E}\exp(tX_j)\le\exp\left(\frac{\sigma^2t^2}{2}\right),\quad\text{for all } j\in\{1,\ldots,d\},\\ \text{or }&(\text{multivariate subgaussian})\quad\mathbb{E}\exp(t\boldsymbol{v}^T\boldsymbol{X})\le\exp\left(\frac{\sigma^2t^2}{2}\right),\quad\text{for all }\boldsymbol{v}\in\mathbb{S}^{d-1},\end{aligned}\tag{3.1.2}$$

for some absolute constant $\sigma^2>0$. Here $\mathbb{S}^{d-1}$ is the d-dimensional unit sphere in $\mathbb{R}^d$.

The moment conditions in (3.1.2) are not satisfied for many distributions. To elaborate how strong this condition is, we consider the student's t distribution. Assuming that T

follows a student's t distribution with degree of freedom ν, it is known (Hogg and Craig, 2012) that

$$\mathbb{E}T^{2k} = \infty \quad \text{for } k \geq \nu/2.$$

Recently, Han and Liu (2014b) advocated using the transelliptical distribution for modeling and analyzing complex and noisy data. They exploited a transformed version of the Kendall's tau sample correlation matrix $\widehat{\mathbf{\Sigma}}$ to estimate the latent Pearson's correlation matrix $\mathbf{\Sigma}$. The transelliptical family assumes that, after a set of unknown marginal transformations, the data follow an elliptical distribution. This family is closely related to the elliptical copula and contains many well known distributions, including multivariate Gaussian, rank-deficient Gaussian, multivariate-t, Cauchy, Kotz, logistic, etc.. Under the transelliptical distribution, without any moment constraint, they showed a transformed Kendall's tau sample correlation matrix $\widehat{\mathbf{\Sigma}}$ approximates the latent Pearson's correlation matrix $\mathbf{\Sigma}$ in a parametric rate:

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right), \tag{3.1.3}$$

which attains the minimax rate of convergence.

Although (3.1.3) is inspiring, in terms of theoretical analysis of many multivariate methods, the rates of convergence under spectral norm and restricted spectral norm are more desired. For example, Bickel and Levina (2008b) and Yuan and Zhang (2013) showed the performances of principal component analysis and a computationally tractable sparse PCA

method are determined by the rates of convergence for the plug-in matrix estimators under spectral and restricted spectral norms. A trivial extension of (3.1.3) gives

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O_P\left(d\sqrt{\frac{\log d}{n}}\right) \quad \text{and} \quad \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{2,s} = O_P\left(s\sqrt{\frac{\log d}{n}}\right),$$

which are both not tight compared to the parametric rates (For more details, check Lounici (2013a) and Bunea and Xiao (2014) for results under the spectral norm, and Vu and Lei (2012) for results under the restricted spectral norm).

In this work we push the results in Han and Liu (2014b) forward, providing improved results of the transformed Kendall's tau correlation matrix under both spectral and restricted spectral norms. We consider the statistical properties of the Kendall's tau sample correlation matrix $\widehat{\mathbf{T}}$ in estimating the Kendall's tau correlation matrix $\mathbf{T}$, and the transformed version $\widehat{\mathbf{\Sigma}}$ in estimating $\mathbf{\Sigma}$.

First, we considering estimating the Kendall's tau correlation matrix $\mathbf{T}$ itself. Estimating Kendall's tau is of its self-interest. For example, Embrechts et al. (2003) claimed in many cases in modeling dependence Pearson's correlation coefficient "might prove very misleading" and advocated using the Kendall's tau correlation coefficient as the "perhaps best alternatives to the linear correlation coefficient as a measure of dependence for non-elliptical distributions". In estimating $\mathbf{T}$, we show that, without any condition, for any

continuous random vector $\boldsymbol{X}$,

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_2 = O_P\left(\|\mathbf{T}\|_2\sqrt{\frac{r_e(\mathbf{T})\log d}{n}}\right),$$

where $r_e(\mathbf{T}) := \mathrm{Tr}(\mathbf{T})/\|\mathbf{T}\|_2$ is called effective rank. Moreover, we provide a new term called "sign subgaussian condition", under which we have

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_{2,s} = O_P\left(\|\mathbf{T}\|_2\sqrt{\frac{s\log d}{n}}\right).$$

Secondly, under the transelliptical family, we consider estimating the Pearson's correlation matrix $\mathbf{\Sigma}$ of the latent elliptical distribution using the transformed Kendall's tau sample correlation matrix $\widehat{\mathbf{\Sigma}} = [\sin(\frac{\pi}{2}\widehat{\mathbf{T}}_{jk})]$. Without any moment condition, we show, as long as $\boldsymbol{X}$ belongs to the transelliptical family,

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O_P\left(\|\mathbf{\Sigma}\|_2\left\{\sqrt{\frac{r_e(\mathbf{\Sigma})\log d}{n}} + \frac{r_e(\mathbf{\Sigma})\log d}{n}\right\}\right),$$

which attains the nearly optimal rate of convergence obtained in Lounici (2013a) and Bunea and Xiao (2014). Moreover, provided the sign subgaussian condition is satisfied, we have

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{2,s} = O_P\left(\|\mathbf{\Sigma}\|_2\sqrt{\frac{s\log d}{n}} + \frac{s\log d}{n}\right),$$

which attains the nearly optimal rate of convergence obtained in Vu and Lei (2012).

3.1.1 Discussion with Related Works

Our work is related to a vast literature in large covariance matrix estimation, with different settings of sparsity assumptions (Cai et al., 2010; Cai and Zhou, 2012; Vu and Lei, 2012; Cai et al., 2014b), or without any sparsity assumption (Lounici, 2013a; Bunea and Xiao, 2014). In particular, this work is closely related to Lounici (2013a) and Bunea and Xiao (2014) with regard to the theoretical analysis of the spectral norm convergence, and the work of Vu and Lei (2012) with regard to the theoretical analysis of the restricted spectral norm convergence.

However, there are various new contributions made in this work given the aforementioned results. We emphasize the advantage of rank-based statistics over moment-based statistics. One new message delivered in this work is, via resorting to the rank-based statistics, the statistical efficiency attained by the aforementioned methods under some stringent moment constraints, can also be attained under some more flexible models. Moreover, we believe the technical developments built in this work, including the analysis of U-statistics, the concentration of matrix-value functions, and the verification of the sign subgaussian condition for several particular models, are distinct from the existing literature and of self-interest.

Our work is also closely related to an expanding literature in extending copula models to the high dimensional settings. These include the use of the nonparanormal (Gaussian copula) and the transelliptical (elliptical copula) distribution families. Methodologically, the Spearman's rho is recommended in the analysis of the nonparanormal family for con-

ducting graphical model estimation (Liu et al., 2012a; Xue and Zou, 2012), classification (Han et al., 2013), and PCA (Han and Liu, 2014a). The Kendall's tau is recommended in the analysis of the transelliptical family for conducting graphical model estimation (Liu et al., 2012c) and PCA (Han and Liu, 2014b).

Our work is motivated by the aforementioned results. But, different from the existing ones, we give a more general study on the convergence of the Kendall's tau matrix itself, and provide more insights into the rank-based statistics. We characterize three types of convergence with regard to the Kendal's tau matrix $\widehat{\mathbf{T}}$ and its transformed version $\widehat{\mathbf{\Sigma}}$: The element-wise supremum norm ($\ell_{\max}$), the spectral norm (ℓ_2), and the restricted spectral norm ($\ell_{2,s}$). In comparison, the existing results only exploited the $\ell_{\max}$ convergence result, which we find is not sufficient in showing the statistical efficiency of many rank-based methods. It is also worth noting the new theories developed here with regard to the ℓ_2 and $\ell_{2,s}$ convergence have broad implications. They can be easily applied to the study of factor model, sparse PCA, robust regression, and many other methods, and can lead to more refined statistical analysis.

3.1.2 Notation System

Let $\mathbf{M} = [\mathbf{M}_{ij}] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$. We denote $\boldsymbol{v}_I$ to be the subvector of $\boldsymbol{v}$ whose entries are indexed by a set I. We also denote $\mathbf{M}_{I,J}$ to be the submatrix of $\mathbf{M}$ whose rows are indexed by I and columns are indexed by J. Let $\mathbf{M}_{I*}$ and $\mathbf{M}_{*J}$ be the submatrix of $\mathbf{M}$ with rows indexed by I, and the submatrix of $\mathbf{M}$ with columns indexed

by J. Let $\text{supp}(\boldsymbol{v}) := \{j : v_j \neq 0\}$. For $0 < q < \infty$, we define the ℓ_0, ℓ_q, and ℓ_∞ vector (pseudo-)norms as

$$\|\boldsymbol{v}\|_0 := \text{card}(\text{supp}(\boldsymbol{v})),\ \|\boldsymbol{v}\|_q := (\sum_{i=1}^{d} |v_i|^q)^{1/q},\ \text{and}\ \|\boldsymbol{v}\|_\infty := \max_{1\leq i\leq d} |v_i|.$$

Let $\lambda_j(\mathbf{M})$ be the j-th largest eigenvalue of $\mathbf{M}$ and $\boldsymbol{\Theta}_j(\mathbf{M})$ be a corresponding eigenvector. In particular, we let $\lambda_{\max}(\mathbf{M}) := \lambda_1(\mathbf{M})$. We define $\mathbb{S}^{d-1} := \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$ to be the d-dimensional unit sphere. We define the matrix element-wise supremum norm ($\ell_{\max}$ norm), spectral norm (ℓ_2 norm), and restricted spectral norm ($\ell_{2,s}$ norm) as

$$\|\mathbf{M}\|_{\max} := \max\{|\mathbf{M}_{ij}|\}, \|\mathbf{M}\|_2 := \sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \|\mathbf{M}\boldsymbol{v}\|_2,\ \text{and}\ \|\mathbf{M}\|_{2,s} := \sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}\cap\|\boldsymbol{v}\|_0\leq s} \|\mathbf{M}\boldsymbol{v}\|_2.$$

We define $\text{diag}(\mathbf{M})$ to be a diagonal matrix with $[\text{diag}(\mathbf{M})]_{jj} = \mathbf{M}_{jj}$ for $j = 1, \ldots, d$. We also denote $\text{vec}(\mathbf{M}) := (\mathbf{M}_{*1}^T, \ldots, \mathbf{M}_{*d}^T)^T$. For any two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, we denote $\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \boldsymbol{a}^T\boldsymbol{b}$ and $\text{sign}(\boldsymbol{a}) := (\text{sign}(a_1), \ldots, \text{sign}(a_d))^T$, where $\text{sign}(x) = x/|x|$ with the convention $0/0 = 0$.

3.1.3 Chapter Organization

The rest of this work is organized as follows. In the next section, we briefly overview the transelliptical distribution family and the main concentration results for the transformed Kendall's tau sample correlation matrix proposed by Han and Liu (2014b). In Section 3.3

we analyze the convergence rates of Kendall's tau sample correlation matrix and its transformed version with regard to the spectral norm. In Section 3.4, we analyze the convergence rates of Kendall's tau sample correlation matrix and its transformed version with regard to the restricted spectral norm. The technical proofs of these results are provided in Section 3.5. More discussions and conclusions are provided in Section 3.6.

3.2 Preliminaries and Background Overview

In this section, we briefly review the transelliptical distribution and the corresponding latent generalized correlation matrix estimator proposed by Han and Liu (2014b).

3.2.1 Transelliptical Distribution Family

The concept of transelliptical distribution builds upon the elliptical distribution. Accordingly, we first provide a definition of the elliptical distribution, using the stochastic representation as in Fang et al. (1990). In the sequel, for any two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, we denote $\boldsymbol{X} \stackrel{d}{=} \boldsymbol{Y}$ if they are identically distributed.

Definition 3.2.1 (Fang et al. (1990))**.** *A random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T$ follows an elliptical distribution if and only if $\boldsymbol{Z}$ has a stochastic representation: $\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{\mu} + \xi \mathbf{A}\boldsymbol{U}$. Here $\boldsymbol{\mu} \in \mathbb{R}^d$, $q := \mathrm{rank}(\mathbf{A})$, $\mathbf{A} \in \mathbb{R}^{d\times q}$, $\xi \geq 0$ is a random variable independent of $\boldsymbol{U}$, $\boldsymbol{U} \in \mathbb{S}^{q-1}$ is uniformly distributed on the unit sphere in $\mathbb{R}^q$. In this setting, letting $\boldsymbol{\Sigma} := \mathbf{A}\mathbf{A}^T$, we denote $\boldsymbol{Z} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. Here $\boldsymbol{\Sigma}$ is called the scatter matrix.*

The elliptical family can be viewed as a semiparametric generalization of the Gaussian family, maintaining the symmetric property of the Gaussian distribution but allowing heavy tails and richer structures. Moreover, it is a natural model for many multivariate methods such as principal component analysis (Boente et al., 2012). The transelliptical distribution family further relaxes the symmetric assumption of the elliptical distribution by assuming that, after unspecified strictly increasing marginal transformations, the data are elliptically distributed. A formal definition of the transelliptical distribution is as follows.

Definition 3.2.2 (Han and Liu (2014b))**.** *A random vector* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ *follows a transelliptical distribution, denoted by* $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$*, if there exist univariate strictly increasing functions* $f_1, \ldots, f_d$ *such that*

$$\big(f_1(X_1), \ldots, f_d(X_d)\big)^T \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi), \quad \text{where } \mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_d \text{ and } \mathbb{P}(\xi = 0) = 0.$$

Here $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ *is the* d*-dimensional identity matrix and* $\boldsymbol{\Sigma}$ *is called the latent generalized correlation matrix.*

We note the transelliptical distribution is closely related to the nonparanormal distribution (Liu et al., 2009, 2012a; Xue and Zou, 2012; Han and Liu, 2014a; Han et al., 2013) and meta-elliptical distribution (Fang et al., 2002). The nonparanormal distribution assumes after unspecified strictly increasing marginal transformations, the data are Gaussian distributed. It is easy to see the transelliptical family contains the nonparanormal family. On the other hand, it is subtle to elaborate the difference between the transelliptical and

meta-elliptical. In short, the transelliptical family contains meta-elliptical family. Compared to the meta-elliptical, the transelliptical family does not require the random vectors to have densities and brings new insight into both theoretical analysis and model interpretability. We refer to Liu et al. (2012c) for more detailed discussion on the comparison between the transelliptical family, nonparanormal, and meta-elliptical families.

3.2.2 Latent Generalized Correlation Matrix Estimation

Following Han and Liu (2014b), we are interested in estimating the latent generalized correlation matrix $\mathbf{\Sigma}$, i.e., the correlation matrix of the latent elliptically distributed random vector $f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T$. By treating both the generating variable ξ and the marginal transformation functions $f = \{f_j\}_{j=1}^d$ as nuisance parameters, Han and Liu (2014b) proposed to use a transformed Kendall's tau sample correlation matrix to estimate the latent generalized correlation matrix $\mathbf{\Sigma}$. More specifically, letting $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n independent and identically distributed observations of a random vector $\boldsymbol{X} \in TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$, the Kendall's tau correlation coefficient between the variables X_j and X_k is defined as

$$\widehat{\tau}_{jk} := \frac{2}{n(n-1)} \sum_{i<i'} \mathrm{sign}\big((\boldsymbol{x}_i - \boldsymbol{x}_{i'})_j (\boldsymbol{x}_i - \boldsymbol{x}_{i'})_k\big).$$

Its population quantity can be written as

$$\tau_{jk} := \mathbb{P}\big((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0\big) - \mathbb{P}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) < 0\big), \tag{3.2.1}$$

where $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ is an independent copy of $\boldsymbol{X}$. We denote

$$\mathbf{T} := [\tau_{jk}] \quad \text{and} \quad \widehat{\mathbf{T}} := [\widehat{\tau}_{jk}]$$

to be the Kendall's tau correlation matrix and Kendall's tau sample correlation matrix.

For the transelliptical family, it is known that $\boldsymbol{\Sigma}_{jk} = \sin(\frac{\pi}{2}\tau_{jk})$ (Check, for example, Theorem 3.2 in Han and Liu (2014b)). A latent generalized correlation matrix estimator $\widehat{\boldsymbol{\Sigma}} := [\widehat{\boldsymbol{\Sigma}}_{jk}]$, called the transformed Kendall's tau sample correlation matrix, is accordingly defined by:

$$\widehat{\boldsymbol{\Sigma}}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right). \tag{3.2.2}$$

Han and Liu (2014b) showed, without any moment constraint,

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right),$$

and accordingly by simple algebra we have

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O_P\left(d\sqrt{\frac{\log d}{n}}\right) \quad \text{and} \quad \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{2,s} = O_P\left(s\sqrt{\frac{\log d}{n}}\right). \tag{3.2.3}$$

The rates of convergence in (3.2.3) are far from optimal (Check Lounici (2013a), Bunea and Xiao (2014), and Vu and Lei (2012) for the parametric rates). In the next two sections, we will push the results in Han and Liu (2014b) forward, showing that better rates of convergence can be built in estimating the Kendall's tau correlation matrix and the latent generalized correlation matrix.

3.3 Rate of Convergence under Spectral Norm

In this section we provide the rate of convergence of the Kendall's tau sample correlation matrix $\widehat{\mathbf{T}}$ to $\mathbf{T}$, as well as the transformed Kendall's tau sample correlation matrix $\widehat{\mathbf{\Sigma}}$ to $\mathbf{\Sigma}$, under the spectral norm. The next theorem shows, without any moment constraint or assumption on the data distribution (as long as it is continuous), the rate of convergence of $\widehat{\mathbf{T}}$ to $\mathbf{T}$ under the spectral norm is $\|\mathbf{T}\|_2\sqrt{r_e(\mathbf{T})\log d/n}$, where for any positive semidefinite matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$,

$$r_e(\mathbf{M}) := \frac{\mathrm{Tr}(\mathbf{M})}{\|\mathbf{M}\|_2}$$

is called the effective rank of $\mathbf{M}$ and must be less than or equal to the dimension d. For notational simplicity, in the sequel we assume the sample size n is even. When n is odd,

we can always use $n-1$ data points without affecting the obtained rate of convergence.

Theorem 3.3.1. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n observations of a d dimensional continuous random vector $\boldsymbol{X}$. Then when $r_e(\mathbf{T}) \log d/n \to 0$, for sufficiently large n and any $0 < \alpha < 1$, with probability larger than $1 - 2\alpha$, we have*

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_2 \leq 4\|\mathbf{T}\|_2 \sqrt{\frac{\{r_e(\mathbf{T}) + 1\} \log(d/\alpha)}{3n}}. \tag{3.3.1}$$

Theorem 3.3.1 shows that, when $r_e(\mathbf{T}) \log d/n \to 0$, we have

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_2 = O_P\left(\|\mathbf{T}\|_2 \sqrt{\frac{r_e(\mathbf{T}) \log d}{n}}\right).$$

This rate of convergence is the same parametric rate as obtained in Vershynin (2010), Lounici (2013a), and Bunea and Xiao (2014) when there is not any additional structure.

In the next theorem, we show that, under the modeling assumption that $\boldsymbol{X}$ is transelliptically distributed, which is of particular interest in real applications as shown in Han and Liu (2014b), we have a transformed version of the Kendall's tau sample correlation matrix can estimate the latent generalized correlation matrix in a nearly optimal rate.

Theorem 3.3.2. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n observations of $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$. Let $\widehat{\boldsymbol{\Sigma}}$ be the transformed Kendall's tau sample correlation matrix defined in* (3.2.2)*. We have, when $r_e(\boldsymbol{\Sigma}) \log d/n \to 0$, for n large enough and $0 < \alpha < 1$, with probability larger than*

$1 - 2\alpha - \alpha^2$,

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 \leq \pi^2 \|\mathbf{\Sigma}\|_2 \left(2\sqrt{\frac{\{r_e(\mathbf{\Sigma}) + 1\}\log(d/\alpha)}{3n}} + \frac{r_e(\mathbf{\Sigma})\log(d/\alpha)}{n} \right). \tag{3.3.2}$$

Theorem 3.3.2 indicates that, when $r_e(\mathbf{\Sigma}) \log d/n \to 0$, we have

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O_P\left(\|\mathbf{\Sigma}\|_2 \sqrt{\frac{r_e(\mathbf{\Sigma})\log d}{n}} \right).$$

By the discussion of Theorem 2 in Lounici (2013a), the obtained rate of convergence is minimax optimal up to a logarithmic factor with respect to a suitable parameter space. However, compared to the conditions in Lounici (2013a), and Bunea and Xiao (2014), which require strong multivariate subgaussian modeling assumption on $\boldsymbol{X}$ (which implies the existence of moments of arbitrary order), $\widehat{\mathbf{\Sigma}}$ attains this parametric rate in estimating the latent generalized correlation matrix without any moment constraints.

Remark 3.3.3. *The* $\log d$ *term presented in the rate of convergence of* $\widehat{\mathbf{T}}$ *and* $\widehat{\mathbf{\Sigma}}$ *is an artifact of the proof, and also appears in the statistical analysis of the sample covariance matrix under the subgaussian model (See, for example, Proposition 3 in Lounici (2013a) and Theorem 2.2 in Bunea and Xiao (2014)). If we would like to highlight the role of the effective rank,* $r_e(\mathbf{T})$ *and* $r_e(\mathbf{\Sigma})$*, to our knowledge there is no work that can avoid the* $\log d$ *term. On the other hand, in estimating* $\mathbf{T}$ *using* $\widehat{\mathbf{T}}$*, a* $O_P(\sqrt{d/n})$ *rate of convergence can be attained under the condition of Theorem 3.4.11 provided in the next section. In estimating* $\mathbf{\Sigma}$ *using* $\widehat{\mathbf{\Sigma}}$*, a* $O_P(\sqrt{d/n})$ *rate of convergence is also attainable under the condition of*

Theorem 3.4.11 when $d(\log d)^2 = O(n)$.

3.4 Rate of Convergence under Restricted Spectral Norm

In this section, we analyze the rates of convergence of the Kendall's tau sample correlation matrix and its transformed version under the restricted spectral norm. The main target is to improve the rate $O_P(s\sqrt{\log d/n})$ shown in (3.2.3) to the rate $O_P(\sqrt{s\log(d/s)/n})$. Such a rate has been shown to be minimax optimal under the Gaussian model (via combining Theorem 2.1 and Lemma 3.2.1 in Vu and Lei (2012)). Obtaining such an improved rate is technically challenging since the data could be very heavy-tailed and the transformed Kendall's tau sample correlation matrix has a much more complex structure than the Pearson's covariance/correlation matrix.

In the following we lay out a venue to analyze the statistical efficiency of $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{\Sigma}}$ under the restricted spectral norm. In particular, we characterize a subset of the transelliptical distributions for which $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{\Sigma}}$ can approximate $\mathbf{T}$ and $\mathbf{\Sigma}$ in an improved rate. More specifically, we provide a "sign subgaussian" condition which is sufficient for $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{\Sigma}}$ to attain the nearly optimal rate. This condition is related to the subgaussian assumption in Vu and Lei (2012), Lounici (2013a), and Bunea and Xiao (2014) (see Assumption 2.2 in Vu and Lei (2012), for example). Before proceeding to the formal definition of this condition, we first define an operator $\psi : \mathbb{R} \to \mathbb{R}$ as follows:

Definition 3.4.1. *For any random variable $Y \in \mathbb{R}$, the operator $\psi : \mathbb{R} \to \mathbb{R}$ is defined as*

$$\psi(Y; \alpha, t_0) := \inf \left\{ c > 0 : \mathbb{E} \exp \left\{ t(Y^\alpha - \mathbb{E}Y^\alpha) \right\} \leq \exp(ct^2), \text{ for } |t| < t_0 \right\}. \quad (3.4.1)$$

The operator $\psi(\cdot)$ can be used to quantify the tail behaviors of random variables. We recall that a zero-mean random variable $X \in \mathbb{R}$ is said to be subgaussian if there exists a constant c such that $\mathbb{E} \exp(tX) \leq \exp(ct^2)$ for all $t \in \mathbb{R}$. A zero-mean random variable $Y \in \mathbb{R}$ with $\psi(Y; 1, \infty)$ bounded is well known to be subgaussian, which implies a tail probability

$$\mathbb{P}(|Y - \mathbb{E}Y| > t) < 2 \exp(-t^2/(4c)),$$

where c is the constant defined in Equation (3.4.1). Moreover, $\psi(Y; \alpha, t_0)$ is related to the Orlicz ψ_2-norm. A formal definition of the Orlicz norm is provided as follows.

Definition 3.4.2. *For any random variable $Y \in \mathbb{R}$, its Orlicz ψ_2-norm is defined as*

$$\|Y\|_{\psi_2} := \inf \left\{ c > 0 : \mathbb{E} \exp \left(|Y/c|^2 \right) \leq 2 \right\}.$$

It is well known that a random variable Y has $\psi(Y; 1, \infty)$ to be bounded if and only if $\|Y\|_{\psi_2}$ in Definition 3.4.2 is bounded (van de Geer and Lederer, 2013).

Another relevant norm to $\psi(\cdot)$ is the subgaussian norm $\|\cdot\|_{\phi_2}$ used in, for example, Vershynin (2010). A former definition of the subgaussian norm is as follows.

Definition 3.4.3. *For any random variable $X \in \mathbb{R}$, its subgaussian norm is defined as*

$$\|X\|_{\phi_2} := \sup_{k \geq 1} k^{-1/2} \left(\mathbb{E}|X|^k\right)^{1/k}.$$

The subgaussian norm is also highly related to the subgaussian random variables. In particular, we have if $\mathbb{E}X = 0$, then $\mathbb{E}\exp(tX) \leq \exp(Ct^2\|X\|_{\phi_2}^2)$.

Using the operator $\psi(\cdot)$, we now proceed to define the sign subgaussian condition. For mathematical rigorousness, the formal definition is posed on $\{\mathcal{F}^d, d = 1, 2, \ldots\}$, where $\mathcal{F}^d$ represents a set of probability measures on $\mathbb{R}^d$. Here for any vector $\boldsymbol{v} = (v_1, \ldots, v_d) \in \mathbb{R}^d$, we remind that $\mathrm{sign}(\boldsymbol{v}) := (\mathrm{sign}(v_1), \ldots, \mathrm{sign}(v_d))^T$. In the following, a random vector $\boldsymbol{X}$ is said to be in a set of probability measures $\mathcal{F}'$ if its distribution is in $\mathcal{F}'$.

Definition 3.4.4 (Sign subgaussian condition)**.** *For $d = 1, 2, \ldots$, let $\mathcal{F}^d$ be a set of probability measures on $\mathbb{R}^d$ where infinitely many sets $\mathcal{F}^d$ are non-empty and $\mathcal{F} := \cup_{d=1}^{\infty}\mathcal{F}^d$. $\mathcal{F}$ is said to satisfy the sign subgaussian condition if and only if for any $\boldsymbol{X}$ in $\mathcal{F}$, we have*

$$\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \psi\left(\left\langle \mathrm{sign}(\boldsymbol{X} - \widetilde{\boldsymbol{X}}), \boldsymbol{v} \right\rangle; 2, t_0\right) \leq K\|\mathbf{T}\|_2^2, \tag{3.4.2}$$

where $\widetilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$, K is an absolute constant, and t_0 is another absolute positive number such that $t_0\|\mathbf{T}\|_2$ is lower bounded by an absolute positive constant. We remind that here $\mathbf{T}$ can be written as

$$\mathbf{T} := \mathbb{E}\mathrm{sign}(\boldsymbol{X} - \widetilde{\boldsymbol{X}}) \cdot (\mathrm{sign}(\boldsymbol{X} - \widetilde{\boldsymbol{X}}))^T.$$

To gain more insights about the sign subgaussian condition, we point out two sets of probability measures of interest satisfying the sign subgaussian condition.

Proposition 3.4.5. *Suppose the set of probability measures $\mathcal{F}$ satisfies that for any random vector $\boldsymbol{X}$ in $\mathcal{F}$ and $\widetilde{\boldsymbol{X}}$ being an independent copy of $\boldsymbol{X}$, we have*

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \left\| \left\langle \operatorname{sign}(\boldsymbol{X}-\widetilde{\boldsymbol{X}}), \boldsymbol{v} \right\rangle^2 - \boldsymbol{v}^T\mathbf{T}\boldsymbol{v} \right\|_{\psi_2} \leq L_1\|\mathbf{T}\|_2, \tag{3.4.3}$$

where L_1 is a fixed constant. Then $\mathcal{F}$ satisfies the sign subgaussian condition by setting $t_0 = \infty$ and $K = 5L_1^2/2$ in Equation (3.4.2)*.*

Proposition 3.4.6. *Suppose the set of probability measure $\mathcal{F}$ satisfies that for any random vector $\boldsymbol{X}$ in $\mathcal{F}$ and $\widetilde{\boldsymbol{X}}$ being an independent copy of $\boldsymbol{X}$, we have there exists an absolute constant L_2 such that*

$$\|\boldsymbol{v}^T\operatorname{sign}(\boldsymbol{X}-\widetilde{\boldsymbol{X}})\|_{\phi_2}^2 \leq \frac{L_2\|\mathbf{T}\|_2}{2} \quad \text{for all } \boldsymbol{v}\in\mathbb{S}^{d-1}. \tag{3.4.4}$$

Then $\mathcal{F}$ satisfies the sign subgaussian condition with $t_0 = c\|\mathbf{T}\|_2^{-1}$ and $K = C$ in Equation (3.4.2)*, where c and C are two fixed absolute constants.*

In the following, for clarity of presentation, we abuse notation a little and write $\boldsymbol{X}$ satisfies the sign subgaussian condition if there exists a set of probability measures $\mathcal{F}$ satisfying the sign subgaussian condition such that for $d = 1, 2, \ldots$, $\boldsymbol{X} \in \mathbb{R}^d$ is in $\mathcal{F}$.

Proposition 3.4.6 builds a bridge between the sign subgaussian condition and Assump-

tion 1 in Bunea and Xiao (2014) and Lounici (2013a). More specifically, saying $\boldsymbol{X}$ satisfies Equation (3.4.4) is equivalent to saying $\mathrm{sign}(\boldsymbol{X} - \widetilde{\boldsymbol{X}})$ satisfies the multivariate subgaussian condition defined in Bunea and Xiao (2014). Therefore, Proposition 3.4.6 can be treated as an explanation of why we call the condition in Equation (3.4.2) "sign subgaussian". However, by Lemma 5.14 in Vershynin (2010), the sign subgaussian condition is weaker than that of Equation (3.4.4), i.e., a set of probability measures satisfying the sign subgaussian condition does not necessarily satisfy the condition in Proposition 3.4.6.

The sign subgaussian condition is intuitive due to its relation to the Orlicz and subgaussian norms. However, it is extremely difficult to verify whether a given set of distributions satisfies this condition. The main difficulty arises because we must sharply characterize the tail behavior of the summation of a sequence of possibly correlated discrete Bernoulli random variables, which is much harder than analyzing the summation of Gaussian random variables as usually done in the literature.

In the following we provide several examples of sets of distributions satisfying the sign subgaussian condition. The next theorem shows the transelliptically distributed random vector $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ such that $\boldsymbol{\Sigma} = \mathbf{I}_d$ (i.e., the underlying is a spherical distribution) for $d = 1, 2, \ldots$ satisfies the sign subgaussian condition.

Theorem 3.4.7. *Suppose that, for $d = 1, 2, \ldots$, $\boldsymbol{X} \sim TE_d(\mathbf{I}_d, \xi; f_1, \ldots, f_d)$ is transellipticallly distributed with a latent spherical distribution. Then $\boldsymbol{X}$ satisfies the sign subgaussian condition.*

In the next theorem, we provide a stronger version of Theorem 3.4.7. We call a square

matrix compound symmetric if the off-diagonal values of the matrix are equal. The next theorem shows the transelliptically distributed $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}.\xi; f_1, \ldots, f_d)$, with $\boldsymbol{\Sigma}$ a compound symmetric matrix, satisfies Equation (3.4.4), and therefore satisfies the sign subgaussian condition.

Theorem 3.4.8. *Suppose that for $d = 1, 2, \ldots$, $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ is transellipticallly distributed such that $\boldsymbol{\Sigma}$ is a compound symmetric matrix (i.e., $\boldsymbol{\Sigma}_{jk} = \rho$ for all $j \neq k$). Then if $0 \leq \rho := \boldsymbol{\Sigma}_{12} \leq C_0 < 1$ for some absolute positive constant C_0, we have that $\boldsymbol{X}$ satisfies the sign subgaussian condition.*

Although Theorem 3.4.7 can be directly proved using the result in Theorem 3.4.8, the proof of Theorem 3.4.7 contains utterly different techniques which are more transparent and illustrate the main challenges of analyzing binary sequences even in the uncorrelated setting. Therefore, we still list this theorem separately. Theorem 3.4.8 leads to the following corollary, which characterizes a subfamily of the transelliptical distributions satisfying the sign subgaussian condition.

Corollary 3.4.9. *Suppose that for $d = 1, 2, \ldots$, $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$ is transellipticallly distributed with $\boldsymbol{\Sigma}$ a block diagonal compound symmetric matrix, i.e.,*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \ddots & \cdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_q \end{pmatrix}, \tag{3.4.5}$$

where $\mathbf{\Sigma}_k \in \mathbb{R}^{d_k \times d_k}$ *for* $k = 1, \ldots, q$ *is compound symmetric matrix with* $\rho_k := [\mathbf{\Sigma}_k]_{12} \geq 0$. *We have, if* q *is upper bounded by an absolute positive constant and* $0 \leq \rho_k \leq C_1 < 1$ *for some absolute positive constant* C_1, $\boldsymbol{X}$ *satisfies the sign subgaussian condition.*

We call the matrix in the form of Equation (3.4.5) block diagonal compound symmetric matrix. Corollary 3.4.9 implies transelliptically distributed random vectors with a latent block diagonal compound symmetric latent generalized correlation matrix satisfy the sign subgaussian condition.

Remark 3.4.10. *The subgaussian condition is an artifact of the proof. Right now, we are not aware of any transelliptical distribution not satisfying this condition. More investigation on the necessity of this condition is challenging due to the discontinuity issue of the sign transformation and will be left for future investigation.*

Using the sign subgaussian condition, we have the following main result, which shows as long as the sign subgaussian condition holds, improved rates of convergence for both $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{\Sigma}}$ under the restricted spectral norm can be attained.

Theorem 3.4.11. *For* $d = 1, 2, \ldots$, *let* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *be* n *observations of* $\boldsymbol{X} \in \mathbb{R}^d$, *for which the sign subgaussian condition holds. We have, when* $s\log(d/s)/n \to 0$,*with probability larger than* $1 - 2\alpha$,

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_{2,s} \leq 4(2K)^{1/2}\|\mathbf{T}\|_2\sqrt{\frac{s(3 + \log(d/s)) + \log(1/\alpha)}{n}}. \tag{3.4.6}$$

Moreover, when we further have $\boldsymbol{X} \sim TE_d(\mathbf{\Sigma}, \xi; f_1, \ldots, f_d)$, *with probability larger* $1 -$

$2\alpha - \alpha^2$,

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{2,s} \leq \pi^2 \left(2(2K)^{1/2} \|\mathbf{\Sigma}\|_2 \sqrt{\frac{s(3 + \log(d/s)) + \log(1/\alpha)}{n}} + \frac{s \log(d/\alpha)}{n} \right). \tag{3.4.7}$$

The results presented in Theorem 3.4.11 show under various settings the rate of convergence for $\widehat{\mathbf{\Sigma}}$ under the restricted spectral norm is $O_P(\sqrt{s \log(d/s)/n})$, which is the parametric and minimax optimal rate shown in Vu and Lei (2012) within the Gaussian family. However, the Kendall's tau sample correlation matrix and its transformed version attains this rate with all moment constraints waived.

3.5 Discussion

This work considers robust estimation of the correlation matrix using the rank-based correlation coefficient estimator Kendall's tau and its transformed version. We showed the Kendall's tau is an very robust estimator in high dimensions, because it can achieve the parametric rate of convergence under various norms without any assumption on the data distribution, and in particular, without assuming any moment constraints. We further consider the transelliptical family proposed in Han and Liu (2014b), showing a transformed version of the Kendall's tau attains the parametric rate in estimating the latent Pearson's correlation matrix without assuming any moment constraints. Moreover, unlike the Gaussian case, the theoretical analysis performed here motivates new understandings on rank-

based estimators as well as new proof techniques. These new understandings and proof techniques are of interest in their own right.

Han and Liu (2013b) studied the performance of the latent generalized correlation matrix estimator on dependent data under some mixing conditions and proved that $\widehat{\mathbf{\Sigma}}$ can attain a $s\sqrt{\log d/(n\gamma)}$ rate of convergence under the restricted spectral norm, where $\gamma \leq 1$ reflects the impact of non-independence on the estimation accuracy. It is also interesting to consider extending the results in this work to dependent data under similar mixing conditions, and see whether a similar $\sqrt{s \log d/(n\gamma')}$ rate of convergence can be attained. However, it is much more challenging to obtain such results in dependent data. The current theoretical analysis based on U-statistics is still not sufficient to achieve this goal.

A problem closely related to the leading eigenvector estimation is principal component detection, which is initiated in the work of Berthet and Rigollet (2012, 2013). It is interesting to study this problem under the transelliptical family. It is worthy pointing out Theorems 3.3.2 and 3.4.11 in this work can be exploited in measuring the statistical performance of the corresponding detection of sparse principal components.

Lastly, we note this chapter consists of theoretical results sharpening the ones obtained in the last chapter. In particular, we show the rank-based methods are indeed minimax optimal for inferring the finite-dimensional parameters in a semiparametric transelliptical model. This further supports our main idea in this thesis: We advocate using semiparametric models combined with nonparametric methods.

Chapter 4

ECA: Elliptical Component Analysis in non-Gaussian Distributions

4.1 Introduction

This chapter considers estimating the leading eigenvectors of the covariance matrix for high dimensional, heavy-tailed data. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n data points of a random vector $\boldsymbol{X} \in \mathbb{R}^d$. We denote $\boldsymbol{\Sigma}$ to be the covariance matrix of $\boldsymbol{X}$, and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ to be its top m leading eigenvectors. We want to find $\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_m$ that can estimate $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ accurately. In this chapter, we assume the covariance matrix of $\boldsymbol{X}$ exists.

The above problem is cloesely related to principal component analysis (PCA). In this chapter we are interested in the high dimensional settings and consider the double asymptotic framework where the dimension d is allowed to increase with the sample size n. Under this framework, the performance of PCA, using the leading eigenvectors of the Pearson's sample covariance matrix, has been studied for subgaussian data. In particular, for any matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, letting $\mathrm{Tr}(\mathbf{M})$ and $\sigma_i(\mathbf{M})$ be the trace and i-th largest singular value of $\mathbf{M}$, Lounici (2013a) showed PCA is consistent when $r^*(\boldsymbol{\Sigma}) := \mathrm{Tr}(\boldsymbol{\Sigma})/\sigma_1(\boldsymbol{\Sigma})$ satisfies $r^*(\boldsymbol{\Sigma})/n \to 0$. $r^*(\boldsymbol{\Sigma})$ is referred to as the effective rank of $\boldsymbol{\Sigma}$ in the literature (Vershynin, 2010; Lounici, 2013a).

When $r^*(\boldsymbol{\Sigma})/n \not\to 0$, PCA might not produce a consistent estimator. The inconsistency phenomenon of PCA under the double asymptotic framework has been pointed out by Johnstone and Lu (2009). In particular, they showed the angle between the PCA estimator and $\boldsymbol{u}_1$ may not converge to 0 if $d/n \to c$ for some constant $c > 0$. To avoid this curse of dimensionality, certain types of sparsity assumptions are needed. For example, in estimating the leading eigenvector $\boldsymbol{u}_1 := (u_{11}, \ldots, u_{1d})^T$, we may assume that $\boldsymbol{u}_1$ is

sparse, i.e., $s := \text{card}(\{j : u_{1j} \neq 0\}) \ll n$. We call the setting that $\boldsymbol{u}_1$ is sparse the "sparse setting", and the setting that $\boldsymbol{u}_1$ is not necessarily sparse the "non-sparse setting".

Under the sparse setting, different variants of sparse PCA methods have been proposed. For example, d'Aspremont et al. (2007) proposed to formulate a convex semidefinite program for calculating the sparse leading eigenvectors. Jolliffe et al. (2003) and Zou et al. (2006) connected PCA to regression and proposed to use lasso-type estimators in parameter estimation. Shen and Huang (2008) and Witten et al. (2009) connected PCA to singular vector decomposition (SVD) and proposed iterative algorithms for estimating the left and right singular vectors. Journée et al. (2010) and Zhang and El Ghaoui (2011) proposed to greedily search the principal submatrices of the covariance matrix. Recently, Ma (2013) and Yuan and Zhang (2013) proposed to use modified versions of the power method to estimate eigenvectors and principal subspaces.

Theoretical properties of these methods have been analyzed under both Gaussian and subgaussian assumptions. On one hand, in terms of computationally efficient methods, under the spike covariance Gaussian model, Amini and Wainwright (2009) showed the consistency in parameter estimation and model selection for sparse PCA computed via the semidefinite program proposed in d'Aspremont et al. (2007) and Ma (2013) justified the use of a modified iterative thresholding method in estimating principal subspaces. Very recently, via exploiting a convex program using the Fantope projection (Overton and Womersley, 1992; Dattorro, 2005), Vu et al. (2013) showed that there exist computationally efficient estimators that attain a rate of convergence $O_P(s\sqrt{\log d/n})$ for general covariance

matrices. In Section 4.5.1 we will discuss the Fantope projection in more details.

On the other hand, there exists another line of research focusing on studying sparse PCA conducted via combinatoric programs. For example, Vu and Lei (2012), Lounici (2013b), and Vu and Lei (2013) studied leading eigenvector and principal subspace estimation problems by exhaustively searching over all submatrices. They showed the optimal $O_P(\sqrt{s\log(d/s)/n})$ rate of convergence can be attained using this computationally expensive approach. Such a global search was also studied in Cai et al. (2014b), where they established the upper and lower bounds in both covariance matrix and principal subspace estimations. Barriers between the aforementioned statistically efficient method and computationally efficient methods in sparse PCA was pointed out by Berthet and Rigollet (2012) using the principal component detection problem. Such barriers were also studied in Ma and Wu (2013).

One limitation for the PCA and sparse PCA theories is that they rely heavily on the Gaussian or subgaussian assumption. If the Gaussian assumption is correct, accurate estimation can be expected, otherwise, the obtained result may be misleading. To relax the Gaussian assumption, Han and Liu (2014b) generalized the Gaussian to the semiparametric transelliptical family (called the "meta-elliptical" in their paper) for modeling the data. The transelliptical family assumes that, after unspecified increasing marginal transformations, the data are elliptically distributed. By resorting to the marginal Kendall's tau statistic, Han and Liu (2014b) proposed a semiparametric alternative to scale-invariant PCA, named transelliptical component analysis (TCA), for estimating the leading eigenvector of

the latent generalized correlation matrix $\mathbf{\Sigma}^0$. In a follow-up work, Han and Liu (2013a) showed: (i) Under the non-sparse setting, TCA attains the $O_P(\sqrt{r^*(\mathbf{\Sigma}^0)\log d/n})$ rate of convergence in parameter estimation, which is the same rate of convergence for PCA under the subgaussian assumption (Lounici, 2013a; Bunea and Xiao, 2014); (ii) Under the sparse setting, sparse TCA, formulated as a combinatoric program, can attain the optimal $O_P(\sqrt{s\log(d/s)/n})$ rate of convergence under the "sign subgaussian" condition. More recently, Vu et al. (2013) showed, sparse TCA, via the Fantope projection, can attain the $O_P(s\sqrt{\log d/n})$ rate of convergence.

Despite of all these efforts, there are two remaining problems for the aforementioned works exploiting the marginal Kendall's tau statistic. First, using marginal ranks, they can only estimate the leading eigenvectors of the correlation matrix but not the covariance matrix. Secondly, the sign subgaussian condition is not easy to verify.

In this chapter, we show, under the elliptical model, the optimal $O_P(\sqrt{s\log(d/s)/n})$ rate of convergence in estimating the leading eigenvector of $\mathbf{\Sigma}$ can be attained without the need of sign subgaussian condition. In particular, we present an alternative procedure, named elliptical component analysis (ECA), to directly estimate the eigenvectors of $\mathbf{\Sigma}$ and treat the corresponding eigenvalues as nuisance parameters. ECA exploits multivariate Kendall's tau for estimating the eigenspace of $\mathbf{\Sigma}$. When the target parameter is sparse, the corresponding ECA procedure is specified to be called sparse ECA.

We show that (sparse) ECA has the following properties:

1. Under the non-sparse setting, ECA attains the efficient $O_P(\sqrt{r^*(\mathbf{\Sigma})\log d/n})$ rate of

convergence;

2. Under the sparse setting, sparse ECA, via a combinatoric program, attains the minimax optimal $O_P(\sqrt{s\log(d/s)/n})$ rate of convergence;

3. Under the sparse setting, sparse ECA, via a computationally efficient program which combines the Fantope projection (Vu et al., 2013) and truncated power algorithm (Yuan and Zhang, 2013), attains the optimal $O_P(\sqrt{s\log(d/s)/n})$ rate of convergence under a suboptimal scaling ($s^2 \log d/n \to 0$).

We compare (sparse) PCA, (sparse) TCA, and (sparse) ECA in Table 4.1.

4.1.1 Related Works

The multivariate Kendall's tau statistic is first introduced in Choi and Marden (1998) for testing independence and is further used in estimating low-dimensional covariance matrices (Visuri et al., 2000; Oja, 2010) and principal components (Marden, 1999; Croux et al., 2002; Jackson and Chen, 2004). In particular, Marden (1999) showed the population multivariate Kendall's tau, $\mathbf{K}$, shares the same eigenspace as the covariance matrix $\mathbf{\Sigma}$. Croux et al. (2002) illustrated the asymptotical efficiency of ECA compared to PCA for the Gaussian data when $d = 2$ and 3. Taskinen et al. (2012) characterized the robustness and efficiency properties of ECA in low dimensions

Table 4.1: The illustration of the results in (sparse) PCA, (sparse) TCA, and (sparse) ECA for the leading eigenvector estimation. Similar results also hold for principal subspace estimation. Here $\mathbf{\Sigma}$ is the covariance matrix, $\mathbf{\Sigma}^0$ is the latent generalized correlation matrix, $r^*(\mathrm{M}) := \mathrm{Tr}(\mathrm{M})/\sigma_1(\mathrm{M})$ represents the effective rank of M, "r.c." stands for "rate of convergence", "n-s setting 1" stands for the "non-sparse setting" and the estimation procedure is conducted via a combinatoric program, "n-s setting 2" stands for the "non-sparse setting" and the estimation procedure is conducted via combining the Fantope projection (Vu et al., 2013) and the truncated power method (Yuan and Zhang, 2013).

	(sparse) PCA	(sparse) TCA	(sparse) ECA
working model:	subgaussian family	transelliptical family	elliptical family
parameter of interest:	eigenvectors of $\mathbf{\Sigma}$	eigenvectors of $\mathbf{\Sigma}^0$	eigenvectors of $\mathbf{\Sigma}$
input statistics:	Pearson's covariance matrix	Kendall's tau	multivariate Kendall's tau
sparse setting (r.c.):	$\sqrt{r^*(\mathbf{\Sigma})\log d/n}$	$\sqrt{r^*(\mathbf{\Sigma}^0)\log d/n}$	$\sqrt{r^*(\mathbf{\Sigma})\log d/n}$
n-s setting 1 (r.c):	$\sqrt{s\log(d/s)/n}$	$s\sqrt{\log d/n}$ (general), $\sqrt{s\log(d/s)/n}$ (sign subgaussian)	$\sqrt{s\log(d/s)/n}$,
n-s setting 2 (r,c):	$\sqrt{s\log(d/s)/n}$ given $s^2\log d/n \to 0$	$s\sqrt{\log d/n}$	$\sqrt{s\log(d/s)/n}$ given $s^2\log d/n \to 0$

Some related methods using multivariate rank-based statistics have also been discussed in Tyler (1982), Tyler (1987), Taskinen et al. (2003), Oja and Randles (2004), Oja and Paindaveine (2005), Oja et al. (2006), and Sirkiä et al. (2007). Theoretical analysis in low dimensions was provided in Hallin and Paindaveine (2002b,a, 2004, 2005, 2006) and Hallin et al. (2006).

Our work has significantly new contributions to high dimensional robust statistics literature. Theoretically, we study the use of multivariate Kendall's tau in high dimensional settings, provide new properties of the multivariate rank statistic, and characterize the performance of ECA under both non-sparse and sparse settings. Computationally, we provide an efficient algorithm for conducting sparse ECA and highlight the "optimal rate, suboptimal scaling" phenomenon in understanding the behavior of the proposed algorithm.

4.1.2 Notation

Let $\mathrm{M} = [\mathrm{M}_{jk}] \in \mathbb{R}^{d\times d}$ be a symmetric matrix and $\boldsymbol{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$ be a vector. We denote $\boldsymbol{v}_I$ to be the subvector of $\boldsymbol{v}$ whose entries are indexed by a set I, and $\mathrm{M}_{I,J}$ to be the submatrix of M whose rows are indexed by I and columns are indexed by J. We denote $\mathrm{supp}(\boldsymbol{v}) := \{j : v_j \neq 0\}$. For $0 < q < \infty$, we define the ℓ_q and ℓ_∞ vector norms as $\|\boldsymbol{v}\|_q := (\sum_{i=1}^d |v_i|^q)^{1/q}$ and $\|\boldsymbol{v}\|_\infty := \max_{1\leq i\leq d} |v_i|$. We denote $\|\boldsymbol{v}\|_0 := \mathrm{card}(\mathrm{supp}(\boldsymbol{v}))$. We define the matrix entry-wise maximum value and Forbenius norms as $\|\mathrm{M}\|_{\max} := \max\{|\mathrm{M}_{ij}|\}$ and $\|\mathrm{M}\|_{\mathsf{F}} = (\sum \mathrm{M}_{jk}^2)^{1/2}$. Let $\lambda_j(\mathrm{M})$ be the j-th largest eigenvalue of M. Let $\boldsymbol{u}_j(\mathrm{M})$ be the eigenvector of M corresponding to $\lambda_j(\mathrm{M})$. With no loss of generality, we assume

the first nonzero entry of $\boldsymbol{u}_j(\mathbf{M})$ is positive. We denote $\|\mathbf{M}\|_2$ to be the spectral norm of $\mathbf{M}$ and $\mathbb{S}^{d-1} := \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$ to be the d-dimensional unit sphere. We define the restricted spectral norm $\|\mathbf{M}\|_{2,s} := \sup_{\boldsymbol{v}\in\mathbb{S}^{d-1},\|\boldsymbol{v}\|_0\leq s} |\boldsymbol{v}^T\mathbf{M}\boldsymbol{v}|$, so for $s \geq d$, we have $\|\mathbf{M}\|_{2,s} = \|\mathbf{M}\|_2$. We denote $f(\mathbf{M})$ to be the matrix with entries $[f(\mathbf{M})]_{jk} = f(\mathbf{M}_{jk})$. We denote $\mathrm{diag}(\mathbf{M})$ to be the diagonal matrix with the same diagonal entries as $\mathbf{M}$.

For any two numbers $a, b \in \mathbb{R}$, we denote $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For any two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$, or equivalently $b_n = \Omega(a_n)$, if there exist some constants N and C such that $a_n \leq Cb_n$ for all $n > N$. We write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We write $a_n = o(b_n)$ if for every positive constant ϵ, there exists a constant N such that $a_n \leq \epsilon b_n$ for all $n \geq N$. We write $b_n = \Omega^o(a_n)$ if $b_n = \Omega(a_n)$ and $b_n \not\asymp a_n$.

4.1.3 Chapter Organization

The rest of this chapter is organized as follows. In the next section, we briefly introduce the elliptical distribution, and review the marginal and multivariate Kendall's tau statistics. In Section 4.3, in the non-sparse setting, we propose the ECA method and study its theoretical performance. In Section 4.4, in the sparse setting, we propose a sparse ECA method via a combinatoric program and study its theoretical performance. A computationally efficient algorithm for conducting sparse ECA is provided in Section 4.5. Experiments on both synthetic and brain imaging data are provided in Section 4.6. The discussions are put in Section 4.7.

4.2 Background

In this section we briefly review the elliptical distribution, and marginal and multivariate Kendall's tau statistics. In the sequel, we denote $\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{Y}$ if random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same distribution.

4.2.1 Elliptical Distribution

The elliptical distribution is defined as follows. Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(\boldsymbol{\Sigma}) = q \leq d$. A d-dimensional random vector $\boldsymbol{X}$ has an elliptical distribution, denoted by $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$, if it has a stochastic representation

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + \xi \mathbf{A}\boldsymbol{U}, \tag{4.2.1}$$

where $\boldsymbol{U}$ is a uniform random vector on the unit sphere in $\mathbb{R}^q$, $\xi \geq 0$ is a scalar random variable independent of $\boldsymbol{U}$, $\mathbf{A} \in \mathbb{R}^{d\times q}$ is a deterministic matrix satisfying $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$. Here $\boldsymbol{\Sigma}$ is called the scatter matrix. In this chapter, we only consider continuous elliptical distributions with $\mathbb{P}(\xi = 0) = 0$.

An equivalent definition of the elliptical distribution is through the characteristic function $\exp(\mathrm{i}\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$, where ψ is a properly defined characteristic function and $\mathrm{i} := \sqrt{-1}$. ξ and ψ are mutually determined. In this setting, we denote by $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \psi)$. The elliptical family is closed to independent sums, and the marginal and conditional distributions of an elliptical distribution are also elliptically distributed.

Compared to the Gaussian family, the elliptical family provides more flexibility in modeling complex data. First, the elliptical family can model heavy-tail distributions (In contrast, Gaussian is light-tailed with exponential tail bounds.). Secondly, the elliptical family can be used to model nontrivial tail dependence between variables (Hult and Lindskog, 2002), i.e., different variables tend to go to extremes together (In contrast, Gaussian family can not capture any dependence in the tails.). The capability to handle heavy-tailed distributions and tail dependence is important for modeling many datasets, including: (1) Financial data (Almost all the financial data are heavy-tailed with nontrivial tail dependence. See Rachev (2003) and Čižek et al. (2005)); (2) Genomics data (See Liu et al. (2003) and Posekany et al. (2011)); (3) Bioimaging data (The fMRI data have heavy tails. See, for example, Ruttimann et al. (1998)).

In the sequel, we assume $\mathbb{E}\xi^2 < \infty$ so the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$ is well defined. For model identifiability, we further assume $\mathbb{E}\xi^2 = q$ so $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

4.2.2 Marginal Rank-Based Estimators

In this section, we briefly review the marginal rank-based estimator using the Kendall's tau statistic. This statistic plays a vital role for estimating the leading eigenvectors of the generalized correlation matrix $\boldsymbol{\Sigma}^0$ in Han and Liu (2014b). Letting $\boldsymbol{X} := (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ with $\widetilde{\boldsymbol{X}} := (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ be an independent copy of $\boldsymbol{X}$, the population Kendall's tau

statistic is defined as:

$$\tau(X_j, X_k) := \mathrm{Cov}(\mathrm{sign}(X_j - \widetilde{X}_j), \mathrm{sign}(X_k - \widetilde{X}_k)).$$

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ with $\boldsymbol{X}_i := (x_{i1}, \ldots, x_{id})^T$ be n independent observations of $\boldsymbol{X}$. The sample Kendall's tau statistic is defined as:

$$\widehat{\tau}_{jk}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) := \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \mathrm{sign}(x_{ij} - x_{i'j}) \mathrm{sign}(x_{ik} - x_{i'k}).$$

It is easy to verify $\mathbb{E}\widehat{\tau}_{jk}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \tau(X_j, X_k)$. Let $\widehat{\mathbf{R}} = [\widehat{\mathbf{R}}_{jk}] \in \mathbb{R}^{d \times d}$, with $\widehat{\mathbf{R}}_{jk} = \sin(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n))$, be the Kendall's tau correlation matrix. The marginal rank-based estimator $\widetilde{\boldsymbol{\theta}}_1$ used by TCA is obtained by plugging $\widehat{\mathbf{R}}$ into the optimization formulation in Vu and Lei (2012). When $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ and under mild conditions, Han and Liu (2014b) showed

$$\mathbb{E}|\sin \angle(\widetilde{\boldsymbol{\theta}}_1, \boldsymbol{u}_1(\boldsymbol{\Sigma}^0))| = O\left(s\sqrt{\frac{\log d}{n}}\right),$$

where $s := \|\boldsymbol{u}_1(\boldsymbol{\Sigma}^0)\|_0$ and $\boldsymbol{\Sigma}^0$ is the generalized correlation matrix of $\boldsymbol{X}$. However, TCA is a variant of the scale-invariant PCA and can only estimate the leading eigenvector of the correlation matrix. Then how to estimate the leading eigenvector of the covariance matrix in high dimensional elliptical models? A straightforward approach is to exploit a

covariance matrix estimator $\widehat{\mathbf{S}} := [\widehat{\mathbf{S}}_{jk}]$, defined as

$$\widehat{\mathbf{S}}_{jk} = \widehat{\mathbf{R}}_{jk} \cdot \widehat{\sigma}_j \widehat{\sigma}_k, \tag{4.2.2}$$

where $\{\widehat{\sigma}_j\}_{j=1}^d$ are sample standard deviations. However, since the elliptical distribution can be heavy-tailed, estimating the standard deviations is challenging and requires strong moment conditions. In this work, we solve this problem by resorting to the multivariate rank-based method.

4.2.3 Multivariate Kendall's tau

Let $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ and $\widetilde{\boldsymbol{X}}$ be an independent copy of $\boldsymbol{X}$. The population multivariate Kendall's tau matrix, denoted by $\mathbf{K} \in \mathbb{R}^{d \times d}$, is defined as:

$$\mathbf{K} := \mathbb{E}\left(\frac{(\boldsymbol{X} - \widetilde{\boldsymbol{X}})(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^T}{\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|_2^2} \right). \tag{4.2.3}$$

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be n independent data points of a random vector $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. The definition of multivariate Kendall's tau in (4.2.3) motivates the following sample version multivariate Kendall's tau estimator, which is a second-order U-statistic:

$$\widehat{\mathbf{K}} := \frac{2}{n(n-1)} \sum_{i' \leq i} \frac{(\boldsymbol{X}_i - \boldsymbol{X}_{i'})(\boldsymbol{X}_i - \boldsymbol{X}_{i'})^T}{\|\boldsymbol{X}_i - \boldsymbol{X}_{i'}\|_2^2}. \tag{4.2.4}$$

It is obvious $\mathbb{E}(\widehat{\mathbf{K}}) = \mathbf{K}$, and both $\mathbf{K}$ and $\widehat{\mathbf{K}}$ are positive semidefinite (PSD) matrices. Moreover, the kernel of the U statistics $k_{\mathsf{MK}}(\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$,

$$k_{\mathsf{MK}}(\boldsymbol{X}_i, \boldsymbol{X}_{i'}) := \frac{(\boldsymbol{X}_i - \boldsymbol{X}_{i'})(\boldsymbol{X}_i - \boldsymbol{X}_{i'})^T}{\|\boldsymbol{X}_i - \boldsymbol{X}_{i'}\|_2^2}, \tag{4.2.5}$$

is bounded under the spectral norm, i.e., $\|k_{\mathsf{MK}}(\cdot)\|_2 \leq 1$. Intuitively, such a boundedness property makes the U-statistic $\widehat{\mathbf{K}}$ more amenable to theoretical analysis. Moreover, it is worth noting the $k_{\mathsf{MK}}(\boldsymbol{X}_i, \boldsymbol{X}_{i'})$ is a distribution-free kernel, i.e., for any continuous $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ with the generating variable ξ,

$$k_{\mathsf{MK}}(\boldsymbol{X}_i, \boldsymbol{X}_{i'}) \overset{\mathsf{d}}{=} k_{\mathsf{MK}}(\boldsymbol{z}_i, \boldsymbol{z}_{i'}),$$

where $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i'}$ follow $\boldsymbol{Z} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This can be proven using the closedness of the elliptical family to independent sums, and $\boldsymbol{Z}$ is a stochastic scaling of $\boldsymbol{X}$. Accordingly, as will be shown later, the convergence of $\widehat{\mathbf{K}}$ to $\mathbf{K}$ does not depend on the generating variable ξ, and hence $\widehat{\mathbf{K}}$ enjoys the same distribution-free property as the Tyler's M estimator (Tyler, 1987). However, multivariate Kendall's tau can be directly extended to analyze high dimensional data, while Tyler's M estimator cannot.

Multivariate Kendall's tau $\mathbf{K}$ can be viewed as the covariance matrix of the self-normalized data $\{(\boldsymbol{X}_i - \boldsymbol{X}_{i'})/\|\boldsymbol{X}_i - \boldsymbol{X}_{i'}\|_2\}_{i>i'}$. It is immediate to see $\mathbf{K}$ is not identical or proportional to the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$. However, the following proposition, coming from Marden (1999) and Croux et al. (2002), states the eigenspace of the multivariate Kendall's

tau statistic $\mathbf{K}$ is identical to the eigenspace of the covariance matrix $\mathbf{\Sigma}$.

Proposition 4.2.1. *Let $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \mathbf{\Sigma}, \xi)$ be a continuous distribution and $\mathbf{K}$ be the population multivariate Kendall's tau statistic. Then if $\operatorname{rank}(\mathbf{\Sigma}) = q$ and $\lambda_j(\mathbf{\Sigma}) \neq \lambda_k(\mathbf{\Sigma})$ for any $k \neq j \in \{1, \ldots, q\}$, we have*

$$\boldsymbol{u}_j(\mathbf{\Sigma}) = \boldsymbol{u}_j(\mathbf{K}) \quad \text{and} \quad \lambda_j(\mathbf{K}) = \mathbb{E}\bigg(\frac{\lambda_j(\mathbf{\Sigma})Y_j^2}{\lambda_1(\mathbf{\Sigma})Y_1^2 + \ldots + \lambda_q(\mathbf{\Sigma})Y_q^2}\bigg), \tag{4.2.6}$$

where $\boldsymbol{Y} := (Y_1, \ldots, Y_q)^T \sim N_q(\mathbf{0}, \mathbf{I}_q)$ is a standard multivariate Gaussian distribution.

Proposition 4.2.1 builds the connection between $\mathbf{K}$ and $\mathbf{\Sigma}$, verifying they share the same eigenspace with the same descending orders of the eigenvalues. Therefore, to recover the eigenspace of the covariance matrix $\mathbf{\Sigma}$, we can resort to recovering the eigenspace of $\mathbf{K}$, which, as is discussed above, can be more efficient in estimation via using $\widehat{\mathbf{K}}$.

Remark 4.2.2. *Equation* (4.2.6) *shows that $\boldsymbol{u}_j(\mathbf{\Sigma}) = \boldsymbol{u}_j(\mathbf{K})$. Given that $\mathbf{\Sigma}$ and $\mathbf{K}$ share the same eigenspace as was shown in Marden (1999), this is equivalent to the statement that $\lambda_j(\mathbf{\Sigma}) > \lambda_k(\mathbf{\Sigma})$ implies $\lambda_j(\mathbf{K}) > \lambda_k(\mathbf{K})$. Actually, we have*

$$\frac{\lambda_k(\mathbf{K})}{\lambda_j(\mathbf{K})} = \frac{\mathbb{E}\frac{\lambda_k(\mathbf{\Sigma})Y_k^2}{\lambda_j(\mathbf{\Sigma})Y_j^2+\lambda_k(\mathbf{\Sigma})Y_k^2+E}}{\mathbb{E}\frac{\lambda_j(\mathbf{\Sigma})Y_j^2}{\lambda_j(\mathbf{\Sigma})Y_j^2+\lambda_k(\mathbf{\Sigma})Y_k^2+E}} < \frac{\mathbb{E}\frac{\lambda_k(\mathbf{\Sigma})Y_k^2}{\lambda_k(\mathbf{\Sigma})Y_j^2+\lambda_k(\mathbf{\Sigma})Y_k^2+E}}{\mathbb{E}\frac{\lambda_j(\mathbf{\Sigma})Y_j^2}{\lambda_j(\mathbf{\Sigma})Y_j^2+\lambda_j(\mathbf{\Sigma})Y_k^2+E}} = \frac{\mathbb{E}\frac{Y_k^2}{Y_j^2+Y_k^2+E/\lambda_k(\mathbf{\Sigma})}}{\mathbb{E}\frac{Y_k^2}{Y_j^2+Y_k^2+E/\lambda_j(\mathbf{\Sigma})}} < 1,$$

where we let $E := \sum_{i \notin \{j,k\}} \lambda_i(\mathbf{\Sigma})Y_i^2$ independent of $\{Y_j, Y_k\}$. This provides a simple proof of the left term of (4.2.6).

Remark 4.2.3. *Proposition 4.2.1 shows the eigenspaces of $\mathbf{K}$ and $\mathbf{\Sigma}$ are identical and the*

eigenvalues of $\mathbf{K}$ *only depend on the eigenvalues of* $\mathbf{\Sigma}$*. Therefore, if we can theoretically calculate the relationships between* $\{\lambda_j(\mathbf{K})\}_{j=1}^d$ *and* $\{\lambda_j(\mathbf{\Sigma})\}_{j=1}^d$*, we can recover* $\mathbf{\Sigma}$ *using* $\widehat{\mathbf{K}}$*. When, for example,* $\lambda_1(\mathbf{\Sigma}) = \cdots = \lambda_q(\mathbf{\Sigma})$*. this relationship is calculable. In particular, it is shown (Check, for example, Section 3 in Bilodeau and Brenner (1999)) that*

$$\frac{Y_j^2}{Y_1^2 + \cdots + Y_q^2} \sim \text{Beta}\Big(\frac{1}{2}, \frac{q-1}{2}\Big), \quad \text{for } j = 1, \ldots, q,$$

where $\text{Beta}(\alpha, \beta)$ *is the beta distribution with parameters* α *and* β*. Accordingly,* $\lambda_j(\mathbf{K}) = \mathbb{E}(Y_j^2/(Y_1^2+\cdots+Y_q^2)) = 1/q$*. The general relationship between* $\{\lambda_j(\mathbf{K})\}_{j=1}^d$ *and* $\{\lambda_j(\mathbf{\Sigma})\}_{j=1}^d$ *is non-linear: For example, when* $d = 2$*, Croux et al. (2002) showed that*

$$\lambda_j(\mathbf{K}) = \frac{\sqrt{\lambda_j(\mathbf{\Sigma})}}{\sqrt{\lambda_1(\mathbf{\Sigma})} + \sqrt{\lambda_2(\mathbf{\Sigma})}}, \quad \text{for } j = 1, 2.$$

4.3 ECA: Non-Sparse Setting

In this section, we propose and study the ECA method under the non-sparse setting, i.e, we do not assume sparsity of $\boldsymbol{u}_1(\mathbf{\Sigma})$. Without the sparsity assumption, we propose to use the leading eigenvector $\boldsymbol{u}_1(\widehat{\mathbf{K}})$ to estimate $\boldsymbol{u}_1(\mathbf{K}) = \boldsymbol{u}_1(\mathbf{\Sigma})$:

$$\text{The ECA estimator} : \boldsymbol{u}_1(\widehat{\mathbf{K}}) \text{ (the leading eigenvector of } \widehat{\mathbf{K}}),$$

where $\widehat{\mathbf{K}}$ is defined in (4.2.4). For notational simplicity, in the sequel we assume the sample size n is even. When n is odd, we can always use $n-1$ data points without affecting the obtained rate of convergence.

The approximation error of $\boldsymbol{u}_1(\widehat{\mathbf{K}})$ to $\boldsymbol{u}_1(\mathbf{K})$ is associated with the convergence of $\widehat{\mathbf{K}}$ to $\mathbf{K}$ under the spectral norm via the Davis-Kahan inequality (Davis and Kahan, 1970; Wedin, 1972). In detail, for any two vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$, let $\sin\angle(\boldsymbol{v}_1, \boldsymbol{v}_2)$ be the sine of the angle between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, with

$$|\sin\angle(\boldsymbol{v}_1, \boldsymbol{v}_2)| := \sqrt{1-(\boldsymbol{v}_1^T\boldsymbol{v}_2)^2}.$$

Davis-Kahan inequality states the approximation error of $\boldsymbol{u}_1(\widehat{\mathbf{K}})$ to $\boldsymbol{u}_1(\mathbf{K})$ is controlled by $\|\widehat{\mathbf{K}}-\mathbf{K}\|_2$ divided by the eigengap between $\lambda_1(\mathbf{K})$ and $\lambda_2(\mathbf{K})$:

$$|\sin\angle(\boldsymbol{u}_1(\widehat{\mathbf{K}}), \boldsymbol{u}_1(\mathbf{K}))| \leq \frac{2}{\lambda_1(\mathbf{K})-\lambda_2(\mathbf{K})}\|\widehat{\mathbf{K}}-\mathbf{K}\|_2. \tag{4.3.1}$$

Accordingly, to analyze the convergence rate of $\boldsymbol{u}_1(\widehat{\mathbf{K}})$ to $\boldsymbol{u}_1(\mathbf{K})$, we can focus on the convergence rate of $\widehat{\mathbf{K}}$ to $\mathbf{K}$ under the spectral norm. The next theorem shows, under the elliptical distribution family, the convergence rate of $\widehat{\mathbf{K}}$ to $\mathbf{K}$ under the spectral norm is $\|\mathbf{K}\|_2\sqrt{r^*(\mathbf{K})\log d/n}$, where $r^*(\mathbf{K}) = \mathrm{Tr}(\mathbf{K})/\lambda_1(\mathbf{K})$ is the effective rank of $\mathbf{K}$ and must be less than or equal to d.

Theorem 4.3.1. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independent observations of $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. Let $\widehat{\mathbf{K}}$ be the sample version multivariate Kendall's tau statistic defined in Equation* (4.2.4).

We have, provided that n is sufficiently large such that

$$n \geq \frac{16}{3} \cdot (r^*(\mathbf{K}) + 1)(\log d + \log(1/\alpha)),$$

with probability larger than $1 - \alpha$,

$$\|\widehat{\mathbf{K}} - \mathbf{K}\|_2 \leq \|\mathbf{K}\|_2 \sqrt{\frac{16}{3} \cdot \frac{(r^*(\mathbf{K}) + 1)(\log d + \log(1/\alpha))}{n}}.$$

There is a vast literature in bounding the spectral norm of a random matrix (See, for example, Vershynin (2010) and the references therein) and our proof relies on the matrix Bernstein inequality proposed in Tropp (2012), with a generalization to U-statistics.

Combining (4.3.1) and Theorem 4.3.1, we immediately have the following corollary, which characterizes the explicit rate of convergence for $|\sin \angle(\boldsymbol{u}_1(\widehat{\mathbf{K}}), \boldsymbol{u}_1(\mathbf{K})|$.

Corollary 4.3.2. *Under the conditions of Theorem 4.3.1, provided that n is sufficiently large such that*

$$n \geq \frac{16}{3} \cdot (r^*(\mathbf{K}) + 1)(\log d + \log(1/\alpha)),$$

we have, with probability larger than $1 - \alpha$,

$$|\sin \angle(\boldsymbol{u}_1(\widehat{\mathbf{K}}), \boldsymbol{u}_1(\mathbf{K}))| \leq \frac{2\lambda_1(\mathbf{K})}{\lambda_1(\mathbf{K}) - \lambda_2(\mathbf{K})} \sqrt{\frac{16}{3} \cdot \frac{(r^*(\mathbf{K}) + 1)(\log d + \log(1/\alpha))}{n}}.$$

Remark 4.3.3. *Corollary 4.3.2 indicates it is not necessary to require $d/n \to 0$ for $\boldsymbol{u}_1(\widehat{\mathbf{K}})$*

to be a consistent estimator of $\boldsymbol{u}_1(\mathbf{K})$. *For example, when* $\lambda_2(\mathbf{K})/\lambda_1(\mathbf{K})$ *is upper bounded by an absolute constant strictly smaller than 1,* $r^*(\mathbf{K}) \log d/n \to 0$ *is sufficient to make* $\boldsymbol{u}_1(\widehat{\mathbf{K}})$ *a consistent estimator of* $\boldsymbol{u}_1(\mathbf{K})$. *Such an observation is consistent to the observations in the PCA theories and the* $\log d$ *term here is generally unavoidable if we wish to highlight the role of the effective rank* $r^*(\mathbf{K})$ *(Lounici, 2013a; Bunea and Xiao, 2014). On the other hand, Theorem 4.4.1 in the next section provides a rate of convergence* $O_P(\lambda_1(\mathbf{K})\sqrt{d/n})$ *for* $\|\widehat{\mathbf{K}} - \mathbf{K}\|_2$. *Therefore, the final rate of convergence for ECA, under various settings, can be expressed as* $O_P(\sqrt{r^*(\mathbf{K}) \log d/n} \wedge \sqrt{d/n})$.

Remark 4.3.4. *We note Theorem 4.3.1 can also help to quantify the subspace estimation error via a variation of the Davis-Kahan inequality. In particular, let* $\mathcal{P}^m(\widehat{\mathbf{K}})$ *and* $\mathcal{P}^m(\mathbf{K})$ *be the projection matrices to the span of* m *leading eigenvectors of* $\widehat{\mathbf{K}}$ *and* $\mathbf{K}$. *Using Lemma 4.2 in Vu and Lei (2013), we have*

$$\|\mathcal{P}^m(\widehat{\mathbf{K}}) - \mathcal{P}^m(\mathbf{K})\|_{\mathsf{F}} \leq \frac{2\sqrt{2m}}{\lambda_m(\mathbf{K}) - \lambda_{m+1}(\mathbf{K})}\|\widehat{\mathbf{K}} - \mathbf{K}\|_2, \tag{4.3.2}$$

so that $\|\mathcal{P}^m(\widehat{\mathbf{K}}) - \mathcal{P}^m(\mathbf{K})\|_{\mathsf{F}}$ *can be controlled via a similar argument as in Corollary 4.3.2.*

The above bounds are all related to the eigenvalues of $\mathbf{K}$. The next theorem connects the eigenvalues of $\mathbf{K}$ to the eigenvalues of $\boldsymbol{\Sigma}$, so we can directly bound $\|\widehat{\mathbf{K}} - \mathbf{K}\|_2$ and $|\sin\angle(\boldsymbol{u}_1(\widehat{\mathbf{K}}), \boldsymbol{u}_1(\mathbf{K}))|$ using $\boldsymbol{\Sigma}$. In the sequel, let's denote $r^{**}(\boldsymbol{\Sigma}) := \|\boldsymbol{\Sigma}\|_{\mathsf{F}}/\lambda_1(\boldsymbol{\Sigma}) \leq \sqrt{d}$ to be the "second-order" effective rank of the matrix $\boldsymbol{\Sigma}$.

Theorem 4.3.5 (The upper and lower bounds of $\lambda_j(\mathbf{K})$)**.** *Letting* $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$*, we have*

$$\lambda_j(\mathbf{K}) \geq \frac{\lambda_j(\boldsymbol{\Sigma})}{\mathrm{Tr}(\boldsymbol{\Sigma}) + 2\|\boldsymbol{\Sigma}\|_{\mathsf{F}}\sqrt{\log d} + 2\|\boldsymbol{\Sigma}\|_2 \log d}\left(1 - \frac{1}{d}\right),$$

and when $\mathrm{Tr}(\boldsymbol{\Sigma}) > 2\|\boldsymbol{\Sigma}\|_{\mathsf{F}}\sqrt{\log d}$*,*

$$\lambda_j(\mathbf{K}) \leq \frac{\lambda_j(\boldsymbol{\Sigma})}{\mathrm{Tr}(\boldsymbol{\Sigma}) - 2\|\boldsymbol{\Sigma}\|_{\mathsf{F}}\sqrt{\log d}} + \frac{1}{d}.$$

Using Theorem 4.3.5, we can replace $r^*(\mathbf{K})$ by $(r^*(\boldsymbol{\Sigma}) + 2r^{**}(\boldsymbol{\Sigma})\sqrt{\log d} + 2\log d) \cdot d/(d-1)$ in Theorem 4.3.1. We also note Theorem 4.3.5 can help understand the scaling of $\lambda_j(\mathbf{K})$ with regard to $\lambda_j(\boldsymbol{\Sigma})$. Actually, when $\|\boldsymbol{\Sigma}\|_{\mathsf{F}} \log d = \mathrm{Tr}(\boldsymbol{\Sigma}) \cdot o(1)$, we have $\lambda_j(\mathbf{K}) \asymp \lambda_j(\boldsymbol{\Sigma})/\mathrm{Tr}(\boldsymbol{\Sigma})$, and accordingly, we can continue to write

$$\frac{\lambda_1(\mathbf{K})}{\lambda_1(\mathbf{K}) - \lambda_2(\mathbf{K})} \asymp \frac{\lambda_1(\boldsymbol{\Sigma})}{\lambda_1(\boldsymbol{\Sigma}) - \lambda_2(\boldsymbol{\Sigma})}.$$

In practice, $\|\boldsymbol{\Sigma}\|_{\mathsf{F}} \log d = \mathrm{Tr}(\boldsymbol{\Sigma}) \cdot o(1)$ is a mild condition. For example, when the condition number of $\boldsymbol{\Sigma}$ is upper bounded by an absolute constant, we have $\mathrm{Tr}(\boldsymbol{\Sigma}) \asymp \|\boldsymbol{\Sigma}\|_{\mathsf{F}} \cdot \sqrt{d}$.

4.4 Sparse ECA via a Combinatoric Program

We analyze the theoretical properties of ECA under the sparse setting, where we assume $\|\boldsymbol{u}_1(\boldsymbol{\Sigma})\|_0 \leq s < d \wedge n$. In this section we study the ECA method using a combinatoric

program. For any matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, we define the best s-sparse vector to $\boldsymbol{u}_1(\mathbf{M})$ as

$$\boldsymbol{u}_{1,s}(\mathbf{M}) := \underset{\|\boldsymbol{v}\|_0 \leq s, \|\boldsymbol{v}\|_2 \leq 1}{\arg\max} |\boldsymbol{v}^T \mathbf{M} \boldsymbol{v}|. \tag{4.4.1}$$

We propose to estimate $\boldsymbol{u}_1(\boldsymbol{\Sigma}) = \boldsymbol{u}_1(\mathbf{K})$ via a combinatoric program:

$$\text{Sparse ECA estimator via a combinatoric program} : \boldsymbol{u}_{1,s}(\widehat{\mathbf{K}}),$$

where $\widehat{\mathbf{K}}$ is defined in (4.2.4). Under the sparse setting, by definition we have $\boldsymbol{u}_{1,s}(\mathbf{K}) = \boldsymbol{u}_1(\mathbf{K}) = \boldsymbol{u}_1(\boldsymbol{\Sigma})$. On the other hand, $\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}})$ can be calculated via a combinatoric program by exhaustively searching over all s by s submatrices of $\widehat{\mathbf{K}}$. This global search is not computationally efficient. However, the result in quantifying the approximation error of $\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}})$ to $\boldsymbol{u}_1(\mathbf{K})$ is of strong theoretical interest. Similar algorithms were also studied in Vu and Lei (2012), Lounici (2013b), Vu and Lei (2013), and Cai et al. (2014b). Moreover, as will be seen in the next section, this will help clarify that a computationally efficient sparse ECA algorithm can attain the same convergence rate, under a suboptimal scaling of (n, d, s) though.

In the following we study the performance of $\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}})$ in conducting sparse ECA. The approximation error of $\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}})$ to $\boldsymbol{u}_1(\mathbf{K})$ is connected to the approximation error of $\widehat{\mathbf{K}}$ to $\mathbf{K}$ under the restricted spectral norm. This is due to the following Davis-Kahan type

inequality provided in Vu and Lei (2012):

$$|\sin\angle(\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}}),\boldsymbol{u}_{1,s}(\mathbf{K}))| \leq \frac{2}{\lambda_1(\mathbf{K})-\lambda_2(\mathbf{K})}\|\widehat{\mathbf{K}}-\mathbf{K}\|_{2,2s}. \tag{4.4.2}$$

Accordingly, for studying $|\sin\angle(\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}}),\boldsymbol{u}_{1,s}(\mathbf{K}))|$, we focus on studying the approximation error $\|\widehat{\mathbf{K}}-\mathbf{K}\|_{2,s}$. Before presenting the main results, we provide some extra notation. For any random variable $X \in \mathbb{R}$, we define the subgaussian ($\|\cdot\|_{\psi_2}$) and subexponential norms ($\|\cdot\|_{\psi_1}$) of X as follows:

$$\|X\|_{\psi_2} := \sup_{k\geq 1} k^{-1/2}(\mathbb{E}|X|^k)^{1/k} \quad \text{and} \quad \|X\|_{\psi_1} := \sup_{k\geq 1} k^{-1}(\mathbb{E}|X|^k)^{1/k}. \tag{4.4.3}$$

Any d-dimensional random vector $\boldsymbol{X} \in \mathbb{R}^d$ is said to be subgaussian distributed with the subgaussian constant σ if

$$\|\boldsymbol{v}^T\boldsymbol{X}\|_{\psi_2} \leq \sigma, \quad \text{for any } \boldsymbol{v} \in \mathbb{S}^{d-1}.$$

Moreover, we define the self-normalized operator $S(\cdot)$ for any random vector to be

$$S(\boldsymbol{X}) := (\boldsymbol{X}-\widetilde{\boldsymbol{X}})/\|\boldsymbol{X}-\widetilde{\boldsymbol{X}}\|_2 \quad \text{where } \widetilde{\boldsymbol{X}} \text{ is an independent copy of X.} \tag{4.4.4}$$

It follows that $\mathbf{K} = \mathbb{E}S(\boldsymbol{X})S(\boldsymbol{X})^T$.

The next theorem provides a general result in quantifying the approximation error of $\widehat{\mathbf{K}}$

to $\mathbf{K}$ with regard to the restricted spectral norm.

Theorem 4.4.1. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. Let $\widehat{\mathbf{K}}$ be the sample version multivariate Kendall's tau statistic defined in Equation* (4.2.4)*. We have, when*

$$s \log(d/s)/n \to 0,$$

for n sufficiently large, with probability larger than $1 - 2\alpha$,

$$\|\widehat{\mathbf{K}} - \mathbf{K}\|_{2,s} \leq \left(\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} 2\|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2 + \|\mathbf{K}\|_2 \right) \cdot C_0 \sqrt{\frac{s(3 + \log(d/s)) + \log(1/\alpha)}{n}},$$

for some absolute constant C_0. Here $\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \|\boldsymbol{v}^T \boldsymbol{X}\|_{\psi_2}$ can be further written as

$$\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2} = \sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \left\| \frac{\sum_{i=1}^d v_i \lambda_i^{1/2}(\boldsymbol{\Sigma}) Y_i}{\sqrt{\sum_{i=1}^d \lambda_i(\boldsymbol{\Sigma}) Y_i^2}} \right\|_{\psi_2} \leq 1, \tag{4.4.5}$$

where $\boldsymbol{v} := (v_1, \ldots, v_d)^T$ and $(Y_1, \ldots, Y_d)^T \sim N_d(\mathbf{0}, \mathbf{I}_d)$.

It is obvious that $S(\boldsymbol{X})$ is subgaussian with a variance proxy 1. However, typically a sharper upper bound can be obtained. The next theorem shows, in various settings, the upper bound can be in the same order of $1/q$, which is much smaller than 1. Combined with Theorem 4.4.1, these results give an upper bound of $\|\widehat{\mathbf{K}} - \mathbf{K}\|_{2,s}$.

Theorem 4.4.2. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ with $\operatorname{rank}(\boldsymbol{\Sigma}) = q$ and $\|\boldsymbol{u}_1(\boldsymbol{\Sigma})\|_0 \leq s$. Let $\widehat{\mathbf{K}}$ be the sample version multivariate Kendall's tau statistic*

defined in Equation (4.2.4)*. We have,*

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2} \leq \sqrt{\frac{\lambda_1(\boldsymbol{\Sigma})}{\lambda_q(\boldsymbol{\Sigma})} \cdot \frac{2}{q}} \wedge 1,$$

and accordingly, when $s\log(d/s)/n = o(1)$*, with probability at least* $1-2\alpha$*,*

$$\|\widehat{\mathbf{K}} - \mathbf{K}\|_{2,s} \leq C_0 \left\{ \left(\frac{4\lambda_1(\boldsymbol{\Sigma})}{q\lambda_q(\boldsymbol{\Sigma})} \wedge 1 \right) + \lambda_1(\mathbf{K}) \right\} \sqrt{\frac{s(3+\log(d/s)) + \log(1/\alpha)}{n}}.$$

Similar as Theorem 4.3.1, we wish to show

$$\|\widehat{\mathbf{K}} - \mathbf{K}\|_{2,s} = O_P(\lambda_1(\mathbf{K})\sqrt{s\log(d/s)/n}).$$

In the following, we provide several examples such that $\sup_{\boldsymbol{v}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2$ is in the same order of $\lambda_1(\mathbf{K})$, so that, via Theorem 4.4.2, the desired rate is attained.

- **Condition number controlled:** Bickel and Levina (2008a) considered the covariance matrix model where the condition number of $\boldsymbol{\Sigma}$, $\lambda_1(\boldsymbol{\Sigma})/\lambda_d(\boldsymbol{\Sigma})$, is upper bounded by an absolute constant. Under this condition, we have

$$\sup_{\boldsymbol{v}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2 \asymp d^{-1},$$

and applying Theorem 4.3.5 we also have

$$\lambda_j(\mathbf{K}) \asymp \frac{\lambda_j(\mathbf{\Sigma})}{\mathrm{Tr}(\mathbf{\Sigma})} \asymp d^{-1}.$$

Accordingly, we conclude $\sup_{\boldsymbol{v}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2$ and $\lambda_1(\mathbf{K})$ are in the same order.

- **Spike covariance model:** Johnstone and Lu (2009) considered the following simple spike covariance model:

$$\mathbf{\Sigma} = \beta \boldsymbol{v}\boldsymbol{v}^T + a^2 \mathbf{I}_d,$$

where $\beta, a > 0$ are two positive real numbers and $\boldsymbol{v} \in \mathbb{S}^{d-1}$. In this case, we have, when $\beta = o(da^2/\sqrt{\log d})$ or $\beta = \Omega(da^2)$,

$$\sup_{\boldsymbol{v}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2 \asymp \frac{\beta + a^2}{da^2} \wedge 1 \ \text{ and } \ \lambda_1(\mathbf{K}) \asymp \frac{\beta + a^2}{\beta + da^2}.$$

A simple calculation shows $\sup_{\boldsymbol{v}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2$ and $\lambda_1(\mathbf{K})$ are in the same order.

- **Multi-Factor Model:** Fan and Fan (2008) considered a multi-factor model, which is also related to the general spike covariance model (Ma, 2013):

$$\mathbf{\Sigma} = \sum_{j=1}^{m} \beta_j \boldsymbol{v}_j \boldsymbol{v}_j^T + \mathbf{\Sigma}_u,$$

where we have $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_m > 0$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbb{S}^{d-1}$ and are orthogonal to each other, and $\mathbf{\Sigma}_u$ is a diagonal matrix. For simplicity, we assume $\mathbf{\Sigma}_u = a^2 \mathbf{I}_d$.

When $\sum \beta_j^2 = o(d^2a^4/\log d)$, we have

$$\sup_{\boldsymbol{v}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2 \asymp \frac{\beta_1 + a^2}{da^2} \wedge 1 \text{ and } \lambda_1(\mathbf{K}) \asymp \frac{\beta_1 + a^2}{\sum_{j=1}^m \beta_j + da^2},$$

and $\sup_{\boldsymbol{v}} \|\boldsymbol{v}^T S(\boldsymbol{X})\|_{\psi_2}^2$ and $\lambda_1(\mathbf{K})$ are in the same order if, for example, $\sum_{j=1}^m \beta_j = O(da^2)$.

Equation (4.4.2) and Theorem 4.4.2 together give the following corollary, which quantifies the convergence rate of the sparse ECA estimator calculated via the combinatoric program in (4.4.1).

Corollary 4.4.3. *Under the condition of Theorem 4.4.2, if we have* $s\log(d/s)/n \to 0$*, for* n *sufficiently large, with probability larger than* $1 - 2\alpha$*,*

$$|\sin\angle(\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}}), \boldsymbol{u}_{1,s}(\mathbf{K}))| \leq \frac{2C_0(4\lambda_1(\boldsymbol{\Sigma})/q\lambda_q(\boldsymbol{\Sigma}) \wedge 1 + \lambda_1(\mathbf{K}))}{\lambda_1(\mathbf{K}) - \lambda_2(\mathbf{K})} \cdot \sqrt{\frac{2s(3 + \log(d/2s)) + \log(1/\alpha)}{n}}.$$

Remark 4.4.4. *The restricted spectral norm convergence result obtained in Theorem 4.4.2 is also applicable to analyze principal subspace estimation accuracy. Following the discussion in Vu and Lei (2013), we define the principal subspace estimator to the space spanned by the toppest* m *eigenvectors of any given matrix* $\mathbf{M} \in \mathbb{R}^{d\times d}$ *as*

$$\mathbf{U}_{m,s}(\mathbf{M}) := \underset{\mathbf{V}\in\mathbb{R}^{d\times m}}{\arg\max} \left\langle \mathbf{M}, \mathbf{V}\mathbf{V}^T \right\rangle, \text{ subject to } \sum_{j=1}^d I(\mathbf{V}_{j*} \neq 0) \leq s, \tag{4.4.6}$$

where V_{j*} is the j-th row of $\mathbf{M}$ and the indicator function is 0 if and only if $\mathbf{V}_{j*} = \mathbf{0}$. We then have

$$\|\mathbf{U}_{m,s}(\widehat{\mathbf{K}})\mathbf{U}_{m,s}(\widehat{\mathbf{K}})^T - \mathbf{U}_{m,s}(\mathbf{K})\mathbf{U}_{m,s}(\mathbf{K})^T\|_{\mathsf{F}} \leq \frac{2\sqrt{2m}}{\lambda_m(\mathbf{K}) - \lambda_{m+1}(\mathbf{K})} \cdot \|\widehat{\mathbf{K}} - \mathbf{K}\|_{2,2ms}.$$

An explicit statement of the above inequality can be found in Wang et al. (2013).

4.5 Sparse ECA via a Computationally Efficient Program

There is a vast literature in studying computationally efficient algorithms for estimating sparse $\boldsymbol{u}_1(\boldsymbol{\Sigma})$. In this section, we focus on such an algorithm for conducting sparse ECA by combining the Fantope projection (Vu et al., 2013) with the truncated power method (Yuan and Zhang, 2013).

4.5.1 Fantope Projection

In this section, we first review the algorithm and theory developed in Vu et al. (2013) for sparse subspace estimation, and then we provide some new analysis in obtaining the sparse leading eigenvector estimators.

Let $\mathbf{\Pi}_m := \mathbf{V}_m\mathbf{V}_m^T$ with $\mathbf{V}_m$ as the combination of the m leading eigenvectors of $\mathbf{K}$. It is well known that $\mathbf{\Pi}_m$ is the optimal rank-m projections to $\mathbf{K}$. Similarly as in (4.4.6), we

define s_Π to be the number of nonzero columns in $\mathbf{\Pi}_m$.

We then introduce the sparse principal subspace estimator $\mathbf{X}_m$ corresponding to the space spanned by the first m leading eigenvectors of the multivariate Kendall's tau matrix $\widehat{\mathbf{K}}$. To induce sparsity, $\mathbf{X}_m$ is defined to be the solution to the following convex program:

$$\mathbf{X}_m := \underset{\mathbf{M}\in\mathbb{R}^{d\times d}}{\arg\max} \big\langle \widehat{\mathbf{K}}, \mathbf{M} \big\rangle - \lambda \sum_{j,k} |\mathbf{M}_{jk}|, \;\; \text{subject to } \mathbf{0} \preceq \mathbf{M} \preceq \mathbf{I}_d \text{ and } \mathrm{Tr}(\mathbf{M}) = m, \tag{4.5.1}$$

where for any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, $\mathbf{A} \preceq \mathbf{B}$ represents $\mathbf{B}-\mathbf{A}$ is positive semidefinite. Here $\{\mathbf{M} : \mathbf{0} \preceq \mathbf{M} \preceq \mathbf{I}_d, \mathrm{Tr}(\mathbf{M}) = m\}$ is a convex set called the Fantope. We then have the following deterministic theorem to quantify the approximation error of $\mathbf{X}_m$ to $\mathbf{\Pi}_m$.

Theorem 4.5.1 (Vu et al. (2013))**.** *If the tuning parameter λ in* (4.5.1) *satisfies that $\lambda \geq \|\widehat{\mathbf{K}} - \mathbf{K}\|_{\max}$, we have*

$$\|\mathbf{X}_m - \mathbf{\Pi}_m\|_{\mathsf{F}} \leq \frac{4 s_\Pi \lambda}{\lambda_m(\mathbf{K}) - \lambda_{m+1}(\mathbf{K})},$$

where we note the s_Π is the number of nonzero columns in $\mathbf{\Pi}_m$.

It is easy to see $\mathbf{X}_m$ is symmetric and the rank of $\mathbf{X}_m$ must be greater or equal to m, but is not necessary to be exactly m. However, in various cases, dimension reduction for example, it is desired to estimate the top m leading eigenvectors of $\mathbf{\Sigma}$, or equivalently, to estimate an exactly rank m projection matrix. Noticing that $\mathbf{X}_m$ is a real symmetric matrix,

we propose to use the following estimate $\widehat{\mathbf{X}}_m \in \mathbb{R}^{d\times d}$:

$$\widehat{\mathbf{X}}_m := \sum_{j\leq m} \boldsymbol{u}_j(\mathbf{X}_m)[\boldsymbol{u}_j(\mathbf{X}_m)]^T. \tag{4.5.2}$$

We then have the next theorem, which quantifies the distance between $\widehat{\mathbf{X}}_m$ and $\mathbf{\Pi}_m$.

Theorem 4.5.2. *If* $\lambda \geq \|\widehat{\mathbf{K}} - \mathbf{K}\|_{\max}$, *we have*

$$\|\widehat{\mathbf{X}}_m - \mathbf{\Pi}_m\|_{\mathsf{F}} \leq 4\|\mathbf{X}_m - \mathbf{\Pi}_m\|_{\mathsf{F}} \leq \frac{16 s_{\Pi}\lambda}{\lambda_m(\mathbf{K}) - \lambda_{m+1}(\mathbf{K})}.$$

4.5.2 A Computationally Efficient Algorithm

In this section, we propose a computationally efficient algorithm to conduct sparse ECA via combining the Fantope projection with the truncated power algorithm proposed in Yuan and Zhang (2013). We focus on estimating the leading eigenvector of $\mathbf{K}$ because the rest can be iteratively estimated using the deflation method (Mackey, 2008).

The main idea here is to exploit the Fantope projection for constructing a good initial parameter for the truncated power algorithm and then perform iterative thresholding as in Yuan and Zhang (2013). We call this the Fantope-truncated power algorithm, or FTPM, for abbreviation. Before proceeding to the main algorithm, we first introduce some extra notation. For any vector $\boldsymbol{v} \in \mathbb{R}^d$ and an index set $J \subset \{1, \ldots, d\}$, we define the truncation

function $\mathrm{TRC}(\cdot,\cdot)$ to be

$$\mathrm{TRC}(\boldsymbol{v}, J) := \big(v_1 \cdot I(1 \in J),\ \ldots,\ v_d \cdot I(d \in J)\big)^T, \tag{4.5.3}$$

where $I(\cdot)$ is the indicator function. The initial parameter $\boldsymbol{v}^{(0)}$, then, is the normalized vector consisting of the largest entries in $\boldsymbol{u}_1(\mathbf{X}_1)$, where $\mathbf{X}_1$ is calculated in (4.5.1):

$$\boldsymbol{v}^{(0)} = \boldsymbol{w}^0/\|\boldsymbol{w}^0\|_2, \text{ where } \boldsymbol{w}^0 = \mathrm{TRC}(\boldsymbol{u}_1(\mathbf{X}_1), J_\delta) \text{ and } J_\delta = \{j : |(\boldsymbol{u}_1(\mathbf{X}_1))_j| \geq \delta\}. \tag{4.5.4}$$

We have $\|\boldsymbol{v}^{(0)}\|_0 = \mathrm{supp}\{j : |(\boldsymbol{u}_1(\mathbf{X}_1))_j| \geq \delta\}$. Algorithm 2 then provides the detailed FTPM algorithm and the final FTPM estimator is denoted as $\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}$.

Algorithm 2 The FTPM algorithm. Within each iteration, a new sparse vector $\boldsymbol{v}^{(t)}$ with $\|\boldsymbol{v}^{(t)}\|_0 \leq k$ is updated. The algorithm terminates when $\|\boldsymbol{v}^{(t)} - \boldsymbol{v}^{(t-1)}\|_2$ is less than a given threshold ϵ.

algorithmECA: $\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}}) \leftarrow \mathsf{FTPM}(\widehat{\mathbf{K}}, k, \epsilon)$
Initialize: $\mathbf{X}_1$ calculated by (4.5.1) with $m = 1$, $\boldsymbol{v}^{(0)}$ is calculated using (4.5.4), and $t \leftarrow 0$
Repeat:
  $t \leftarrow t + 1$
  $\boldsymbol{X}_y \leftarrow \widehat{\mathbf{K}}\boldsymbol{v}^{(t-1)}$
  If $\|\boldsymbol{X}_t\|_0 \leq k$, then $\boldsymbol{v}^{(t)} = \boldsymbol{X}_t/\|\boldsymbol{X}_t\|_2$
  Else, let A_t be the indices of the elements in $\boldsymbol{X}_t$ with the largest k absolute values
  $\boldsymbol{v}^{(t)} = \mathrm{TRC}(\boldsymbol{X}_t, A_t)/\|\,\mathrm{TRC}(\boldsymbol{X}_t, A_t)\|_2$
Until convergence: $\|\boldsymbol{v}^{(t)} - \boldsymbol{v}^{(t-1)}\|_2 \leq \epsilon$
$\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}}) \leftarrow \boldsymbol{v}^{(t)}$
Output: $\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}})$

In the rest of this section, we study the approximation accuracy of $\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}$ to $\boldsymbol{u}_1(\mathbf{K})$.

Via observing Theorems 4.5.1 and 4.5.2, it is immediate the approximation accuracy of $\boldsymbol{u}_1(\mathbf{X}_1)$ is related to $\|\widehat{\mathbf{K}} - \mathbf{K}\|_{\max}$. The next theorem gives a nonasymptotic upper bound of $\|\widehat{\mathbf{K}} - \mathbf{K}\|_{\max}$, and accordingly, combined with Theorems 4.5.1 and 4.5.2, gives an upper bound on $|\sin\angle(\boldsymbol{u}_1(\mathbf{X}_1), \boldsymbol{u}_1(\mathbf{K}))|$.

Theorem 4.5.3. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n observations of $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ with $\operatorname{rank}(\boldsymbol{\Sigma}) = q$ and $\|\boldsymbol{u}_1(\boldsymbol{\Sigma})\|_0 \leq s$. Let $\widehat{\mathbf{K}}$ be the sample version multivariate Kendall's tau statistic defined in Equation* (4.2.4)*. If $\log d/n = o(1)$, we have there exists some positive absolute constant C_1 such that for sufficiently large n, with probability at least $1 - \alpha^2$,*

$$\|\widehat{\mathbf{K}} - \mathbf{K}\|_{\max} \leq C_1\Big(\frac{8\lambda_1(\boldsymbol{\Sigma})}{q\lambda_q(\boldsymbol{\Sigma})} + \|\mathbf{K}\|_{\max}\Big)\sqrt{\frac{\log d + \log(1/\alpha)}{n}}.$$

Accordingly, if

$$\lambda \geq C_1\Big(\frac{8\lambda_1(\boldsymbol{\Sigma})}{q\lambda_q(\boldsymbol{\Sigma})} + \|\mathbf{K}\|_{\max}\Big)\sqrt{\frac{\log d + \log(1/\alpha)}{n}}, \tag{4.5.5}$$

we have, with probability at least $1 - \alpha^2$,

$$|\sin\angle(\boldsymbol{u}_1(\mathbf{X}_1), \boldsymbol{u}_1(\mathbf{K}))| \leq \frac{8\sqrt{2}s\lambda}{\lambda_1(\mathbf{K}) - \lambda_2(\mathbf{K})}.$$

Theorem 4.5.3 builds sufficient conditions under which $\boldsymbol{u}_1(\mathbf{X}_1)$ is a consistent estimator of $\boldsymbol{u}_1(\mathbf{K})$. Under multiple settings, the "condition number controlled", "spike covariance model", and "multi-factor model" settings considered in Section 4.4 for example, when

$\lambda \asymp \lambda_1(\mathbf{K})\sqrt{\log d/n}$, we have $|\sin\angle(\boldsymbol{u}_1(\mathbf{X}_1), \boldsymbol{u}_1(\mathbf{K}))| = O_P(s\sqrt{\log d/n})$. This is summarized in the next corollary.

Corollary 4.5.4. *Under the conditions of Theorem 4.5.3, if we further have* $\lambda_1(\mathbf{\Sigma})/q\lambda_q(\mathbf{\Sigma}) = O(\lambda_1(\mathbf{K}))$, $\|\mathbf{\Sigma}\|_{\mathsf{F}} \log d = \mathrm{Tr}(\mathbf{\Sigma}) \cdot o(1)$, $\lambda_2(\mathbf{\Sigma})/\lambda_1(\mathbf{\Sigma})$ *is upper bounded by an absolute constant less than* 1*, and* $\lambda \asymp \lambda_1(\mathbf{K})\sqrt{\log d/n}$*, then*

$$|\sin\angle(\boldsymbol{u}_1(\mathbf{X}_1), \boldsymbol{u}_1(\mathbf{K}))| = O_P\bigg(s\sqrt{\frac{\log d}{n}}\bigg).$$

Corollary 4.5.4 is a direct consequence of Theorem 4.5.3 and Theorem 4.3.5, and its proof is omitted. We then turn to study the estimation error of $\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}})$. By examining Theorem 4 in Yuan and Zhang (2013), for theoretical guarantee of fast rate of convergence, it is enough to show that $(\boldsymbol{v}^{(0)})^T\boldsymbol{u}_1(\mathbf{K})$ is lower bounded by an absolute constant larger than zero. In the next theorem, we show, under mild conditions, this is true with high probability, and accordingly we can exploit the result in Yuan and Zhang (2013) to show that $\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}})$ attains the same optimal convergence rate as that of $\boldsymbol{u}_{1,s}(\widehat{\mathbf{K}})$.

Theorem 4.5.5. *Under the conditions of Corollary 4.5.4, let* $J_0 := \{j : |(\boldsymbol{u}_1(\mathbf{K}))_j| = \Omega^0(s\log d/\sqrt{n})\}$*. Set* δ *in* (4.5.4) *to be* $\delta = C_2 s(\log d)/\sqrt{n}$ *for some positive absolute constant* C_2*. If* $s\sqrt{\log d/n} \to 0$*, and* $\|(\boldsymbol{u}_1(\mathbf{K}))_{J_0}\|_2 \geq C_3 > 0$ *is lower bounded by an absolute positive constant, then, with probability tending to 1,* $\|\boldsymbol{v}^{(0)}\|_0 \leq s$ *and* $|(\boldsymbol{v}^{(0)})^T\boldsymbol{u}_1(\mathbf{K})|$ *is lower bounded by* $C_3/2$*. Accordingly under the condition of Theorem 4 in Yuan and Zhang*

(2013), for $k \geq s$, we have

$$|\sin\angle(\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}}), \boldsymbol{u}_1(\mathbf{K}))| = O_P\Big(\sqrt{\frac{(k+s)\log d}{n}}\Big).$$

Remark 4.5.6. *Although a similar second step of truncation is performed, the assumption that the largest entries in $\boldsymbol{u}_1(\mathbf{K})$ satisfy $\|(\boldsymbol{u}_1(\mathbf{K}))_{J_0}\|_2 \geq C_3$ is much weaker than the assumption in Theorem 3.2 of Vu et al. (2013), because we allow a lot of entries in the leading eigenvector to be small and not detectable. This is possible since our aim is parameter estimation instead of guaranteeing the consistency in model selection.*

Remark 4.5.7. *In practice, we can adaptively select the tuning parameter k in Algorithm 2. One possible way is to use the criterion set up in Yuan and Zhang (2013), selecting k to maximize $(\widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}}))^T \cdot \widehat{\mathbf{K}}_{\mathrm{val}} \cdot \widehat{\boldsymbol{u}}_{1,k}^{\mathsf{FT}}(\widehat{\mathbf{K}})$, where $\widehat{\mathbf{K}}_{\mathrm{val}}$ is an independent empirical multivariate Kendall's tau statistic based on a separated sample set of the data. Yuan and Zhang (2013) showed such a heuristic performed quite well in applications.*

Remark 4.5.8. *In Corollary 4.5.4 and Theorem 4.5.5, we assume λ is in the same scale of $\lambda_1(\mathbf{K})\sqrt{\log d/n}$. In practice, λ is a tuning parameter. Here we can select λ using similar data driven estimation procedures as proposed in Lounici (2013b) and Wegkamp and Zhao (2013). The main idea is to replace the population quantities with their corresponding empirical versions in* (4.5.5)*. We hypothesize similar theoretical behaviors can be anticipated using a data driven way to select λ, as were shown in Lounici (2013b) and Wegkamp and Zhao (2013).*

4.6 Numerical Experiments

In this section we use both synthetic and real data to investigate the empirical usefulness of ECA. We use the FTPM algorithm described in Algorithm 2 for parameter estimation. To estimate more than one leading eigenvectors, we exploit the deflation method proposed in Mackey (2008). Here the cardinalities of the support sets of the leading eigenvectors are treated as tuning parameters. The following three methods are considered:

- TP: Sparse PCA method on the Pearson's sample covariance matrix;
- TCA: Transelliptical component analysis based on the transformed Kendall's tau covariance matrix shown in Equation (4.2.2);
- ECA: Elliptical component analysis based on the multivariate kendall's tau matrix.

For fairness of comparison, TCA and TP also exploit the FTPM algorithm, while using the Kendall's tau covariance matrix and Pearson's sample covariance matrix as the input matrix. The tuning parameter λ in (4.5.1) is selected using the method discussed in Remark 4.5.8, and the truncation value δ in (4.5.4) is selected such that $\|\boldsymbol{v}^{(0)}\|_0 \leq 10$.

4.6.1 Simulation Study

In this section, we conduct simulation study to back up the theoretical results and further investigate the empirical performance of ECA.

4.6.1.1 Dependence on Sample Size and Dimension

We first illustrate the dependence of the estimation accuracy of the sparse ECA estimator on the triplet (n, d, s). We adopt the data generating schemes of Yuan and Zhang (2013) and Han and Liu (2014b). More specifically, we first create a covariance matrix $\mathbf{\Sigma}$ whose first two eigenvectors $\boldsymbol{v}_j := (v_{j1}, \ldots, v_{jd})^T$ are specified to be sparse:

$$v_{1j} = \begin{cases} \frac{1}{\sqrt{10}} & 1 \leq j \leq 10 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad v_{2j} = \begin{cases} \frac{1}{\sqrt{10}} & 11 \leq j \leq 20 \\ 0 & \text{otherwise} \end{cases}.$$

Then we let $\mathbf{\Sigma}$ be $\mathbf{\Sigma} = 5\boldsymbol{v}_1\boldsymbol{v}_1^T + 2\boldsymbol{v}_2\boldsymbol{v}_2^T + \mathbf{I}_d$, where $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix. We have $\lambda_1(\mathbf{\Sigma}) = 6$, $\lambda_2(\mathbf{\Sigma}) = 3$, $\lambda_3(\mathbf{\Sigma}) = \ldots = \lambda_d(\mathbf{\Sigma}) = 1$. Using $\mathbf{\Sigma}$ as the covariance matrix, we generate n data points from a Gaussian distribution or a multivariate-t distribution with degrees of freedom 3. Here the dimension d varies from 64 to 256 and the sample size n varies from 10 to 500. Figure 4.1 plots the averaged angle distances $|\sin\angle(\widetilde{\boldsymbol{v}}_1, \boldsymbol{v}_1)|$ between the sparse ECA estimate $\widetilde{\boldsymbol{v}}_1$ and the true parameter $\boldsymbol{v}_1$, for dimensions $d = 64, 100, 256$, over 1,000 replications. In each setting, $s := \|\boldsymbol{v}_1\|_0$ is fixed to be a constant $s = 10$.

By examining the two curves in Figure 4.1 (A) and (B), the averaged distance between $\boldsymbol{v}_1$ and $\widetilde{\boldsymbol{v}}_1$ starts at almost zero (for sample size n large enough), and then transits to almost one as the sample size decreases (In other words, $1/n$ increases simultaneously.). Figure 4.1 reports all curves almost overlapped with each other when the averaged distances are plotted against $\log d/n$. This phenomenon confirms the results in Theorem 4.5.5. Consequently, the ratio $n/\log d$ acts as an effective sample size in controlling the prediction

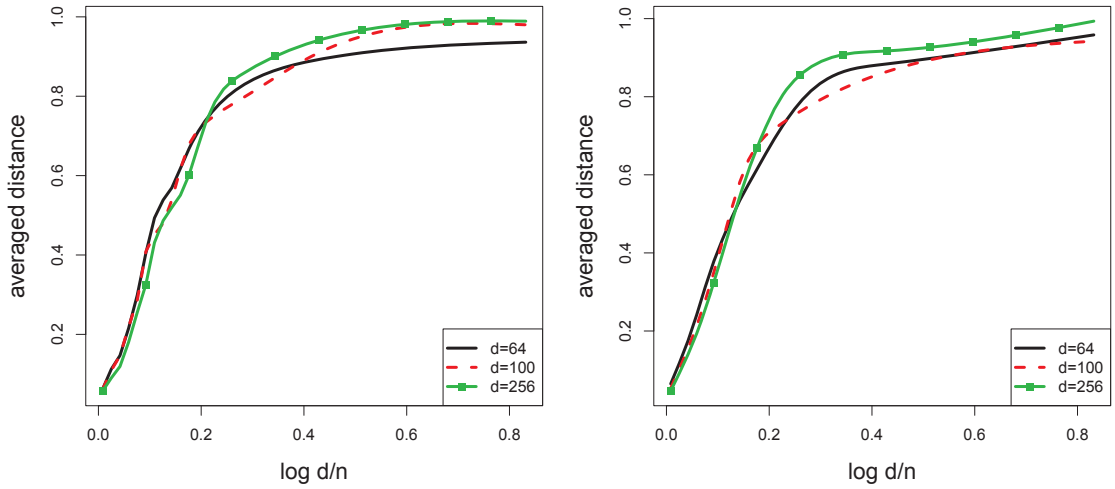


Figure 4.1: Simulation for two different distributions (normal and multivariate-t) with varying numbers of dimension d and sample size n. Plots of averaged distances between the estimators and the true parameters are conducted over 1,000 replications. (A) Normal distribution; (B) Multivariate-t distribution.

accuracy of the eigenvectors.

4.6.1.2 Estimating the Leading Eigenvector of the Covariance Matrix

We now focus on estimating the leading eigenvector of the covariance matrix $\mathbf{\Sigma}$. The first three rows in Table 4.2 list the simulation schemes of (n, d) and $\mathbf{\Sigma}$. In detail, let $\omega_1 > \omega_2 > \omega_3 = \ldots = \omega_d$ be the eigenvalues and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ be the eigenvectors of $\mathbf{\Sigma}$ with $\boldsymbol{v}_j := (v_{j1}, \ldots, v_{jd})^T$. The top m leading eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ of $\mathbf{\Sigma}$ are specified to be

sparse such that $s_j := \|\boldsymbol{v}_j\|_0$ is small and

$$v_{jk} = \begin{cases} 1/\sqrt{s_j}, & 1+\sum_{i=1}^{j-1} s_i \le k \le \sum_{i=1}^{j} s_i, \\ 0, & \text{otherwise.} \end{cases}$$

Accordingly, $\boldsymbol{\Sigma}$ is generated as

$$\boldsymbol{\Sigma} = \sum_{j=1}^{m} (\omega_j - \omega_d) \boldsymbol{v}_j \boldsymbol{v}_j^T + \omega_d \mathbf{I}_d.$$

Table 4.2 shows the cardinalities $s_1, \ldots, s_m$ and eigenvalues $\omega_1, \ldots, \omega_m$ and ω_d. In this section we set $m = 2$ (for the first three schemes) and $m = 4$ (for the later three schemes).

Table 4.2: Simulation schemes with different n, d and $\boldsymbol{\Sigma}$. Here the eigenvalues of $\boldsymbol{\Sigma}$ are set to be $\omega_1 > \ldots > \omega_m > \omega_{m+1} = \ldots = \omega_d$ and the top m leading eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ of $\boldsymbol{\Sigma}$ are specified to be sparse with $s_j := \|\boldsymbol{v}_j\|_0$ and $u_{jk} = 1/\sqrt{s_j}$ for $k \in [1 + \sum_{i=1}^{j-1} s_i, \sum_{i=1}^{j} s_i]$ and zero for all the others. $\boldsymbol{\Sigma}$ is generated as $\boldsymbol{\Sigma} = \sum_{j=1}^{m} (\omega_j - \omega_d) \boldsymbol{v}_j \boldsymbol{v}_j^T + \omega_d \mathbf{I}_d$. The column "Cardinalities" shows the cardinality of the support set of $\{\boldsymbol{v}_j\}$ in the form: "$s_1, s_2, \ldots, s_m, *, *, \ldots$". The column "Eigenvalues" shows the eigenvalues of $\boldsymbol{\Sigma}$ in the form: "$\omega_1, \omega_2, \ldots, \omega_m, \omega_d, \omega_d, \ldots$". In the first three schemes, m is set to be 2; In the second three schemes, m is set to be 4.

Scheme	n	d	Cardinalities	Eigenvalues
Scheme 1	50	100	$10, 10, *, *, \ldots$	$6, 3, 1, 1, 0, 0, \ldots$
Scheme 2	100	100	$10, 10, *, * \ldots$	$6, 3, 1, 1, 0, 0, \ldots$
Scheme 3	100	200	$10, 10, *, *, \ldots$	$6, 3, 1, 1, 0, 0, \ldots$
Scheme 4	50	100	$10, 8, 6, 5, *, *, \ldots$	$8, 4, 2, 1, 0.01, 0.01, \ldots$
Scheme 5	100	100	$10, 8, 6, 5, *, *, \ldots$	$8, 4, 2, 1, 0.01, 0.01, \ldots$
Scheme 6	100	200	$10, 8, 6, 5, *, *, \ldots$	$8, 4, 2, 1, 0.01, 0.01, \ldots$

We consider the following four different elliptical distributions:

(Normal) $\boldsymbol{X} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi_1 \cdot \sqrt{d/\mathbb{E}\xi_1^2})$ with $\xi_1 \overset{\mathrm{d}}{=} \chi_d$. Here χ_d is the chi-distribution

with degrees of freedom d. For $Y_1, \ldots, Y_d \overset{i.i.d.}{\sim} N(0,1)$,

$$\sqrt{Y_1^2 + \ldots + Y_d^2} \overset{\mathsf{d}}{=} \chi_d.$$

In this setting, $\boldsymbol{X}$ follows a Gaussian distribution (Fang et al., 1990).

(Multivariate-t) $\boldsymbol{X} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi_2 \cdot \sqrt{d/\mathbb{E}\xi_2^2})$ with $\xi_2 \overset{\mathsf{d}}{=} \sqrt{\kappa}\xi_1^*/\xi_2^*$. Here $\xi_1^* \overset{\mathsf{d}}{=} \chi_d$ and $\xi_2^* \overset{\mathsf{d}}{=} \chi_\kappa$ with $\kappa \in \mathbb{Z}^+$. In this setting, $\boldsymbol{X}$ follows a multivariate-t distribution with degrees of freedom κ (Fang et al., 1990). Here we consider $\kappa = 3$.

(EC1) $\boldsymbol{X} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi_3)$ with $\xi_3 \sim F(d, 1)$, i.e., ξ_3 follows an F-distribution with degrees of freedom d and 1 (Here ξ_3 has no finite mean. But ECA could still estimate the eigenvectors of the scatter matrix and is thus robust).

(EC2) $\boldsymbol{X} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi_4 \cdot \sqrt{d/\mathbb{E}\xi_4^2})$ with $\xi_4 \sim \mathrm{Exp}(1)$, i.e., ξ_4 follows an exponential distribution with the rate parameter 1.

We repeatedly generate n data points according to the schemes 1 to 3 and the four distributions discussed above for 1,000 times. To show the estimation accuracy, Figure 4.2 plots the averaged distances between the estimate $\widehat{\boldsymbol{v}}_1$ and the true $\boldsymbol{v}_1$, defined as $|\sin\angle(\widehat{\boldsymbol{v}}_1, \boldsymbol{v}_1)|$, against the numbers of estimated nonzero entries (defined as $\|\widehat{\boldsymbol{v}}_1\|_0$), for three different methods: TP,TCA, and ECA.

To show the feature selection results for estimating the support set of the leading eigenvector $\boldsymbol{v}_1$, Figure 4.3 plots the false positive rates against the true positive rates for the three different estimators under different schemes of (n, d), $\boldsymbol{\Sigma}$, and different distributions.

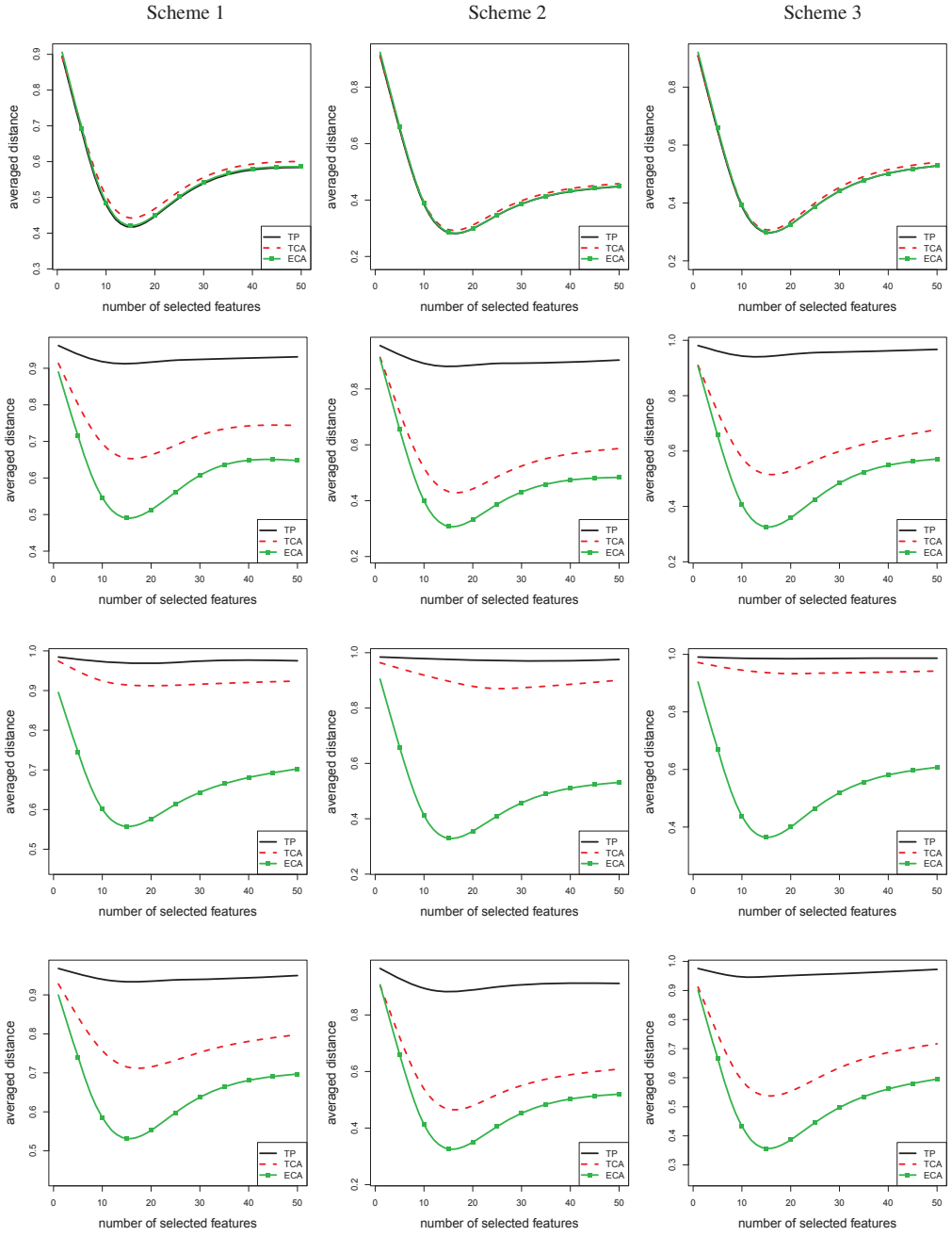


Figure 4.2: Curves of averaged distances between the estimates and true parameters for different schemes and distributions (normal, multivariate-t, EC1, and EC2, from top to bottom) using the FTPM algorithm. Here we are interested in estimating the leading eigenvector. The horizontal-axis represents the cardinalities of the estimates' support sets and the vertical-axis represents the averaged distances.

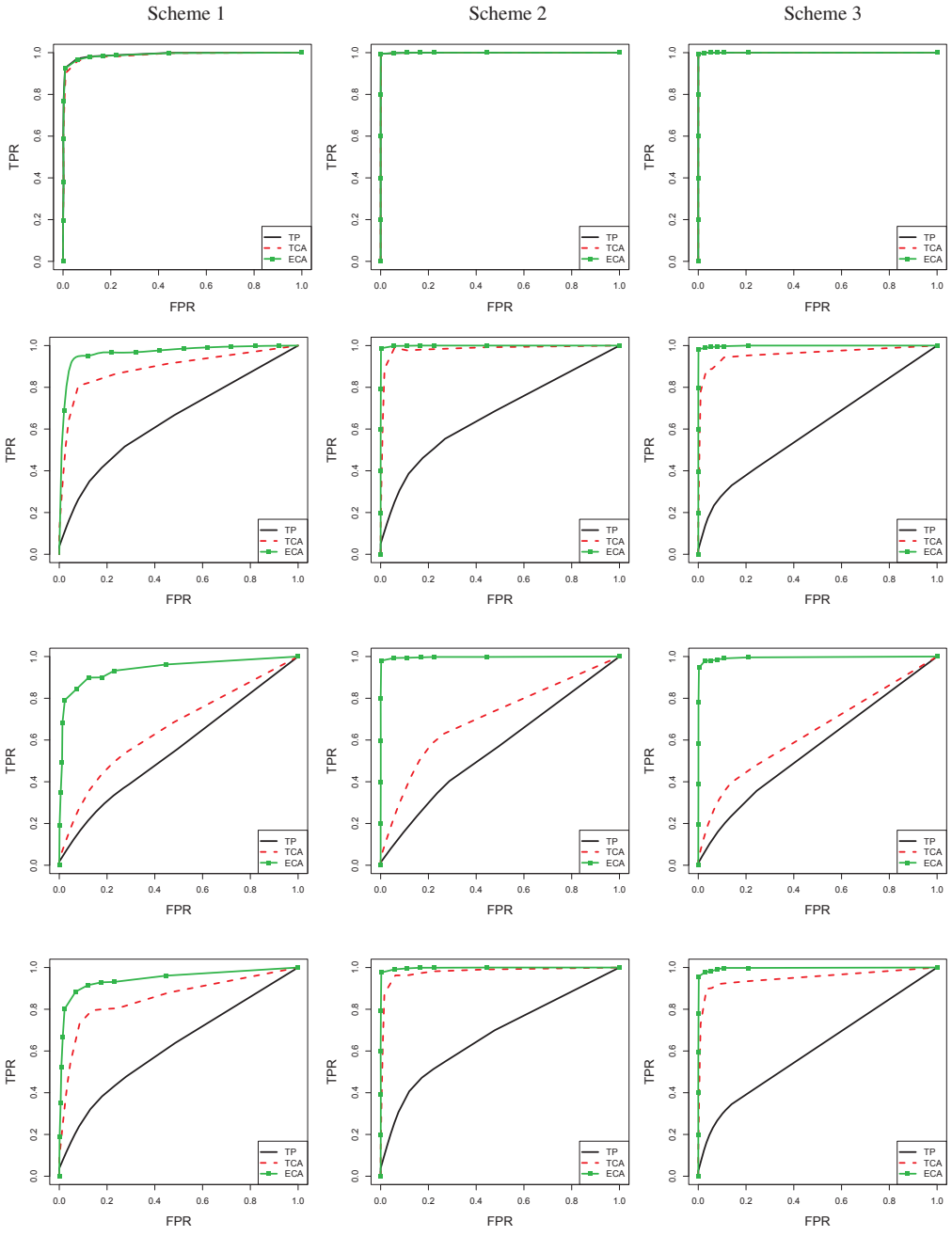


Figure 4.3: ROC curves for different methods in schemes 1 to 3 and different distributions (normal, multivariate-t, EC1, and EC2, from top to bottom) using the FTPM algorithm. Here we are interested in estimating the sparsity pattern of the leading eigenvector.

Figure 4.2 shows when the data are non-Gaussian but follow an elliptical distribution, ECA consistently outperforms TCA and TP in estimation accuracy. Moreover, when the data are indeed normal, there is no obvious difference between ECA and TP, indicating that ECA is a safe alternative to sparse PCA within the elliptical family. Furthermore, Figure 4.3 verifies that, in term of feature selection, the same conclusion can be drawn.

4.6.1.3 Estimating the Top m Leading Eigenvectors of the Covariance Matrix

Next, we focus on estimating the top m leading eigenvectors of the covariance matrix $\mathbf{\Sigma}$. We generate $\mathbf{\Sigma}$ in a similar way as in Section 4.6.1.2. We adopt the schemes 4 to 6 in Table 4.2 and the four distributions discussed in Section 4.6.1.2. We consider the case $m = 4$. We use the iterative deflation method and exploit the FTPM algorithm in each step to estimate the eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_4$. The tuning parameter remains the same in each iterative delation step.

As in the last section, Figure 4.4 plots the distances between the estimates $\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_4$ and the true parameters $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_4$ against the numbers of estimated nonzero entries. Here the distance is defined as $\sum_{j=1}^{4} |\sin \angle(\boldsymbol{v}_j, \widehat{\boldsymbol{v}}_j)|$ and the number is defined as $\sum_{j=1}^{4} \|\widehat{\boldsymbol{v}}_j\|_0$. We see the averaged distance starts from 4 and decreases first, then increases with the number of estimated nonzero entries. The minimum achieves when the number of nonzero entries is 40. The same conclusions drawn in the last section hold here, indicating ECA is a safe alternative to sparse PCA when the data are elliptically distributed.

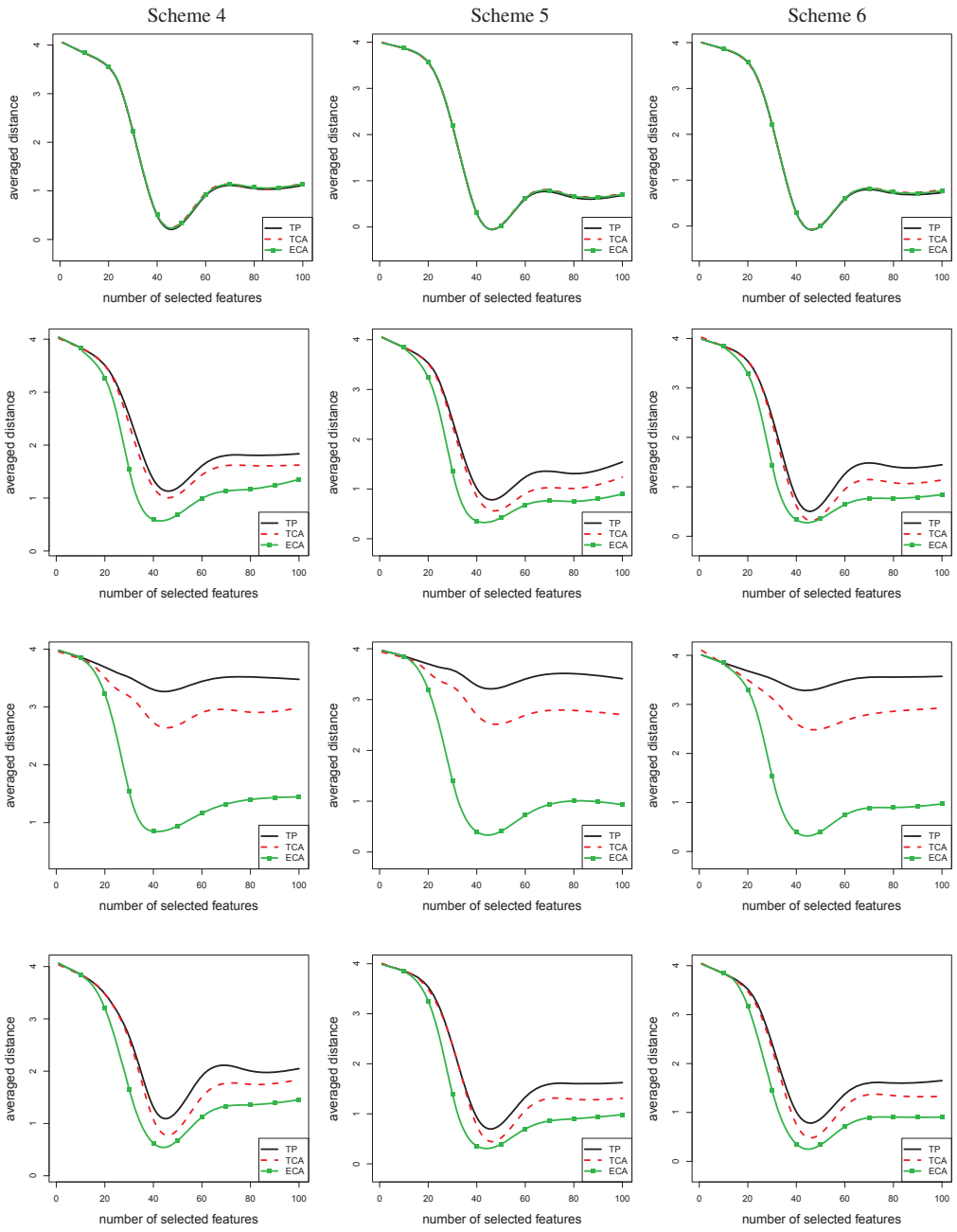


Figure 4.4: Curves of averaged distances between the estimates and true parameters for different methods in schemes 4 to 6 and different distributions (normal, multivariate-t, EC1, and EC 2, from top to bottom) using the FTPM algorithm. Here we are interested in estimating the top 4 leading eigenvectors. The horizontal-axis represents the cardinalities of the estimates' support sets and the vertical-axis represents the averaged distances.

4.6.2 Brain Imaging Data Study

In this section, we apply ECA and the other two methods to a brain imaging data obtained from the Autism Brain Imaging Data Exchange (ABIDE) project (Di Martino et al., 2013). The ABIDE project shares over 1,000 functional and structural scans from people with and without autism. This dataset includes 1,043 subjects, of which 544 are control and the rest are diagnosed with autism. Each subject is scanned for multiple time points, ranging from 72 to 290. The data were pre-processed to correct for the motion and eliminating noises. We refer to Di Martino et al. (2013) and Kang (2013) for more details in data preprocessing procedures.

Based on the 3D scans, we extract 116 regions of interest (Tzourio-Mazoyer et al., 2002) and broadly cover the brain. This gives us 1,043 matrices, each with 116 columns and number of rows ranging from 72 to 290. We then followed the idea in Eloyan et al. (2012) and Han et al. (2013) to compress the information of each subject by taking the median of each column for each matrix. In this study, we are mainly interested in studying the control group of subjects without autism. This gives a 544×116 matrix.

Table 4.3: Testing for normality of the ABIDE data. This table illustrates the number of voxels (out of a total number 116) rejecting the null hypothesis of normality at the significance level of 0.05 with or without Bonferroni's adjustment.

Critical value	**Kolmogorov-Smirnov**	**Shapiro-Wilk**	**Lilliefors**
0.05	88	115	115
0.05/116	61	113	92

First, we explore the obtained dataset to unveil several characteristics. In general, we find the observed data are non-Gaussian and marginally symmetric. We first illustrate the non-Gaussian issue. Table 4.3 provides the results of marginal normality tests. Here we conduct the three marginal normality tests at the significant level of 0.05. It is clear that at most 28 out of 116 voxels would pass any of three normality test. With Bonferroni correction there are still over half voxels failing to pass any normality tests. This indicates these imaging data are not Gaussian distributed.

We then show these data are marginally symmetric. For this, we first calculate the marginal skewness of each column in the data matrix. We then compare the empirical distribution function based on the marginal skewness values of the data matrix with that based on the simulated data from the standard Gaussian ($N(0, 1)$), t distribution with degree freedom 3 ($t(\text{df} = 3)$), t distribution with degree freedom 5 ($t(\text{df} = 5)$), and the exponential distribution with the rate parameter 1 ($\exp(1)$). Here the first three distributions are symmetric, and the exponential distribution is skewed to the right. Figure 4.5 plots the five estimated distribution functions. We see the distribution function for the marginal skewness of the imaging data is very close to that of the $t(\text{df} = 3)$ distribution . This indicates the data are marginally symmetric. Moreover, the distribution function based on the imaging data is far away from that based on the Gaussian distribution, indicating these data can be heavy-tailed.

The above data exploration results reveal the ABIDE data are non-Gaussian, symmetric, and heavy-tailed, which make elliptical distribution an very appealing way in modeling the

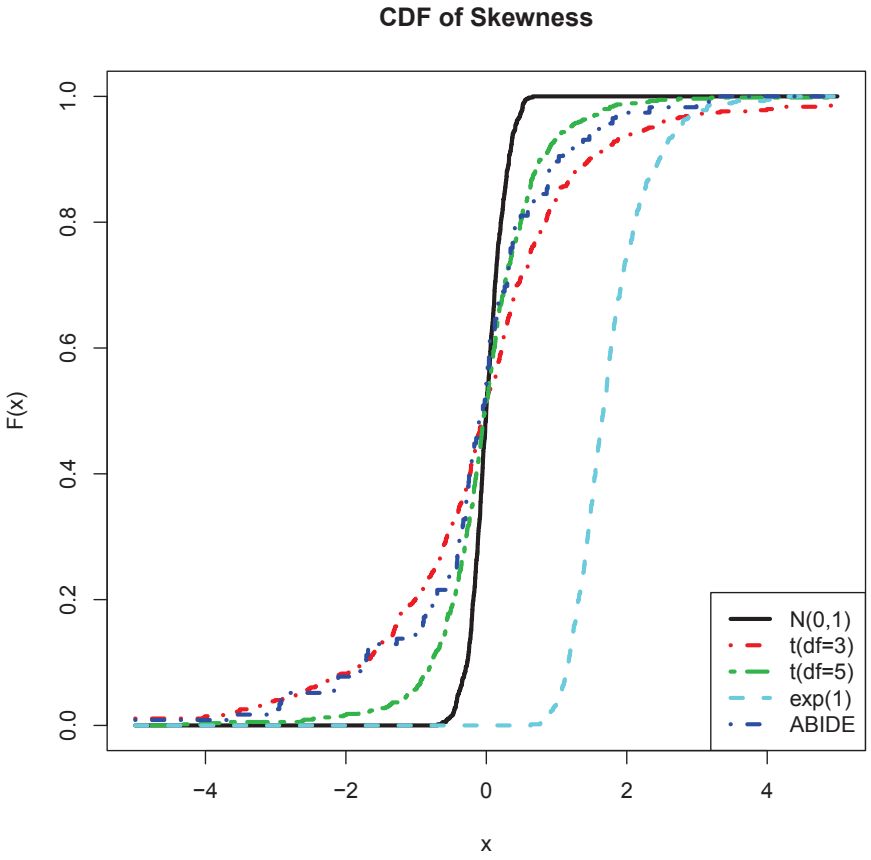


Figure 4.5: Illustration of the symmetric and heavy-tailed properties of the brain imaging data. The estimated cumulative distribution functions (CDF) of the marginal skewness based on the ABIDE data and four simulated distributions are plotted against each other.

data. We then apply TP, TCA and ECA to this dataset. We extract the top three eigenvectors and set the tuning parameter of the truncated power method to be 40. We project any two principal components of the ABIDE data into 2D plots, shown in Figure 4.6. Here the red dots represent the possible outliers that could have strong leverage influence. The leverage strength is defined as the diagonal values of the hat matrix in the linear model obtained by regressing the first principal component on the second one (Neter et al., 1996). High

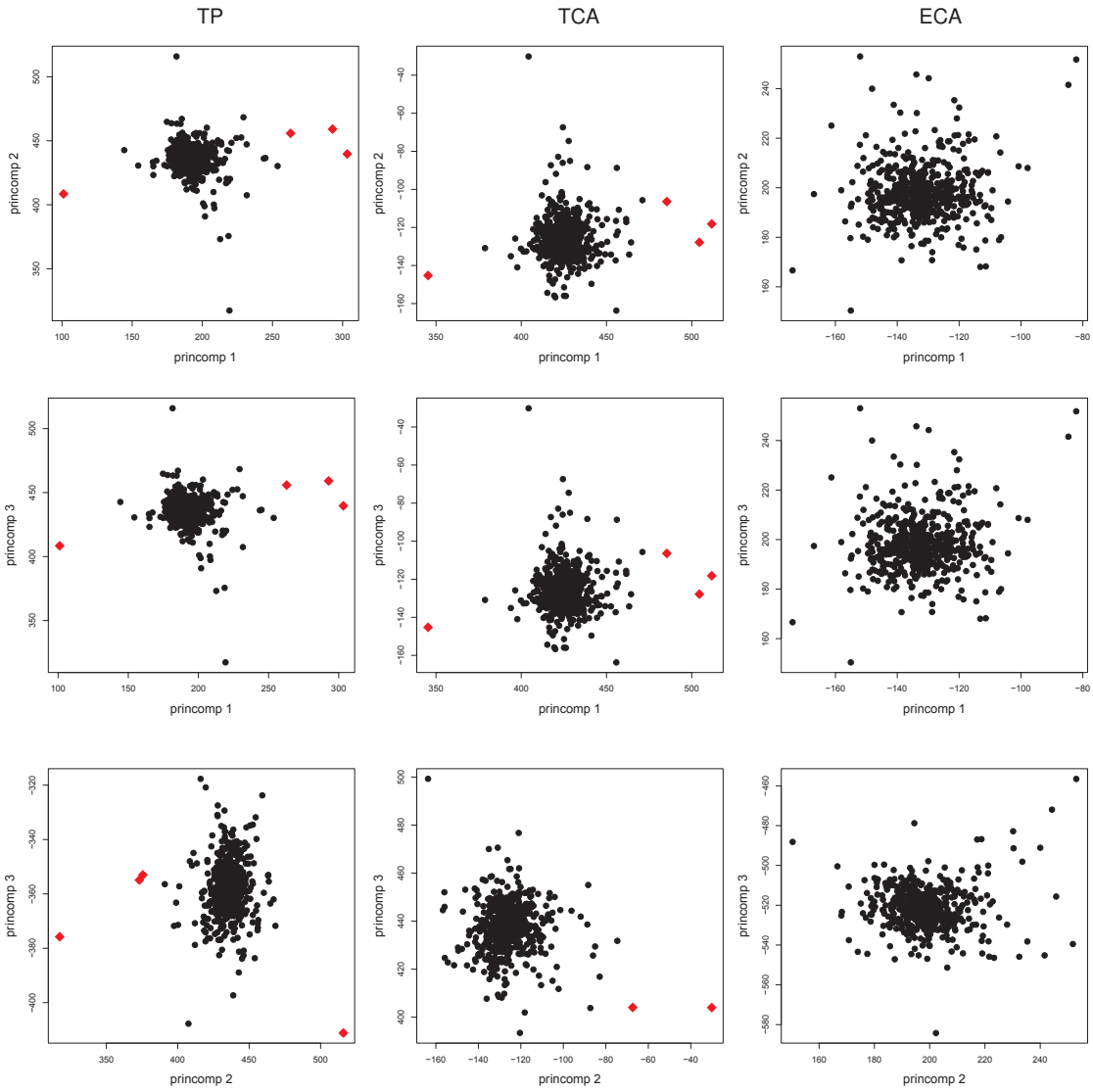


Figure 4.6: Plots of principal components 1 against 2, 1 against 3, 2 against 3 from top to bottom. The methods used are TP, TCA and ECA. Here red dots represent the points with strong leverage influence.

leverage strength means including these points will severely affect the linear regression estimates applied to principal components from the data. A data point is said to have strong leverage influence, if its leverage strength is higher than a chosen threshold value. Here we choose the threshold value to be $0.05(\approx 27/544)$.

It can be seen there are points with strong leverage influence for both statistics learnt by TP and TCA, while none for ECA (noted by red in Figure 4.6). This implies ECA has the potential to deliver better results for inference based on these estimated principal components.

4.7 Discussion

In this chapter, we propose elliptical component analysis (ECA) for estimating the eigenspace of the covariance matrix within the elliptical family, and study the statistical properties of ECA. For handling heavy tailed distributions, we focus on the multivariate Kendall's tau statistic. In the previous sections, we provided theoretical results to justify the use of both ECA and sparse ECA in analyzing high dimensional elliptical data. Table 4.1 summaries the theoretical performances of (sparse) ECA and provides comparisons to (sparse) PCA and (sparse) TCA. It could be observed that (sparse) ECA, although built on a significantly larger distribution family than the Gaussian, maintains similar statistical properties as (sparse) PCA under the subgaussian model in various settings.

Existing theory of multivariate Kendall's tau has been confined in the low dimensional settings (The dimension d is fixed). See for example, Marden (1999), Croux et al. (2002), and Jackson and Chen (2004). Instead, the ECA theory is for high dimensional regimes (where d could be even larger than n). Two other related methods — TCA and Copula Component Analysis (COCA) — have been recently proposed by Han and Liu (2014b)

and Han and Liu (2014a), in which they studied the performance of marginal rank-based statistics (including Kendall's tau and Spearmans' rho) on the transelliptical and nonparanromal models. These methods can efficiently estimate the leading eigenvectors of the (latent) correlation matrix. However, ECA is fundamentally different from TCA and COCA in the following aspects.

On one hand, the main differences between ECA and TCA include: (i) TCA can only estimate the leading eigenvectors of the correlation matrix, while ECA estimates the leading eigenvectors of the covariance matrix; (ii) Unlike ECA, TCA cannot estimate the principal components; (iii) ECA has a theoretically guaranteed faster rate of convergence compared to TCA under the elliptical model.

On the other hand, the differences between ECA and COCA include: (i) COCA assumes the nonparanormal model while ECA assumes the elliptical model. In fact, the only distribution that is within the nonparanormal and elliptical families is Gaussian (Liu et al., 2012b); (ii) To estimate the leading eigenvectors of the covariance matrix, COCA requires a strong marginal subgaussian assumption, which is not needed for ECA.

Lastly, compared to Chapters 2 and 3, Chapter 4 discusses an alternative method, ECA, on an alternative elliptical model. Under the elliptical model, we show ECA, built on the nonparametric multivariate rank statistic, is a minimax optimal procedure for conducting principal component analysis. Thusly, this further strengthens our main idea: Semiparametric modeling coupled with nonparametric methods could be an appealing approach for tacking high dimensional complex data.

Chapter 5

Distribution-Free Tests of Independence with Applications to Testing More Structures

5.1 Introduction

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independent observations of a continuous random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$. We aim to test the null hypothesis:

$$\mathbf{H_0}: \ X_1, X_2, \ldots, X_d \text{ are mutually independent.} \tag{5.1.1}$$

This problem plays a fundamental role in many fields, including the false discovery rate (FDR) control (Benjamini and Hochberg, 1995), naive Bayes classification (Tibshirani et al., 2002; Fan and Fan, 2008), and independent component analysis (Comon, 1994).

The problem of testing (5.1.1) has been intensively studied when $\boldsymbol{X}$ is multivariate Gaussian. In low dimensions, major methods include the likelihood ratio test (LRT) (Anderson, 2003), Roy's largest root test (Roy, 1957), and Nagao's test (Nagao, 1973). They test Pearson's covariance matrix $\boldsymbol{\Sigma}$ or correlation matrix $\mathbf{R}$ to be the identity $\mathbf{I}_d$ matrix using their sample counterparts. When $d \to \infty$ as $n \to \infty$ and $d/n \not\to 0$, the classic LRTs suffer poor performance due to the inconsistency of the eigenvalues of sample covariance matrix to their population quantities (Bai and Yin, 1993). This observation motivates a variety of works in the high dimensional settings, summarized as below.

- When $d/n \to \gamma \in (0, 1]$, Bai et al. (2009) and Jiang and Yang (2013) proposed corrected LRT statistics and proved their asymptotic normality[1]. Johnstone (2001) and Bao et al. (2012) proved the Tracy-Widom law for the null limiting distributions

[1] Bai et al. (2009) only considered the regime $\gamma \in (0, 1)$. Jiang et al. (2012) proved it covers the extreme $\gamma = 1$. Of note, for testing $\boldsymbol{\Sigma} = \mathbf{I}_d$ and $\mathbf{R} = \mathbf{I}_d$, we require $d < n$ and $d \leq n - 5$ separately.

of the Roy's largest root test statistics[2].

- When $d/n \to \gamma \in (0, \infty)$, Ledoit and Wolf (2002) and Schott (2005) proposed corrected Nagao's test statistics and prove their asymptotic normality. Jiang (2004) proposed a test statistic based on the maximum magnitude of the Pearson's sample correlation coefficients and show it converges to an extreme value type I distribution. With some adjustments, Birke and Dette (2005) and Cai and Jiang (2012) proved the tests in Ledoit and Wolf (2002) and Jiang (2004) are extendable to the case when $\gamma = \infty$. To the best of our knowledge, there is no result generalizing the test in Schott (2005) to the regime $\gamma = \infty$.

- When $d/n \in [0, \infty]$ with the limit of d/n possibly not existing, Srivastava (2006) proposed a corrected LRT using only nonzero sample eigenvalues. Srivastava (2005) introduced a test using unbiased estimators of the covariance matrix's higher powers' traces. Cai et al. (2014b) studied the test in Chen et al. (2010) and showed it uniformly dominates the corrected LRT tests in Bai et al. (2009) and Jiang and Yang (2013). The aforementioned three test statistics are asymptotically normal. Zhou (2007) modified Jiang (2004)'s test and show that the null limiting distribution of the test statistic is an extreme value type I distribution.

Most of the aforementioned tests are designed only under the Gaussian assumption. Testing (5.1.1) in high dimensions for nonGaussian data is not as well studied as for the

[2]Bao et al. (2012)'s result is only valid when $\gamma \in (0, 1)$, while the result in Johnstone (2001) applies to the case $\gamma = 1$. Their results are further generalized to $\gamma > 1$ in Péché (2009) and Pillai and Yin (2012) with possibly nonGaussian observations.

Gaussian data. Péché (2009) and Pillai and Yin (2012) studied Roy's largest root test for sub-Gaussian data. Bao et al. (2013) studied the Spearman's rho statistic. Jiang (2004) studied the largest off-diagonal entry in the Pearson's sample correlation matrix. In particular, Jiang (2004) showed, for testing a simplified null hypothesis of (5.1.1):

$$\mathbf{H_0'}: \ X, X_1, \ldots, X_d \text{ are independently and identically distributed}, \tag{5.1.2}$$

the Gaussian assumption could be relaxed to a moment assumption $\mathbb{E}|X|^r < \infty$ for some $r > 30$. Later, Zhou (2007) modified Jiang (2004)'s test and relaxed the moment assumption to be $r = 6$. For more advances in this track, we refer to Li and Rosalsky (2006), Zhou (2007), Liu et al. (2008), Li et al. (2010), Cai and Jiang (2011), Cai and Jiang (2012), Shao and Zhou (2014), among others.

In this work, we investigate testing (5.1.1) in high dimensions. The asymptotic regime of interest is $d, n \to \infty$ and $d/n \in [0, \infty]$. Our main focus is on nonparametric rank-based tests and their optimality. The major contributions of this work are in threefold. First, we consider a large family of rank-based test statistics including the Spearman's rho (Spearman, 1904) and Kendall's tau (Kendall, 1938) statistics. We prove they all converge weakly to an extreme value type I distribution. Secondly, we provide power analysis and optimality of the propose tests against the sparse alternative (Explicit definition is in Section 5.4). In particular, we show the tests based on Spearman's rho and Kendall's tau are rate optimal. Thirdly, we generalize these results to testing m-dependence and data homogeneity, for

which we propose new rank-based tests and show these tests are rate optimal against certain alternatives. Techniques in Arcones and Gine (1993) and Zaïtsev (1987) for studying the tails of U-statistics and approximation of summations of independent random vectors to the Gaussian are key ingredients in the analysis. We also discuss approximating the exact distributions of the test statistics for accelerating the rate of convergence.

5.1.1 Other Related Work

Testing (5.1.1) is related to testing bivariate independence. For testing independence of two random variables, Hotelling and Pabst (1936) and Kendall (1938) proposed using the Spearman's rho and Kendall's tau statistics, and Hoeffding (1948b) proposed a modified Cramér-von Mises type statistic, the Hoeffding's D statistic. For testing independence of two random vectors (with possibly very large dimensions d_1 and d_2), Bakirov et al. (2006), Székely and Rizzo (2013), and Jiang et al. (2013) proposed tests based on spatial signs, distance correlations, and modified likelihood ratios. However, we cannot directly apply these results to test (5.1.1) without making multivariate adjustment. For testing (5.1.1), a notable alternative to Pearson's correlation coefficient is Spearman's rho (Zhou, 2007). More specifically, Zhou (2007) established the limiting distribution of the largest off-diagonal entry of the Spearman's rho correlation matrix but without power analysis. This work includes the results in Zhou (2007) as a special case.

5.1.2 Chapter Organization

In Section 5.2, we introduce the proposed families of tests of (5.1.1). In Section 5.3, we prove the proposed test statistics converge weakly to an extreme value type I distribution. In Section 5.4, we give power analysis and present the optimality properties of the proposed tests. We present the numerical results in Section 5.5. In Section 5.6, we generalize the results to testing more hypotheses and propose rate optimal rank-based tests against certain alternatives. In Section 5.6, we also propose new methods which achieve improved rates of convergence. In Section 5.7, we summarize the results and discuss the relevant work.

5.2 Testing Procedures

In this section, we introduce nonparametric rank-based tests of (5.1.1). Suppose we observe n independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of a d-dimensional continuous[3] random vector $\boldsymbol{X} \in \mathbb{R}^d$. We write $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$. For any two entries $j \neq k \in \{1, \ldots, d\}$, let Q_{ni}^j be the rank of X_{ij} in $\{X_{1j}, \ldots, X_{nj}\}$ and Q_{ni}^k be the rank of X_{ik} in $\{X_{1k}, \ldots, X_{nk}\}$. We denote $\{R_{ni}^{jk}, i = 1, \ldots .n\}$ to be the relative ranks of the k-th entry corresponding to the j-th entry, satisfying

$$R_{ni}^{jk} = Q_{ni'}^k \text{ subject to } Q_{ni'}^j = i, \text{ for } i = 1, \ldots, n.$$

We propose two families of nonparametric tests based on the relative ranks. The first is

[3]We pose the assumption of continuity thereafter to avoid the possible ties in the data.

based on the simple linear rank statistics of the form:

$$T_{jk} := \sum_{i=1}^{n} c_{ni} g(R_{ni}^{jk}/(n+1)). \tag{5.2.1}$$

Here $\{c_{ni}, i = 1, \ldots, n\}$ are an array of constants called the regression constants. $g(\cdot)$ is a Lipchitz function and called the score function. For avoiding the trivial case, we assume $\sum_i c_{ni}^2 > 0$. For accommodating tests of independence, we further pose the alignment assumption:

$$c_{ni} = n^{-1} \cdot f(i/(n+1)), \text{ where } f(\cdot) \text{ is a Lipchitz function.} \tag{5.2.2}$$

Under this assumption, the corresponding simple linear rank statistic is a general measurement of agreements in the ranks between two sequences of values[4]. Spearman's rho is in the family of simple linear rank statistics.

The second is based on the rank type U-statistics. A rank type U-statistic is a function of relative ranks $\{R_{ni}^{jk}, 1 \le i \le n\}$ and at the same time a U-statistic with order $m < n$:

$$U_{jk} := \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{i_1 \neq i_2 \neq \cdots \neq i_m} h(\boldsymbol{X}_{i_1,\{j,k\}}, \ldots, \boldsymbol{X}_{i_m,\{j,k\}}), \tag{5.2.3}$$

only depending on $\left\{R_{ni}^{jk}\right\}_{i=1}^{n}$. Here for any vector $\boldsymbol{X}_i$ and some set $\mathcal{A} \subset \{1, \ldots, d\}$, we denote $\boldsymbol{X}_{i,\mathcal{A}}$ to be the sub-vector of $\boldsymbol{X}_i$ with entries indexed by $\mathcal{A}$. The kernel function

[4]The alignment assumption in (5.2.2) is not required in deriving the null limiting distribution. However, they play an important role in power analysis.

$h(\cdot) : \underbrace{\mathbb{R}^2 \otimes \cdots \otimes \mathbb{R}^2}_{m} \to \mathbb{R}$ is assumed to be bounded, but not necessarily symmetric. Because our focus is on measuring correlation instead of covariance, the boundedness assumption is mild.

First, we focus on the simple linear rank statistics. Hotelling and Pabst (1936) first introduced the family of simple linear rank statistics for testing homogeneity. Wald and Wolfowitz (1940) proved they are asymptotically normal. Some more recent developments in proving the moderate deviation (sometimes referred to as Cramér's large deviation) properties of the simple linear rank statistics are in Kallenberg (1982), Vandemaele and Veraverbeke (1982), Seoh et al. (1985), and Inglot (2012).

Of note, under (5.1.1) the distribution of T_{jk} is irrelevant to the specific distribution of $\boldsymbol{X}$. Accordingly, we can analytically calculate the mean and variance of T_{jk} without any prior knowledge about the data. Let $\mathbb{E}_{\mathbf{H_0}}(\cdot)$ and $\mathrm{Var}_{\mathbf{H_0}}(\cdot)$ be the expectation and variance of a certain statistic under $\mathbf{H_0}$. We have

$$\mathbb{E}_{\mathbf{H_0}} T_{jk} = \bar{g}_n \sum_{i=1}^{n} c_{ni} \text{ and } \mathrm{Var}_{\mathbf{H_0}} T_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \Big(g(i/(n+1)) - \bar{g}_n \Big)^2 \cdot \sum_{i=1}^{n} (c_{ni} - \bar{c}_n)^2, \tag{5.2.4}$$

where $\bar{g}_n := \frac{1}{n} \sum_{i=1}^{n} g(i/(n+1))$ is the sample mean of $\{g(R_{ni}^{jk}/(n+1)), i = 1, \ldots, n\}$. Based on $\{T_{jk}, 1 \le j < k \le d\}$, we propose the following extreme value statistic for

testing (5.1.1):

$$L_n := \max_{j<k} |T_{jk} - \mathbb{E}_{\mathbf{H_0}} T_{jk}|. \tag{5.2.5}$$

Secondly, we focus on the rank type U-statistics. They belong to the general family of U-statistics (Serfling, 2002). Similar to simple linear rank statistics, we can calculate $\mathbb{E}_{\mathbf{H_0}} U_{jk}$ and $\mathrm{Var}_{\mathbf{H_0}} U_{jk}$. We test (5.1.1) using

$$\widetilde{L}_n := \max_{j<k} |U_{jk} - \mathbb{E}_{\mathbf{H_0}} U_{jk}|. \tag{5.2.6}$$

Detailed studies of L_n and $\widetilde{L}_n$'s null limiting distributions are in Section 5.3. We leave the theoretical results until then, but give some intuition here. Under some regularity conditions, the standardized version of T_{jk} (U_{jk}) is asymptotically normal. Accordingly, the standardized version of L_n^2 ($\widetilde{L}_n^2$) is asymptotically "close" to the maximum of $d(d-1)/2$ independent chi-square distributed random values with degree of freedom 1. The later converges weakly to an extreme value type I distribution after certain adjustment.

Let σ_T^2 and σ_U^2 be the variances of $\sqrt{n}T_{jk}$ and $\sqrt{n}U_{jk}$ under (5.1.1):

$$\sigma_T^2 := n\mathrm{Var}_{\mathbf{H_0}} T_{jk} \ \text{ and } \ \sigma_U^2 := n\mathrm{Var}_{\mathbf{H_0}} U_{jk}. \tag{5.2.7}$$

We propose the size α tests T_α and $\widetilde{\mathsf{T}}_\alpha$ as follows:

$$\begin{aligned}\mathsf{T}_\alpha &:= I(nL_n^2/\sigma_T^2 - 4\log d + \log\log d \geq q_\alpha) \text{ and} \\ \widetilde{\mathsf{T}}_\alpha &:= I(n\widetilde{L}_n^2/\sigma_U^2 - 4\log d + \log\log d \geq q_\alpha). \end{aligned} \tag{5.2.8}$$

Here $I(\cdot)$ represents the indicator function and

$$q_\alpha := -\log 8\pi - 2\log\log(1-\alpha)^{-1} \tag{5.2.9}$$

is the $1-\alpha$ quantile of the extreme value type I distribution with the distribution function $\exp(-\exp(-y/2)/\sqrt{8\pi})$[5]. The null hypothesis is rejected if T_α (or $\widetilde{\mathsf{T}}_\alpha$) returns value one.

In the following, we provide four examples of distribution-free tests of independence. They are based on either simple linear rank or rank type U-statistics.

Example 5.2.1 (Spearman's rho)**.** *Remind that Q_{ni}^j and Q_{ni}^k are the ranks of X_{ij} and X_{ik} among $\{X_{1j},\ldots,X_{nj}\}$ and $\{X_{1k},\ldots,X_{nk}\}$. The Spearman's rho correlation coefficient is defined as*

$$\rho_{jk} :== \frac{12}{n(n^2-1)}\sum_{i=1}^{n}\Big(i-\frac{n+1}{2}\Big)\Big(R_{ni}^{jk}-\frac{n+1}{2}\Big), \tag{5.2.10}$$

where $\bar{Q}_n^j = \bar{Q}_n^i := (n+1)/2$. It follows Spearman's rho is a simple linear rank statistic.

[5]In practice we can conduct simulation to approximate the exact distribution of $nL_n^2/\sigma_T^2 - 4\log d + \log\log d$ and choose q_α to be the $1-\alpha$ quantile of the corresponding empirical distribution. This is a simulation-based approach to select the threshold value. Section 5.6.2 discusses the details of this approach.

In particular, using (5.2.4)*, we have*

$$\mathbb{E}_{\mathbf{H_0}}\rho_{jk} = 0 \text{ and } \mathrm{Var}_{\mathbf{H_0}}\rho_{jk} = 1/(n-1).$$

Accordingly, the proposed test statistic based on Spearman's rho is:

$$\mathsf{T}^{\rho}_{\alpha} = I((n-1)\max_{j<k}\rho^2_{jk} - 4\log d + \log\log d \geq q_{\alpha}).$$

Example 5.2.2 (Kendall's tau)**.** *The Kendall's tau correlation coefficient is defined as*

$$\tau_{jk} := \frac{2}{n(n-1)}\sum_{i<i'}\mathrm{sign}(X_{i'j} - X_{ij})\mathrm{sign}(X_{i'k} - X_{ik}) = \frac{2}{n(n-1)}\sum_{i<i'}\mathrm{sign}(R^{jk}_{ni'} - R^{jk}_{ni}),$$

where the sign function $\mathrm{sign}(\cdot)$ *is defined as* $\mathrm{sign}(x) := x/|x|$ *with the convention* $0/0 = 0$*. The Kendall's tau statistic is a function of the relative ranks* $\{R^{jk}_{ni}, i = 1, \ldots, n\}$ *and a U-statistics with a bounded kernel function. Accordingly, Kerdall's tau is an rank type U-statistic. Moreover, we have*

$$\mathbb{E}_{\mathbf{H_0}}\tau_{jk} = 0 \quad \text{and} \quad \mathrm{Var}_{\mathbf{H_0}}\tau_{jk} = \frac{2(2n+5)}{9n(n-1)}.$$

Accordingly, the proposed test statistic based on Kendall's tau is:

$$\mathsf{T}^{\tau}_{\alpha} = I\left(\frac{9n(n-1)}{2(2n+5)}\max_{j<k}\tau^2_{jk} - 4\log d + \log\log d \geq q_{\alpha}\right).$$

Example 5.2.3 (A major part of Spearman's rho)**.** *Spearman's rho is not a U-statistic. But by Hoeffding (1948a), we can write*

$$\rho_{jk} = \frac{n-2}{n+1} \cdot \underbrace{\frac{3}{n(n-1)(n-2)} \sum_{i \neq i' \neq i''} \text{sign}(X_{ij} - X_{i'j})\text{sign}(X_{ik} - X_{i''k})}_{\widehat{\rho}_{jk}} + \frac{3\tau_{jk}}{n+1}. \tag{5.2.11}$$

$\widehat{\rho}_{jk}$ is a U-statistic with degree 3, an asymmetric bounded kernel function, and

$$\mathbb{E}_{\mathbf{H_0}}(\widehat{\rho}_{jk}) = 0 \quad \text{and} \quad \text{Var}_{\mathbf{H_0}}\widehat{\rho}_{jk} = \frac{n^2-3}{n(n-1)(n-2)}.$$

We propose the test based on $\{\widehat{\rho}_{jk}, 1 \leq j < k \leq d\}$ as

$$\mathsf{T}_\alpha^{\widehat{\rho}} = I\left(\frac{n(n-1)(n-2)}{n^2-3} \max_{j<k} \widehat{\rho}_{jk}^2 - 4\log d + \log\log d \geq q_\alpha\right).$$

Example 5.2.4 (Projection of Kendall's tau to simple linear rank statistics)**.** *Kendall's tau does not belong to the family of simple linear rank statistics. However, by the projection argument in Hájek (1968), τ_{jk} can be approximated by the following simple linear rank statistic:*

$$\widehat{\tau}_{jk} = \frac{8}{n^2(n-1)} \sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)\left(R_{ni}^{jk} - \frac{n+1}{2}\right).$$

Using the variance of ρ_{jk} and relation between ρ_{jk} and $\widehat{\tau}_{jk}$, it is easy to obtain

$$\mathbb{E}_{\mathbf{H_0}}\widehat{\tau}_{jk} = 0 \quad \text{and} \quad \text{Var}_{\mathbf{H_0}}\widehat{\tau}_{jk} = \frac{4(n+1)^2}{9n^2(n-1)}^6,$$

The proposed test statistic based on $\{\widehat{\tau}_{jk}, 1 \leq j < k \leq d\}$ then is

$$\mathsf{T}_{\alpha}^{\widehat{\tau}} = I\left(\frac{9n^2(n-1)}{4(n+1)^2}\max_{j<k}\widehat{\tau}_{jk}^2 - 4\log d + \log\log d \geq q_{\alpha}\right).$$

Remark 5.2.5. *In this section, we consider two families of test statistics: the family of simple linear rank statistics and the family of rank type U-statistics. When the sample size n is small or large, Waerden (1957) and Woodworth (1970) separately studied the performance of Spearman's rho and Kendall's tau in testing bivariate independence under Gaussian assumption. They showed that Spearman's rho is more favorable than Kendall's tau when n is small, while the reverse is true if n is large. Accordingly, the advantage of one over the other is determined on a case-by-case basis.*

5.3 Limiting Null Distributions

In this section, we characterize the limiting distributions of L_n and $\widetilde{L}_n$ under (5.1.1). We start with an introduction of some necessary notation. Let $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$ be a d-dimensional vector and $\mathbf{M} = [\mathbf{M}_{jk}] \in \mathbb{R}^{d\times d}$ be a d by d square matrix. For any sets

[6] We have $\text{Var}_{\mathbf{H_0}}(\widehat{\tau}_{jk})/\text{Var}_{\mathbf{H_0}}(\tau_{jk})$ goes to 1 as n goes to infinity, indicating $\widehat{\tau}_{jk}$ is asymptotically equivalent to τ_{jk} under the null hypothesis $\mathbf{H_0}$.

$I, J \subset \{1, \ldots, d\}$, let $\boldsymbol{v}_I$ be the sub-vector of $\boldsymbol{v}$ with entries indexed by I, and $\mathbf{M}_{I,J}$ be the sub-matrix of $\mathbf{M}$ with rows indexed by I and columns indexed by J. For $0 < q < \infty$, let $\|\boldsymbol{v}\|_q := (\sum |v_i|^q)^{1/q}$ be the vector L_q norm. Let $\|\mathbf{M}\|_q := \sup_{\boldsymbol{v}} \|\mathbf{M}\boldsymbol{v}\|_q / \|\boldsymbol{v}\|_q$ be the matrix operator q-norm and let $\|\mathbf{M}\|_{\max} := \max_{jk} |\mathbf{M}_{jk}|$ be the matrix elementwise maximum norm. Let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the largest and smallest eigenvalues of $\mathbf{M}$. For two sequences of numbers $\{a_1, a_2, \ldots\}$ and $\{b_1, b_2, \ldots\}$, we write $a_n = O(b_n)$ if we have $|a_n| \leq C|b_n|$ for some positive generic constant C and all sufficiently large n. We write $a_n = o(b_n)$ if for any positive constant c, for all sufficient large n, $|a_n| \leq c|b_n|$. We write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. For any two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, we write $\boldsymbol{X} \overset{D}{=} \boldsymbol{Y}$ if they are identically distributed. We study the asymptotics of triangular arrays (Greenshtein and Ritov, 2004) and allow the dimension d_n to increase with n. We write d to be the abbreviation of d_n. Throughout the chapter, c and C represent generic absolute positive constants. The actual values of c and C may vary at different locations.

Let's first consider the simple linear rank statistic T_{jk}. The next theorem shows, under (5.1.1) and some regularity conditions for the regression constants $\{c_{n1}, \ldots, c_{nn}\}$, but without any assumption on $\boldsymbol{X}$, $nL_n^2/\sigma_T^2 - 4\log d + \log\log d$ converges weakly to an extreme value type I distribution.

Theorem 5.3.1 (Simple linear rank statistics)**.** *Suppose the simple linear rank statistics* $\{T_{jk}, 1 \leq j < k \leq d\}$ *take the form* (5.2.1)*, the regression constants* $\{c_{n1}, \ldots, c_{nn}\}$

satisfy:

$$\max_{1\le i\le n}|c_{ni}-\bar{c}_n| \le C_1 n^{-1/2}\Big(\sum_{i=1}^{n}(c_{ni}-\bar{c}_n)^2\Big)^{1/2},$$
$$\Big|\sum_{i=1}^{n}(c_{ni}-\bar{c}_n)^3\Big| \le C_2 n^{-1/2}\Big(\sum_{i=1}^{n}(c_{ni}-\bar{c}_n)^2\Big)^{3/2}, \tag{5.3.1}$$

where $\bar{c}_n := \sum_{i=1}^{n} c_{ni}$ *represents the sample mean of the regression constants and* C_1, C_2 *are two absolute constants*[7]*, and the score function satisfies the Lipchitz condition*[8]*:*

$$g(\cdot) \text{ is differentiable, with bounded Lipchitz constant } \Delta < \infty. \tag{5.3.2}$$

We then have, under the regime $\log d = o(n^{1/3})$*,* $d \to \infty$*, and* (5.1.1)*, for any* $y \in \mathbb{R}$*,*

$$|\mathbb{P}(nL_n^2/\sigma_T^2 - 4\log d + \log\log d \le y) - \exp(-\exp(-y/2)/\sqrt{8\pi})| = o(1).$$

Here L_n *and* σ_T^2 *are separately defined in* (5.2.5) *and* (5.2.7)*.*

For testing (5.1.1), Theorem 5.3.1 gives a distribution-free type result. In other words, the asymptotic or nonasymptotic behavior of a testing procedure is irrelevant to the specific data distribution (Kendall and Stuart, 1977). In comparison, the tests based on the Pearson's sample covariance and correlation matrices (Jiang, 2004; Li et al., 2010; Cai and Jiang,

[7]Regularity conditions in the form (5.3.1) are common for the simple linear statistics to be asymptotically normal and moderately deviated from the Gaussian. We refer to Hájek et al. (1999) and Kallenberg (1982) for details. We also mention that Seoh et al. (1985) propose a similar (intrinsically the same) set of regularity conditions for $\{c_{ni}, 1 \le i \le n\}$.

[8]The Lipchitz condition rules out the normal score (Fisher-Yates) statistic where $g(\cdot)$ is proportional to $\Phi^{-1}(\cdot/(n+1))$. Here $\Phi^{-1}(\cdot)$ represents the quantile function of the standard Gaussian.

2011; Shao and Zhou, 2014) are not distribution-free. Check, for example, the moment requirements for $\boldsymbol{X}$ in Li et al. (2010) and Shao and Zhou (2014).

Of note, Spearman's rho is in the family of simple linear rank statistics, and it is straightforward to check that it satisfies the regularity condition (5.3.1). Therefore, Theorem 5.3.1 is a strict generalization to Theorem 1.2 in Zhou (2007).

We then turn to study the null limiting distribution of the rank type U-statistics. The next theorem gives a result similar to Theorem 5.3.1. In particular, there is also no distributional assumption required.

Theorem 5.3.2 (Rank type U-statistics)**.** *Suppose the rank type U-statistics* $\{U_{jk}, 1 \le j < k \le d\}$ *are of the form* (5.2.3)*, of degree* m*, and satisfy the kernel function* $h(\cdot)$ *is bounded and*

$$\widetilde{\sigma}_U^2 := m^2 \cdot \mathrm{Var}_{\mathbf{H_0}}\left[\mathbb{E}_{\mathbf{H_0}}\left(h(\boldsymbol{X}_{1,\{1,2\}}, \ldots, \boldsymbol{X}_{m,\{1,2\}}) | \boldsymbol{X}_{1,\{1,2\}}\right)\right] > 0 \tag{5.3.3}$$

is fixed[9]*. We then have, under the regime* $\log d = o(n^{1/3})$*,* $d \to \infty$*, and* (5.1.1)*, for any* $y \in \mathbb{R}$*,*

$$\left|\mathbb{P}(n\widetilde{L}_n^2/\sigma_U^2 - 4\log d + \log\log d \le y) - \exp(-\exp(-y/2)/\sqrt{8\pi})\right| = o(1).$$

Here $\widetilde{L}_n$ *and* $\widetilde{\sigma}_T^2$ *are separately defined in* (5.2.6) *and* (5.2.7)*.* σ_U^2 *can be replaced by* $\widetilde{\sigma}_U^2$

[9]This is equivalent to assuming that the rank type U-statistic is non-degenerate, and hence rules out the Hoeffding's D statistic (Hoeffding, 1948b).

without changing the limiting distribution.

As a consequence of Theorems 5.3.1 and 5.3.2, we immediately have the following corollary.

Corollary 5.3.3. *Suppose the conditions in Theorems 5.3.1 or 5.3.2 hold. We then have*

$$\mathbb{P}(\mathsf{T}_\alpha = 1|\mathbf{H_0}) = \alpha + o(1) \text{ or } \mathbb{P}(\widetilde{\mathsf{T}}_\alpha = 1|\mathbf{H_0}) = \alpha + o(1).$$

Corollary 5.3.3 justifies the tests T_α and $\widetilde{\mathsf{T}}_\alpha$ can effectively control the type I error. As an immediate consequence of Theorems 5.3.1 and 5.3.2, all the test statistics in Examples 5.2.1 to 5.2.4 converge weakly to an extreme value type I distribution.

Corollary 5.3.4. *Under the regime* $\log d = o(n^{1/3})$ *and* $d \to \infty$*, we have*

$$\mathbb{P}(\mathsf{T}^a_\alpha = 1|\mathbf{H_0}) = \alpha + o(1)$$

for $a \in \{\rho, \tau, \widehat{\rho}, \widehat{\tau}\}$*, corresponding to test statistics introduced in Examples 5.2.1 to 5.2.4.*

5.4 Power Analysis and Optimality Properties

In this section, we first provide power analysis of the proposed tests against a certain alternative. We then justify the optimality of these tests.

5.4.1 Power Analysis

We first introduce some additional notation. Let's consider the following set of matrices:

$$\mathcal{U}(c) := \Big\{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathrm{diag}(\mathbf{M}) = \mathbf{I}_d, \mathbf{M} = \mathbf{M}^T, \max_{1\le j<k\le d} |\mathbf{M}_{jk}| \ge c\sqrt{\log d/n}\Big\}. \quad (5.4.1)$$

Here $c > 0$ is a generic positive constant. For the simple linear rank statistics $\{T_{jk}, 1 \le j < k \le d\}$, we define the random matrix corresponding to it as $\widehat{\mathbf{T}} = [\widehat{\mathbf{T}}_{jk}] \in \mathbb{R}^{d\times d}$ with

$$\widehat{\mathbf{T}}_{jk} = \widehat{\mathbf{T}}_{kj} = \frac{T_{jk} - \mathbb{E}_{\mathbf{H_0}} T_{jk}}{\sigma_T} \text{ and } \widehat{\mathbf{T}}_{jj} = 1, \text{ for } 1 \le j < k \le d,$$

where σ_T is defined in (5.2.7). We define the population version of $\widehat{\mathbf{T}}$ to be $\mathbf{T} := \mathbb{E}\widehat{\mathbf{T}}$. The next theorem characterizes the conditions under which the power of the test T_α tends to zero as $n \to \infty$. This is when $\mathbf{T}$ is within a certain set of matrices $\mathcal{U}(C)$ for some large enough constant C.

Theorem 5.4.1 (Powers of tests based on simple linear rank statistics)**.** *Assume the alignment assumption in* (5.2.2) *holds. Moreover, assume the following three conditions hold for some positive absolute constants A_1, A_2, and Δ:*

$$\sigma_T^2 = A_1(1+o(1)), \ \max\{|f(0)|, |g(0)|\} \le A_2, \text{ and } f(\cdot), g(\cdot) \text{ having Lipchitz constant } \Delta.$$

We have, there exists some large enough constant B_1 only depending on A_1, A_2, Δ, such

that

$$\inf_{\mathbf{T}\in\mathcal{U}(B_1)} \mathbb{P}(\mathsf{T}_\alpha = 1) = 1 - o(1).$$

Here the infimum is taken over all distributions $\mathbf{T} \in \mathcal{U}(B_1)$.

For the rank type U-statistics $\{U_{jk}, 1 \le j < k \le d\}$, we can similarly define the random matrix corresponding to it as $\widehat{\mathbf{U}} = [\widehat{\mathbf{U}}_{jk}] \in \mathbb{R}^{d\times d}$, with

$$\widehat{\mathbf{U}}_{jk} = \widehat{\mathbf{U}}_{kj} = \frac{U_{jk} - \mathbb{E}_{\mathbf{H_0}} U_{jk}}{\widetilde{\sigma}_U} \text{ and } \widehat{\mathbf{U}}_{jj} = 1, \text{ for } 1 \le j < k \le d,$$

where $\widetilde{\sigma}_U$ is defined in (5.3.3). Let the population version of $\widehat{\mathbf{U}}$ be $\mathbf{U} := \mathbb{E}\widehat{\mathbf{U}}$. Similar to the simple linear rank statistics, the test statistic $\widetilde{\mathsf{T}}_\alpha$ attains the power tending to one when $\mathbf{U}$ belongs to $\mathcal{U}(C)$ for some large enough generic constant C.

Theorem 5.4.2 (Powers of tests based on rank type U-statistics)**.** *Suppose the kernel function* $h(\cdot)$ *in* (5.2.3) *is bounded with* $|h(\cdot)| \le A_3$ *and*

$$m^2 \cdot \mathrm{Var}_{\mathbf{H_0}}\big[\mathbb{E}_{\mathbf{H_0}}(h(\boldsymbol{X}_{1,\{1,2\}}, \ldots, \boldsymbol{X}_{m,\{1,2\}})|\boldsymbol{X}_{1,\{1,2\}})\big] = A_4(1 + o(1))$$

holds for some absolute positive constants A_3, A_4*. There exists some large enough constant* $B_2 > 0$ *only depending on* A_3, A_4*, and* m*, such that*

$$\inf_{\mathbf{U}\in\mathcal{U}(B_2)} \mathbb{P}(\widetilde{\mathsf{T}}_\alpha = 1) = 1 - o(1).$$

Here the infimum is taken over all distributions such that $\mathbf{U} \in \mathcal{U}(B_2)$.

In Theorems 5.4.1 and 5.4.2, we consider a sparse alternative, i.e., at least one entry of $\mathbf{T}$ or $\mathbf{U}$ has the magnitude larger than $C\sqrt{\log d/n}$ for some large enough constant C. Such an alternative can be very close to the null. Actually, we can have all but a small number of entries in $\mathbf{T}$ or $\mathbf{U}$ to be exactly zero. The above theorems show the proposed tests are very sensitive to such small perturbations.

Next we consider the examples discussed in Section 5.2. Let $\mathbf{T}^{\rho}, \mathbf{T}^{\widehat{\tau}}$ and $\mathbf{U}^{\tau}, \mathbf{U}^{\widehat{\rho}}$ be matrices corresponding to Spearman's rho, Kendall's tau, and their variants outlined in Examples 5.2.1 to 5.2.4. They all have power tending to one against the sparse alternative.

Corollary 5.4.3. *There exists constant* $B_a > 0$ *such that*

$$\inf_{\mathbf{M}^a \in \mathcal{U}(B_a)} \mathbb{P}(\mathsf{T}_{\alpha}^{a} = 1) = 1 - o(1),$$

for $a \in \{\rho, \tau, \widehat{\rho}, \widehat{\tau}\}$ *and the matrix* $\mathbf{M}^a \in \{\mathbf{T}^{\rho}, \mathbf{T}^{\widehat{\tau}}, \mathbf{U}^{\tau}, \mathbf{U}^{\widehat{\rho}}\}$.

5.4.2 Optimality Properties

In this section, we prove the optimality of the proposed test statistics. Recall T_{α} and $\widetilde{\mathsf{T}}_{\alpha}$ can correctly reject the sparse alternative as long as at least one entry of $\mathbf{T}$ or $\mathbf{U}$'s magnitudes is larger than $C\sqrt{\log d/n}$ for some large enough constant C. In the following we show such a bound is rate optimal.

We first introduce some additional notation. For testing the null hypothesis $\mathbf{H_0}$, for each n, let's define $\mathcal{T}_\alpha$ to be the set of all size α tests T_α with

$$\mathcal{T}_\alpha := \{T_\alpha : \mathbb{P}(T_\alpha = 1|\mathbf{H_0}) \leq \alpha\}.$$

Consider the Pearson's population correlation matrix $\mathbf{R}$. Note under the Gaussian model, the null hypothesis (5.1.1) is equivalent to the null hypothesis that $\mathbf{R}$ is an identity matrix. The next theorem shows that, over any distribution family including the Gaussian as a subset, any size α test cannot differentiate the null hypothesis $\mathbf{H_0}$ and the sparse alternative when $\max_{j<k} |\mathbf{R}_{jk}| \leq c\sqrt{\log d/n}$ for some constant $c < 1$.

Theorem 5.4.4. *Assume $c_0 < 1$ is an absolute positive constant and we have $\log d/n = o(1)$. Let $\alpha, \beta > 0$ and $\alpha + \beta < 1$ be any two absolute constants. For all large enough n and d, we have*

$$\inf_{T_\alpha \in \mathcal{T}_\alpha} \sup_{\mathbf{R} \in \mathcal{U}(c_0)} \mathbb{P}(T_\alpha = 0) \geq 1 - \alpha - \beta,$$

where the supremum is taken over any distribution family including the Gaussian as a subset and such that $\mathbf{R} \in \mathcal{U}(c_0)$.

For proving Theorem 5.4.4, we adopt the general framework in Baraud (2002). The proof technique is relevant in deriving the lower bound in two sample covariance tests (Cai et al., 2013). However, different from the test for covariance matrix, the test of independence focuses on off-diagonal entries of the correlation/covariance matrices. For incorporating the particular structure of the correlation matrix, we need to construct a different set

of parameters.

Theorem 5.4.4 then directly leads to the next theorem. It justifies the optimality of the proposed tests against the sparse alternative under any distribution family including the Gaussian as a subset.

Theorem 5.4.5. *Suppose the simple linear rank statistics* $\{T_{jk}, 1 \leq j < k \leq d\}$ *satisfy all regularity conditions in Theorems 5.3.1, 5.4.1, and the corresponding matrix* $\mathbf{T}$ *satisfies for all* n, d *large enough, we have*

$$c\mathbf{T}_{jk} \leq \mathbf{R}_{jk} \leq C\mathbf{T}_{jk}.$$

under the Gaussian assumption. We then have, under the regime $\log d = o(n^{1/3})$ *and* $d \to \infty$*, the corresponding size* α *test* T_α *is rate optimal. In other words, there exist two absolute constants* $D_1 < D_2$ *such that: (i)*

$$\sup_{\mathbf{T} \in \mathcal{U}(D_2)} \mathbb{P}(\mathsf{T}_\alpha = 0) = o(1),$$

over any distribution family such that $\mathbf{T} \in \mathcal{U}(D_2)$*; (ii) For any* $\alpha, \beta > 0$ *with* $\alpha + \beta < 1$*, for all large enough* n, d*, we have*

$$\inf_{T_\alpha \in \mathcal{T}_\alpha} \sup_{\mathbf{T} \in \mathcal{U}(D_1)} \mathbb{P}(T_\alpha = 0) \geq 1 - \alpha - \beta,$$

where the supremum is taken over any distribution family including the Gaussian as a

subset and $\mathbf{T} \in \mathcal{U}(D_1)$. *For all rank type U-statistics satisfying the regularity conditions in Theorems 5.3.2 and 5.4.2, and* $\mathbf{U}$ *satisfying the constraint that for all* n, d *large enough, we have*

$$c\mathbf{U}_{jk} \leq \mathbf{R}_{jk} \leq C\mathbf{U}_{jk},$$

under the Gaussian assumption, the same optimality property holds.

As an example, combined with Corollaries 5.3.4 and 5.4.3, the next corollary justifies the tests statistics in Examples 5.2.1 to 5.2.4 are all rate optimal against the sparse alternative.

Corollary 5.4.6. *Over any distribution family including the Gaussian as a subset, the tests statistics in Examples 5.2.1 to 5.2.4 are all rate optimal against the sparse alternative.*

5.5 Numerical Results

In this section, we present the numeric results for investigating the finite sample behaviors of the proposed tests and their competitors under the null and alternative hypotheses[10]. We compare the empirical performance of these proposed tests to the classic likelihood ratio test (Morrison, 2004) and to the alternatives in Jiang (2004) (Here we use the modified version in Zhou (2007).) and in Mao (2014).. We are interested in testing (5.1.1). We denote $\widehat{\mathbf{R}} = [r_{jk}]$ to be the Pearson's sample correlation matrix. The classic likelihood ratio

[10]For fair comparison, we use the theoretical threshold q_α in (5.2.9) instead of the simulation-based one introduced in Section 5.6.2.

test rejects (5.1.1) if we have

$$-\Big(n-1-\frac{2d+5}{6}\Big)\log|\widehat{\mathbf{R}}| > F^{-1}\big(1-\alpha;\chi^2_{d(d-1)/2}\big), \tag{5.5.1}$$

where $F^{-1}(1-\alpha;\xi^2_{d(d-1)/2})$ represents the $1-\alpha$ quantile of a chi-square distribution with degree of freedom $d(d-1)/2$. Jiang (2004) (modified by Zhou (2007)) proposed to reject $\mathbf{H_0}$ if we have

$$n\max_{j<k} r^2_{jk} - 4\log d + \log\log d \geq q_\alpha, \tag{5.5.2}$$

where we note q_α is the theoretical threshold value as defined in (5.2.9). Mao (2014) modified the original test in Schott (2005) and proposed the test rejecting the null hypothesis if

$$\Big(\sum_{j<k}\frac{r^2_{jk}}{1-r^2_{jk}} - \frac{d(d-1)}{2(n-4)}\Big)\cdot\sqrt{\frac{(n-4)^2(n-6)}{d(d-1)(n-3)}} \geq \Phi^{-1}(1-\alpha), \tag{5.5.3}$$

where $\Phi^{-1}(\cdot)$ represents the quantile function of the standard Gaussian. Of note, the test statistic in Mao (2014) has theoretically guaranteed size control only under the Gaussian model. We compare the empirical performance of the following five tests of independence:

- R1: The proposed test using the Spearman's rho statistic, outlined in Example 5.2.1;
- R2: The proposed test using the Kendall's tau statistic, outlined in Example 5.2.2;

- LRT: The likelihood ratio test in (5.5.1);
- Jiang: The test in Jiang (2004), shown in (5.5.2);
- Mao: The test in Mao (2014), shown in (5.5.3).

The next two sections provide experimental comparisons of these methods on both synthetic and real data. Throughout this section, we set the significance level to be $\alpha = 0.05$.

5.5.1 Synthetic Data Analysis

In this section, we illustrate the size and power comparisons among five competing tests introduced earlier. We first focus on evaluating the empirical sizes of the competing tests. To this end, we consider the following four models. Under each model, the data points are independent observations of a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$.

- Model 1 (Gaussian): The data are generated from a Gaussian distribution with $\boldsymbol{X} \sim N_d(\mathbf{0}, \mathbf{I}_d)$.
- Model 2 (Gaussian copula): The data are generated from a Gaussian copula distribution with $X_j = Z_j^{1/3}$ for $j = 1, \ldots, d$. Here $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T \sim N_d(\mathbf{0}, \mathbf{I}_d)$ is standard Gaussian.
- Model 3 (Gaussian copula): The data are generated from a Gaussian copula distribution with $X_j = Z_j^3$ for $j = 1, \ldots, d$. Here $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T \sim N_d(\mathbf{0}, \mathbf{I}_d)$ is standard Gaussian.

- Model 4 (t distribution): The data are generated from a t distribution with $X_1, \ldots, X_d$ independently and identically distributed to a univariate t distribution with degree of freedom three.

We conduct simulation studies based on the above four models, with sample size $n = 60$ or 100, and the dimension d changing from 50 to 800. We derive the empirical results based on 5,000 replications. The empirical sizes of the five competing tests are present in the first half columns of Table 5.1. We will discuss the obtained results in the end of this section.

We then turn to evaluate the sizes of the five competing tests. To this end, we consider the following four models. Under each model, the data points are independent observations of $\boldsymbol{X}$.

- Model 5 (Gaussian): The data are generated from a Gaussian distribution with $\boldsymbol{X} \sim N_d(\mathbf{0}, \mathbf{R}^*)$. Here $\mathbf{R}^*$ is generated as follows: Consider an random matrix $\boldsymbol{\Delta}$ with all but eight random nonzero entries. We select the locations of four nonzero entries randomly from the upper triangle of $\boldsymbol{\Delta}$, each with a magnitude randomly drawn from the uniform distribution in $[0, 1]$. The other four nonzero entries in the lower triangle are determined by symmetry. $\mathbf{R}^* := \mathbf{I}_d + \boldsymbol{\Delta} + \delta \mathbf{I}_d$, where $\delta = (-\lambda_{\min}(\mathbf{I}_d + \boldsymbol{\Delta}) + 0.05) \cdot I(\lambda_{\min}(\mathbf{I}_d + \boldsymbol{\Delta}) \leq 0)$.

- Model 6 (Gaussian copula): The data are generated from a Gaussian copula distribution with $X_j = Z_j^{1/3}$ for $j = 1, \ldots, d$. Here $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T \sim N_d(\mathbf{0}, \mathbf{R}^*)$ is multivariate Gaussian.

- Model 7 (Gaussian copula): The data are generated from a Gaussian copula distribution with $X_j = Z_j^3$ for $j = 1, \ldots, d$. Here $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T \sim N_d(\mathbf{0}, \mathbf{R}^*)$ is multivariate Gaussian.

- Model 8 (multivariate t distribution): The data are generated from a multivariate t distribution with covariance matrix $\mathbf{R}^*$ and degree of freedom 3.

The second half columns of Table 5.1 show the empirical powers of the five competing tests based on 5,000 replications. Here the sample size ranges from 60 to 100, and the dimension changes from 50 to 800.

Checking Table 5.1, there are some notable observations. First, with regard to the Gaussian setting, the results in Model 1 show all tests can effectively control the sizes under all different n and d considered here. On the other hand, the results in Model 5 show, Jiang attains the highest power against the sparse alternative, and it is closely followed by R1 and R2. Compared to CJ, R1, and R2, Mao has relatively lower power. This is as expected because in Model 5 the Pearson's correlation matrix has only 8 nonzero entries. By averaging over all entries in the correlation matrix, Mao is less sensitive to such a sparse alternative.

With regard to the nonGaussian case, we consider three settings: the light tailed Gaussian copula data (Models 2 and 6), the heavy tailed Gaussian copula data (Models 3 and 7), and the heavy tailed elliptical data (Models 4 and 8). With regard to the light tailed data, the results for Model 2 show that R1, R2, and Jiang can effectively control the sizes, while for Mao it is slightly inflated. On the other hand, the results in model 7 illustrate R2

Table 5.1: Comparison of five competing tests on Models 1 to 8. The sample size n is changing from 60 to 100, and the dimension d ranges from 50 to 800. The results are derived under 5,000 replications.

n	d	R1	R2	LRT	Jiang	Mao	R1	R2	LRT	Jiang	Mao
		Model 1 (empirical size)					Model 5 (empirical power)				
60	50	0.027	0.027	0.941	0.026	0.053	0.900	0.915	0.997	0.924	0.748
	100	0.021	0.025	-	0.021	0.047	0.879	0.896	-	0.902	0.598
	200	0.021	0.025	-	0.019	0.056	0.842	0.863	-	0.865	0.421
	400	0.014	0.015	-	0.013	0.043	0.800	0.829	-	0.837	0.274
	800	0.012	0.011	-	0.010	0.058	0.785	0.829	-	0.836	0.192
100	50	0.032	0.032	0.221	0.027	0.056	0.969	0.969	0.900	0.971	0.865
	100	0.030	0.034	-	0.023	0.058	0.952	0.960	-	0.965	0.766
	200	0.025	0.025	-	0.024	0.046	0.941	0.944	-	0.950	0.578
	400	0.018	0.023	-	0.018	0.050	0.935	0.936	-	0.956	0.396
	800	0.021	0.030	-	0.019	0.048	0.909	0.923	-	0.930	0.260
		Model 2 (empirical size)					Model 6 (empirical power)				
60	50	0.027	0.027	0.953	0.032	0.057	0.900	0.915	0.995	0.880	0.647
	100	0.021	0.025	-	0.032	0.058	0.879	0.896	-	0.843	0.511
	200	0.021	0.025	-	0.026	0.066	0.842	0.863	-	0.808	0.350
	400	0.014	0.015	-	0.021	0.069	0.800	0.829	-	0.739	0.247
	800	0.012	0.011	-	0.018	0.111	0.785	0.829	-	0.722	0.226
100	50	0.032	0.032	0.236	0.041	0.053	0.969	0.969	0.848	0.950	0.811
	100	0.030	0.034	-	0.035	0.058	0.952	0.960	-	0.944	0.677
	200	0.025	0.025	-	0.042	0.059	0.941	0.944	-	0.927	0.463
	400	0.018	0.023	-	0.032	0.049	0.935	0.936	-	0.917	0.320
	800	0.021	0.030	-	0.030	0.066	0.909	0.923	-	0.879	0.218
		Model 3 (empirical size)					Model 7 (empirical power)				
60	50	0.027	0.027	0.955	0.981	0.509	0.900	0.915	0.986	0.997	0.822
	100	0.021	0.025	-	1.000	0.827	0.879	0.896	-	1.000	0.927
	200	0.021	0.025	-	1.000	0.999	0.842	0.863	-	1.000	1.000
	400	0.014	0.015	-	1.000	1.000	0.800	0.829	-	1.000	1.000
	800	0.012	0.011	-	1.000	1.000	0.785	0.829	-	1.000	1.000
100	50	0.032	0.032	0.392	0.989	0.326	0.969	0.969	0.818	0.998	0.828
	100	0.030	0.034	-	1.000	0.542	0.952	0.960	-	1.000	0.843
	200	0.025	0.025	-	1.000	0.860	0.941	0.944	-	1.000	0.954
	400	0.018	0.023	-	1.000	1.000	0.935	0.936	-	1.000	1.000
	800	0.021	0.030	-	1.000	1.000	0.909	0.923	-	1.000	1.000
		Model 4 (empirical size)					Model 8 (empirical power)				
60	50	0.032	0.033	0.944	0.399	0.126	0.901	0.945	1.000	0.994	1.000
	100	0.023	0.034	-	0.639	0.181	0.883	0.919	-	0.995	1.000
	200	0.020	0.025	-	0.915	0.302	0.878	0.910	-	0.995	1.000
	400	0.013	0.017	-	1.000	0.648	0.863	0.905	-	0.998	1.000
	800	0.010	0.009	-	1.000	0.985	0.834	0.882	-	0.995	1.000
100	50	0.035	0.035	0.256	0.475	0.113	0.965	0.978	1.000	1.000	1.000
	100	0.034	0.034	-	0.797	0.130	0.957	0.971	-	1.000	1.000
	200	0.036	0.038	-	0.991	0.234	0.941	0.964	-	1.000	1.000
	400	0.023	0.028	-	1.000	0.427	0.911	0.948	-	1.000	1.000
	800	0.019	0.023	-	1.000	0.852	0.907	0.947	-	1.000	1.000

has the highest power, followed by R1. This is as expected because the current data are Gaussian copula distributed, while the rank-based methods are very effective in handling such data and their performance is identical to that under the Gaussian model. With regard to the heavy tailed data, the results in Models 3 and 4 show the sizes of Jiang and Mao are severely inflated, while R1 and R2 can still effectively control the sizes. Because of these inflations, the comparison of the powers between R1, R2 and others are unfair.

In summary, R1 and R2 perform the best across all settings, with sizes consistently under control and attaining averagely highest powers. R2 performs slightly better than R1 in terms of having higher powers. This is consistent to the theoretical results in Woodworth (1970). They showed that Kendall's tau is more powerful than Spearman's rho in testing independence in terms of having Bahadur efficiency bounded in $(1, 1.05]$ under the Gaussian model. The performance of Jiang and Mao is severely influenced by the data structure, and cannot effectively control the sizes for heavy tailed data. This is as expected because the theory of Mao relies heavily on the Gaussian assumption, while the performance of Jiang is related to the moments of the data. LRT does not perform well across all settings because it is not designed for high dimensional data.

5.5.2 Real Data Analysis

We study the empirical performance of the competing methods on an real stock market monthly log return data. We collect the daily closing prices of 452 stocks consistently in the Standard and Poor 500 index from January 1, 2003 to January 1, 2008 (`finance.`

yahoo.com). This gives us all together 59 data points, each with dimension 452. The corresponding data matrix has the numbers of rows and columns 59 and 452 respectively.

For evaluating the size control ability of these methods, we generate the datasets with independent columns based on the above monthly return data matrix. This is via random permutation as follows: Each time, we independently and randomly permute the entries in each column of the data matrix. Then it is obvious that the corresponding columns are completely independent given the observed original stock data matrix.

We cannot apply LRT to this randomly permuted dataset because the dimension is higher than the sample size. We apply the rest four competing tests to the randomly permuted dataset and record the resulting p-values. We repeat this procedure for 1,000 times. Figure 5.1 shows the histograms of p-values for these four tests.

Because the data matrix is permuted within each column, the corresponding 452 entries are completely independent and the histograms should be close to that of the uniform distribution in $[0, 1]$. We find the histograms of R1 and R2 are relatively flat and the proposed tests can effectively control the size. In comparison, there is a strong skewness to the left for the histograms of Jiang and Mao, indicating the tests tend to reject the null hypothesis. This is as expected since the log return data contain extreme events and are heavy tailed (Rachev, 2003). Random permutation cannot eliminate such extreme events. Following the discussions in Shao and Zhou (2014) and observations in Section 5.5.1, Jiang and Mao are very sensitive to such extreme events.

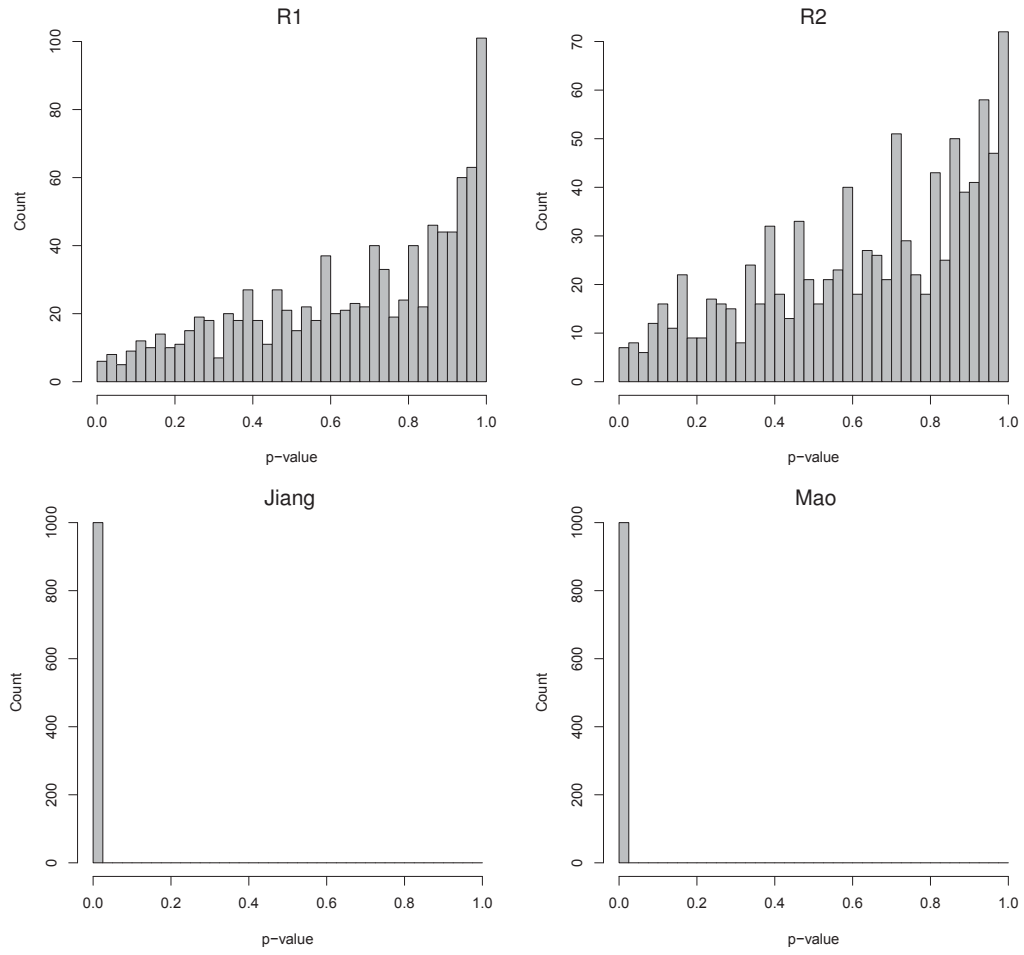


Figure 5.1: Histograms of the p-values of four competing methods on the randomly permuted monthly return data. The results are derived based on 1,000 replications. The empirical probabilities of the pvalues less than 0.05 are 0.014, 0.015, 1.000, and 1.000 for R1, R2, Jiang, and Mao respectively.

5.6 Additional Results

In this section, we provide some additional results in generalizing the aforementioned statistics to testing other structural hypotheses. We further discuss approximating the exact distributions of the proposed test statistics, which focuses on accelerating the rate of convergence.

5.6.1 Generalizations to Other Structural Testing Problems

In this section we discuss and generalize the results in testing independence to other settings, including testing m-dependence and data homogeneity.

5.6.1.1 Test of m-dependence

In this section we study the semiparametric Gaussian copula distribution family. A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ follows a Gaussian copula distribution if and only if we have

$$(F_1(X_1), F_2(X_2), \ldots, F_d(X_d))^T \stackrel{D}{=} (\Phi(Z_1), \ldots, \Phi(Z_d))^T,$$

where $F_1, \ldots, F_d$ are the marginal distribution functions of $X_1, \ldots, X_d$, $\Phi(\cdot)$ represents the distribution function of the standard Gaussian, and we have

$$\boldsymbol{Z} = (Z_1, \ldots, Z_d)^T \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}^0)$$

with $\boldsymbol{\Sigma}^0$ having diagonal entries all equal to 1. The Gaussian copula family includes the Gaussian family, and is a semiparamtric model since we do not specify the marginal distributions of $\boldsymbol{X}$. We refer $\boldsymbol{\Sigma}^0$ to be the latent correlation matrix of $\boldsymbol{X}$. For simplicity, we only consider $\boldsymbol{X}$ to be continuous for avoiding possible ties.

We want to test the null hypothesis $\mathbf{A_0}$:

$$\mathbf{A_0} : \mathbf{\Sigma}^0_{jk} = 0, \quad \text{for all } |j-k| \geq m.$$

Because $\boldsymbol{X}$ is assumed to be Gaussian copula distributed, the dependence structure among $\{X_1, \ldots, X_d\}$ is fully encoded in $\mathbf{\Sigma}^0$. Therefore, the aforementioned null hypothesis is equivalent to testing m-dependence of entries in $\boldsymbol{X}$:

$$X_j \text{ is independent of } X_k, \text{ for all } |j-k| \geq m.$$

Cai and Jiang (2011) were the first to consider testing $\mathbf{A_0}$ in high dimensions. The theory there applies only to the Gaussian data. Later, the result was extended to the non-Gaussian data with a moment assumption (Shao and Zhou, 2014). In this section, we show how, by resorting to the rank-based statistics, the moment assumption can be relaxed.

Under $\mathbf{A_0}$, there are many entries among $\{X_1, \ldots, X_d\}$ that are correlated as long as $m > 1$. For testing $\mathbf{A_0}$, instead of resorting to the Pearson's sample correlation coefficients as in Cai and Jiang (2011) and Shao and Zhou (2014), we consider using the Kendall's tau correlation coefficients $\{\tau_{jk}, 1 \leq j < k \leq d\}$ introduced in Example 5.2.2. It is well known that Kendall's tau is irrelevant to the marginal distributions of $\boldsymbol{X}$ (Nelsen, 1999). Accordingly, within the Gaussian copula family, Kendall's tau is a more natural measurement of dependence than the Pearson's sample correlation coefficient. Moreover,

it is known that, under the Gaussian copula family, we have

$$\mathbf{\Sigma}^0_{jk} = \sin\Big(\frac{\pi}{2}\tau^0_{jk}\Big), \quad \text{where } \tau^0_{jk} := \mathbb{E}\tau_{jk}.$$

Therefore, within the Gaussian copula family, testing $\mathbf{A_0}$ is equivalent to testing $\tau^0_{jk} = 0 \text{ for all } |j-k| \geq m$. We then propose the following test statistic

$$\mathsf{T}^\tau_{\alpha,m} := I\Big(\frac{9n}{4} \cdot (L^\tau_{n,m})^2 - 4\log d + \log\log d \geq q_\alpha\Big), \tag{5.6.1}$$

where q_α is in (5.2.9) and the extreme statistic $L^\tau_{n,m} := \max_{|j-k|\geq m} |\tau_{jk}|$. $L^\tau_{n,m}$ is an extreme value statistic similar to L^τ_n and we expect it to have similar null limiting distribution as L^τ_n given some proper conditions on m. We reject $\mathbf{A_0}$ if and only if $\mathsf{T}^\tau_{\alpha,m} = 1$.

The following theorem justifies the test $\mathsf{T}^\tau_{\alpha,m}$ can asymptotically control the type I error as n, d increase to infinity, provided that certain conditions hold.

Theorem 5.6.1. *Suppose that we have* $\log d = o(n^{1/3})$, $d \to \infty$, $m = o(d^c)$ *for any* $c > 0$, *and for some fixed constant* $\delta \in (0,1)$, *we have*

$$\mathrm{Card}\big(\big\{1 \leq j \leq d : |\mathbf{\Sigma}^0_{jk}| > 1-\delta \text{ for some } 1 \leq k \leq d \text{ and } j \neq k\big\}\big) = o(d).$$

Provided $\boldsymbol{X}$ *is Gaussian copula distributed and continuous, under* $\mathbf{A_0}$, *we have*

$$\left|\mathbb{P}\Big(\frac{9n}{4} \cdot (L^\tau_{n,m})^2 - 4\log d + \log\log d \leq y\Big) - \exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{y}{2}\right)\right)\right| = o(1).$$

Accordingly, the test $\mathsf{T}^{\tau}_{\alpha,m}$ can asymptotically control the type I error as $n, d \to \infty$, i.e.,

$$\mathbb{P}(\mathsf{T}^{\tau}_{\alpha,m} = 1 | \mathbf{A_0}) = \alpha + o(1).$$

Remark 5.6.2. *From the proof of this theorem, we can see the assumption, $m = o(d^c)$ for any $c > 0$, can be easily relaxed. Specifically, we only need to require $m = o(d^{\epsilon(\delta)})$ for some small enough constant $\epsilon(\delta)$ depending on δ. This can be verified by checking the proof. But for simplicity, we use this assumption for discussion.*

Similar to the power analysis in Section 5.4.1, we study the power of the test statistic $\mathsf{T}^{\tau}_{\alpha,m}$ against a certain sparse alternative set. To this end, let's consider the following set of matrices

$$\mathcal{U}_m(c) := \Big\{ \mathbf{M} \in \mathbb{R}^{d\times d} : \mathrm{diag}(\mathbf{M}) = \mathbf{I}_d, \mathbf{M} = \mathbf{M}^T, \max_{|j-k|\geq m} |\mathbf{M}_{jk}| \geq c\sqrt{\log d/n} \Big\}.$$

The following theorem shows under the Gaussian copula family, as long as the latent correlation matrix $\mathbf{\Sigma}^0$ is within a set $\mathcal{U}_m(C)$ for some large enough constant C, the type II error of the proposed test will tend to zero.

Theorem 5.6.3. *Suppose we observe n independent observations of a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ following a Gaussian copula distribution with the latent*

correlation matrix $\mathbf{\Sigma}^0$. Then, there exists some large enough constant D_3 such that

$$\sup_{\mathbf{\Sigma}^0 \in \mathcal{U}_m(D_3)} \mathbb{P}(\mathsf{T}^{\tau}_{\alpha,m} = 0) = o(1),$$

as n and d go to infinity. Here the supremum is taken over the Gaussian copula family such that $\mathbf{\Sigma}^0 \in \mathcal{U}_m(D_3)$

We prove Theorem 5.6.3 using the similar technique as in the proof of Theorem 5.4.1. The proof is omitted accordingly.

We then turn to study the optimality of $\mathsf{T}^{\tau}_{\alpha,m}$. For testing $\mathbf{A_0}$, for each n, we define $\mathcal{T}_{\alpha,m}$ to be the set of all size α tests $T_{\alpha,m}$ with

$$\mathbb{P}(T_{\alpha,m} = 1 | \mathbf{A_0}) \leq \alpha.$$

The following theorem gives the detection lower bound in differentiating the null hypothesis and the sparse alternative.

Theorem 5.6.4. *Assume $c_0' < 1$, $\log d/n = o(1)$, $d \to \infty$, and $m = o(d^c)$ for any $c > 0$. Let $\alpha, \beta > 0$ with $\alpha + \beta < 1$. For all large enough n and d, we have*

$$\inf_{T_{\alpha,m} \in \mathcal{T}_{\alpha,m}} \sup_{\mathbf{\Sigma}^0 \in \mathcal{U}_m(c_0')} \mathbb{P}(T_{\alpha,m} = 0) \geq 1 - \alpha - \beta,$$

where the supremum is taken over any distribution family including the Gaussian as a subset, and such that $\mathbf{\Sigma}^0 \in \mathcal{U}_m(c_0')$.

Therefore, as in the discussion in Section 5.4, we conclude $\mathsf{T}^{\tau}_{\alpha,m}$ is rate optimal when testing the null hypothesis $\mathbf{A_0}$ against the sparse alternative.

Of note, for any constant $c > 0$, the matrix set $\mathcal{U}(c)$ defined in (5.4.1) includes $\mathcal{U}_m(c)$. Accordingly, the lower bound derived in Section 5.4.2 cannot be trivially exploited in deriving the lower bound for testing the bandedness of $\mathbf{\Sigma}^0$. However, using the fact $m = o(d^c)$ for any $c > 0$, we can prove the lower bound for testing $\mathbf{A_0}$ via designing a similar set of parameters as in the proof of Theorem 5.4.4.

5.6.1.2 Test of Homogeneity

In this section, we consider testing the complete homogeneity of the data. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be n independently but not necessarily identically distributed random vectors with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$ for $i = 1, \ldots, n$. We aim at testing

$$\mathbf{B_0}: \ \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \text{ are identically distributed.} \tag{5.6.2}$$

Testing (5.6.2) is of fundamental interest in many statistical fields, including linear discriminant analysis, principal component analysis, and graphical model estimation.

General tests of homogeneity in high dimensions is very complicated. The works in this field are very few and most of them are to test the equity of two sample means and covariance matrices. For example, Bai and Saranadasa (1996), Srivastava and Du (2008), Chen and Qin (2010), and Cai et al. (2014a) considered comparing the means of two high

dimensional Gaussian vectors with unknown covariance matrices. Chen et al. (2010) and Cai et al. (2014a) developed tests of equity of two covariance matrices.

In this section we consider a simplified setting of (5.6.2). That is, the entries in each $\boldsymbol{X}_i$ are mutually independent. Under this simplified setting, we reduce the test of $\mathbf{B_0}$ to the test that $X_{1j}, X_{2j}, \ldots, X_{nj}$ are identically distributed for any $j \in \{1, \ldots, d\}$. For each j, we test the homogeneity using an rank-based test statistic. In the end, we formulate an extreme value statistic via combining the d separate rank-based test statistics.

In detail, let H_n be an extreme value statistic summarizing the d separate rank-based test statistics:

$$H_n := \max_{j\in\{1,\ldots,d\}} |h_j|, \quad \text{with } h_j := \frac{2}{n(n-1)} \sum_{i<i'} \operatorname{sign}(X_{i'j} - X_{ij}).$$

Here h_j is an rank-based statistic counting the number of inequalities $X_{i'j} > X_{ij}$ across all pairs $i < i'$[11]. For testing (5.6.2), we propose the following statistic based on H_n:

$$\mathsf{T}^h_\alpha := I\Big(\frac{9n}{4} H_n^2 - 2\log d + \log\log d \geq \widetilde{q}_\alpha\Big),$$

where $\widetilde{q}_\alpha := -\log\pi - 2\log\log(1-\alpha)^{-1}$ is the $1-\alpha$ quantile of the extreme value distribution with the distribution function $\exp(-\exp(-y/2)/\sqrt{\pi})$.

Next we justify the test T^h_α has theoretically controlled size. Under $\mathbf{B_0}$, we have

[11] Mann (1945) is the first to introduce the test statistic h_j for testing homogeneity. They characterize the sufficient conditions for h_j to be consistent and unbiased. They show that this statistic is powerful against a trend alternative (We will introduce the trend alternative in more details later). For more discussions on the rationale of using h_j for testing homogeneity, we refer to Kendall and Stuart (1977).

$X_{1j}, \ldots, X_{nj}$ are identically distributed and hence each $\mathrm{sign}(X_{i'j} - X_{ij})$ should be close to zero, and the ranks of $X_{1j}, \ldots, X_{nj}$ are uniformly sampled from the set of all permutations of $\{1, \ldots, n\}$. Accordingly, h_j is identically distributed to the Kendall's tau statistic under (5.1.1). Therefore, using Example 5.2.2, we derive

$$\mathbb{E}(h_j|\mathbf{B_0}) = 0 \text{ and } \mathrm{Var}(h_j|\mathbf{B_0}) = \frac{2(2n+5)}{9n(n-1)} = \frac{4}{9n} \cdot (1 + o(1)),$$

and the limiting distribution of H_n shall resemble the Kendall's tau counterpart. Specifically, the following theorem provides the null limiting distribution of H_n.

Theorem 5.6.5. *Suppose* $\log d = o(n^{1/3})$ *and* $d \to \infty$. *Under* $\mathbf{B_0}$*, we have, for any* $y \in \mathbb{R}$*,*

$$\left|\mathbb{P}\Big(\frac{9n}{4}H_n^2 - 2\log d + \log\log d\Big) - \exp\Big(-\frac{1}{\sqrt{\pi}}\exp\Big(-\frac{y}{2}\Big)\Big)\right| = o(1).$$

Accordingly, the test T_α^h *can asymptotically control the type I error as* $n, d \to \infty$*, i.e.,*

$$\mathbb{P}(\mathsf{T}_\alpha^h = 1|\mathbf{B_0}) = \alpha + o(1).$$

It is worth noting, similar to Corollary 5.3.3, Theorem 5.6.5 holds without any distributional assumption on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

We then study the power of the proposed test. We consider a particular trend alternative. That is, for at least one entry $j \in \{1, \ldots, d\}$, the mean of X_{ij} is a linear function of i for a

certain entry $j \in \{1, \ldots, d\}$:

$$\mathbf{B_1} : \text{there exists some } j \in \{1, \ldots, d\} \text{ such that } \mathbb{E}X_{ij} = \beta_0 + \beta_1 i/n \text{ with } \mathrm{Var}(X_{ij}) = \sigma^2, \tag{5.6.3}$$

for $i = 1, \ldots, n$ and $\beta_0, \beta_1, \sigma^2 \in \mathbb{R}$. Under $\mathbf{B_1}$, the variance σ^2 is identical across samples while the means are monotonically increasing or decreasing with regard to the label i. Such an alternative is of interest in areas including quality control, finance, and longitudinal data analysis. For example, in quality control we are interested in inspecting whether machines keep performing well. There one alternative of interest is: At least one machine's performance keeps decreasing.

Under $\mathbf{B_1}$, let's consider the following set of real numbers (a_1, a_2):

$$\mathcal{B}(c) := \Big\{(a_1, a_2) : a_2 > 0 \text{ and } |a_1|/a_2 \geq c\sqrt{\log d/n}\Big\}.$$

The following theorem shows, uniformly over the alternative hypothesis set $\mathcal{B}(C)$, for some large enough constant $C > 0$, the power of the proposed test tends to 1 as $n \to \infty$.

Theorem 5.6.6. *Suppose there exists at least one entry $j \in \{1, \ldots, d\}$ satisfying* (5.6.3) *with parameters of interest (β_1, σ). Moreover, for $i = 1, \ldots, n$, the density function $p_{ij}(\cdot)$*

of $(X_{ij} - \mathbb{E}X_{ij})/(\mathrm{Var}X_{ij})^{1/2}$ *is identical to some density function* $p(\cdot)$*, which satisfies that*

$$p(x) \geq D_4 > 0 \text{ for all } x \in [-M, M], \tag{5.6.4}$$

for some fixed constant $M > 0$*. Then there exists some large enough constant* D_5 *only depending on* D_4 *and* M *such that*

$$\sup_{(\beta_1,\sigma)\in\mathcal{B}(D_5)} \mathbb{P}(\mathsf{T}_\alpha^h = 0) = o(1).$$

In the following we show the detection boundary $|\beta_1|/\sigma \geq C\sqrt{\log d/n}$ is rate optimal. To this end, let's introduce some additional notation. We define $\mathcal{T}_\alpha^h$ to be the set of all size α tests T_α^h satisfying

$$\mathbb{P}(T_\alpha^h = 1|\mathbf{B_0}) \leq \alpha.$$

The following theorem shows the proposed test is rate optimal in testing against the trend alternative $\mathbf{B_1}$.

Theorem 5.6.7. *Assume* $c_0'' < \sqrt{3}$*,* $\log d/n = o(1)$*, and* $d \to \infty$*. Let* $\alpha, \beta > 0$ *with* $\alpha + \beta < 1$*. For all large enough* n, d*, we have*

$$\inf_{T_\alpha^h \in \mathcal{T}_\alpha^h} \sup_{(\beta_1,\sigma)\in\mathcal{B}(c_0'')} \mathbb{P}(T_\alpha^h = 0) \geq 1 - \alpha - \beta,$$

where T_α^h *represents any size* α *test under* $\mathbf{B_0}$*, and the supremum is taken over any distri-*

bution family of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ including the Gaussian as a subset.

It is clear, when $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are Gaussian distributed, Equation (5.6.4) in Theorem 5.6.6 is satisfied. Accordingly, combining Theorems 5.6.5, 5.6.6, and 5.6.7 concludes that T_α^h is rate optimal in testing the null hypothesis $\mathbf{B_0}$ against the trend alternative.

5.6.2 Approximation to the Exact Distributions

In this section, we study the convergence rates of the proposed test statistics to the null limiting distribution, and further discuss accelerating the rate of convergence. The focus is on testing independence, although the generalization to testing homogeneity is straightforward.

Theorems 5.3.1 and 5.3.2 show the proposed test statistics L_n and $\widetilde{L}_n$ converge weakly to the type I extreme distribution. The next theorem explicitly characterizes the convergence rates for L_n and $\widetilde{L}_n$.

Theorem 5.6.8. *For all rank type U-statistics, if conditions in Theorem 5.3.2 hold and* $\log d = o(n^{1/3})$*, we have*

$$\Big|\mathbb{P}\Big(\frac{n\widetilde{L}_n^2}{\sigma_U^2} - 4\log d + \log\log d \le y\Big) - \exp\Big(-\frac{1}{\sqrt{8\pi}}\exp\Big(-\frac{y}{2}\Big)\Big)\Big| = O\Big(\frac{(\log d)^{3/2}}{n^{1/2}} + \frac{1}{(\log d)^{3/2}}\Big).$$

For all simple linear rank statistics, if conditions in Theorem 5.3.1 hold and $\log d =$

$O(n^{1/3-\epsilon})$ for some fixed constant $\epsilon \in (0, 1/3)$, we have

$$\Big|\mathbb{P}\Big(\frac{nL_n^2}{\sigma_T^2} - 4\log d + \log\log d \le y\Big) - \exp\Big(-\frac{1}{\sqrt{8\pi}}\exp\Big(-\frac{y}{2}\Big)\Big)\Big| = O\Big(\frac{(\log d)^{3/2}}{n^{1/2}} + \frac{1}{(\log d)^{3/2}} + \frac{(\log d)^{1/2}}{n^{1/6}}\Big).$$

Theorem 5.6.8 shows two points: (i) When $\log d \asymp n^{\kappa}$ for some $\kappa < 1/3$, the proposed tests based on simple linear rank and rank type U-statistics achieve a polynomial rate of convergence[12]. (ii) When $d \asymp n^C$ or some $C \in (0, \infty)$, Theorem 5.6.8 only guarantees an $O((\log n)^{-3/2})$ rate of convergence. In the following we show the convergence rate can be accelerated by approximating the exact distributions of the test statistics.

Of note, under (5.1.1) $\{T_{jk}, j < k\}$ and $\{U_{jk}, j < k\}$ are independent and only depend on the relative ranks $\{R_{ni}^{jk}, i = 1, \ldots, n, j < k\}$. In particular, the relative ranks are uniformly distributed in the set of all permutations of $\{1, \ldots, n\}$. Therefore, we can conduct simulation to approximate the exact distributions of $\{T_{jk}, j < k\}$ and $\{U_{jk}, j < k\}$.

More specifically, each time for $i = 1, \ldots, M$, we generate $\mathbf{X}^{(i)} \in \mathbb{R}^{n\times d}$ to be an n by d data matrix with all entries in $\mathbf{X}^{(i)}$ independently and identically drawn from the standard Gaussian. For each $i \in \{1, \ldots, M\}$, we calculate the values of pairwise simple linear rank statistics $\{T_{jk}^{(i)}, j < k\}$ and rank type U-statistics $\{U_{jk}^{(i)}, j < k\}$. Based on them, we calculate the values of $n(L_n^{(i)})^2/\sigma_T^2 - 4\log d + \log\log d$ and $n(\widetilde{L}_n^{(i)})^2/\sigma_U^2 - 4\log d + \log\log d$. Here $L_n^{(i)}$ and $\widetilde{L}_n^{(i)}$ are the extreme value statistics based on $\{T_{jk}^{(i)}, j < k\}$ and $\{U_{jk}^{(i)}, j < k\}$.

[12]Compared to the tests based on the rank type U-statistics, the tests based on simple linear rank statistics lose an extra $O(\sqrt{\log d}/n^{1/6})$ term in the rate of convergence. This is the cost of approximating the population ranks using the empirical ones.

Let $\widehat{F}^T_{n,d;M}(\cdot)$ and $\widehat{F}^U_{n,d,;M}(\cdot)$ be the corresponding empirical distributions. Let $F^T_{n,d}(\cdot)$ and $F^U_{n,d}(\cdot)$ be their population counterparts.

The Dvoretzky-Kiefer-Wolfowitz inequality for discrete random variables (Dvoretzky et al., 1956; Massart, 1990) guarantees, for each (n, d), we have

$$\begin{aligned}\mathbb{P}\Big(\sup_{x\in\mathbb{R}}|\widehat{F}^T_{n,d;M}(x)-F^T_{n,d}(x)|>\sqrt{\frac{\log M}{M}}\Big)\leq&\frac{2}{M^2},\\ \mathbb{P}\Big(\sup_{x\in\mathbb{R}}|\widehat{F}^U_{n,d;M}(x)-F^U_{n,d}(x)|>\sqrt{\frac{\log M}{M}}\Big)\leq&\frac{2}{M^2}.\end{aligned}\tag{5.6.5}$$

In (5.2.8), we replace q_α using $\widehat{q}^T_{\alpha;n,d}$ and $\widehat{q}^U_{\alpha;n,d}$, which are the $1-\alpha$ quantiles of $\widehat{F}^T_{n,d;M}(\cdot)$ and $\widehat{F}^U_{n,d;M}(\cdot)$ separately:

$$\widehat{q}^T_{\alpha;n,d}:=\inf\{x:\widehat{F}^T_{n,d;M}(x)\geq 1-\alpha\}\ \text{ and }\ \widehat{q}^U_{\alpha;n,d}:=\inf\{x:\widehat{F}^U_{n,d;M}(x)\geq 1-\alpha\}.$$

We refer the tests using the simulation-based threshold values $\widehat{q}^T_{\alpha;n,d}$ and $\widehat{q}^U_{\alpha;n,d}$ to be the exact tests.

Using (5.6.5), we immediately have the next theorem, which guarantees that the exact tests have asymptotically controlled sizes.

Theorem 5.6.9. *Under* (5.1.1)*, the simple linear rank statistics satisfy that, for each* (n, d)*, with probability larger than* $1-2/M^2$*, we have*

$$\sup_{\alpha\in[0,1]}\Big|\mathbb{P}\Big(\frac{nL_n^2}{\sigma_T^2}-4\log d+\log\log d\geq\widehat{q}^T_{\alpha;n,d}|\{\mathbf{X}^{(i)}\}_{i=1}^M\Big)-(1-\widehat{F}^U_{n,d;M}(\widehat{q}^T_{\alpha;n,d}))\Big|\leq\sqrt{\frac{\log M}{M}}.$$

The same inequality also applies to the rank type U-statistics. Moreover, as $n, d \to \infty$, $\widehat{q}^{T}_{\alpha;n,d}$ *and* $\widehat{q}^{U}_{\alpha;n,d}$ *are both consistent estimators of* q_α *in* (5.2.9) *as* $M = M_n \to \infty$ *with* n.

Theorem 5.6.9 shows, with high probability, we can have arbitrarily fast convergence rate to the above intermediate approximation by setting the number of simulations M large enough. It is typically much faster than the rate of convergence $O((\log n)^{5/2}/\sqrt{n})$ derived in Liu et al. (2008). On the other hand, to attain this arbitrarily fast rate of convergence, we need to conduct M simulations for estimating the threshold value. This increases the computational burden compared to all tests in (5.2.8). For the test of m-dependence, it is impossible to simulate the null exact distribution and we stick to the test in (5.6.1).

In the following, we provide the empirical sizes and powers of such exact tests by studying the finite sample Gaussian distributed datasets. We adopt the same generating models 1 and 5 in Section 5.5.1. We compare the performance of LRT, Jiang, and Mao to two exact tests:

- R1e: The proposed test based on Spearman's rho and simulation-based threshold value;
- R2e: The proposed test based on Kendall's tau and simulation-based threshold value.

Table 5.2 provides the results. We see the type I errors of the exact tests R1e and R2e are well controlled, and the powers are higher than the tests R1 and R2, which use the theoretical threshold value q_α. Moreover, the powers of R1e, R2e are averagely comparable to and sometimes higher than Jiang. This indicates the extra gain by resorting to the exact

Table 5.2: Comparison of five competing tests on Models 1 and 5. Here we conduct exact tests for tests based on Spearman's rho and Kendall's tau. The sample size n is changing from 60 to 100, and the dimension d ranges from 50 to 800. The results are derived under 5,000 replications.

n	d	R1e	R2e	LRT	Jiang	Mao	R1e	R2e	LRT	Jiang	Mao
		Model 1 (empirical size)					Model 5 (empirical power)				
60	50	0.056	0.054	0.941	0.026	0.053	0.919	0.923	0.997	0.924	0.748
	100	0.056	0.050	-	0.021	0.047	0.899	0.907	-	0.902	0.598
	200	0.048	0.040	-	0.019	0.056	0.890	0.888	-	0.865	0.421
	400	0.048	0.058	-	0.013	0.043	0.879	0.880	-	0.837	0.274
	800	0.041	0.048	-	0.010	0.058	0.845	0.844	-	0.836	0.192
100	50	0.059	0.058	0.221	0.027	0.056	0.970	0.971	0.900	0.971	0.865
	100	0.042	0.040	-	0.023	0.058	0.962	0.964	-	0.965	0.766
	200	0.046	0.053	-	0.024	0.046	0.953	0.952	-	0.950	0.578
	400	0.056	0.046	-	0.018	0.050	0.944	0.949	-	0.956	0.396
	800	0.050	0.048	-	0.019	0.048	0.944	0.942	-	0.930	0.260

tests.

5.7 Discussion

This work provides a set of distribution-free tests, as well as providing power analysis and justifying certain optimality for them. Cai et al. (2013) and Cai et al. (2014a) provide related lower bounds for covariance matrix and mean equity tests. We have clearly stated the difference between the results in this work and theirs in Section 5.4.2.

In this work, we assume the regression constants $\{c_{ni}\}_{i=1}^{n}$ and the score function $g(\cdot)$ in (5.2.1), as well as the kernel function $h(\cdot)$ in (5.2.3), to be identical across different pairs of entries. Of note, they can also vary according to the entry (j, k). We do not pursue this direction merely for the clearness of presentation.

We note the problem studied in this work is related to one sample and two sample tests of equality of covariance/correlation matrices and sphericity tests in high dimensions. An extensive literature exists in this line of research. See, Ledoit and Wolf (2002), Chen et al. (2010), Srivastava and Yanagihara (2010), Fisher et al. (2010), Li and Chen (2012), Fisher (2012), Zhang et al. (2013), Cai et al. (2013), among many others. For equity and sphericity tests, the existing methods mostly focus on Pearson's sample covariance matrix. The proposed methods are then based on the statistics characterizing the difference between two sample covariance matrices under different norms, Frobenious and maximum norms for example. As an alternative to the Pearson's sample covariance matrix, Zou et al. (2014) propose a test of sphericity using the multivariate signs. However, the theoretical results in their paper are valid only under the regime $d = O(n^2)$.

Along the research line of this work, an immediate problem is to explore one sample and two sample tests of equality of dependence under different measures of dependence, Kendall's tau and Spearman's rho for example. This is nontrivial because of the possible dependence among random variables. In this work, focusing on the one sample test, we test the bandedness of the latent correlation matrix under the semiparametric Gaussian copula model. We show the test statistic built on the Kendall's tau statistic can asymptotically control the type I error. This test is rate optimal against the sparse alternative. In the future, it is of interest to test the equity of two latent correlation matrices under the Gaussian copula family.

Lastly, in comparison to the focus on estimation procedures in the previous chapters,

we note this chapter considers high dimensional testing problems and highlights robust nonparametric approaches. Because of the particularly nice structure, for conducting tests of independence, we could further relax the semiparametric models to a fully nonparametric one with no modeling constraints at all. On the other hand, the robust nonparametric statistics (such as Kendall's tau and Spearman's rho) still prove to be performing extremely well, and induce minimax optimal tests against sparse alternatives.

Chapter 6

Sparse Median Graphs Estimation in a High Dimensional Semiparametric Model

6.1 Introduction

Undirected graphs provide a powerful tool for understanding the interrelationships among random variables. Given a random vector, $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathcal{R}^d$, the associated conditional independence graph, say $\mathcal{G} \in \{0,1\}^{d \times d}$, is the undirected binary graph so the entry $\mathcal{G}_{jk}$ (for $j \neq k$) is equal to 0 if and only if X_j is conditionally independent of X_k given the remaining variables, $\{X_{\setminus\{j,k\}}\}$. For estimation, it is typically assumed there are n independent and identically distributed realizations of $\boldsymbol{X}$ to infer independence relationships, and thus the associated graph, $\mathcal{G}$.

When $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has a multivariate distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, one obtains the key observation that the non-zero entries of the so-called concentration matrix $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$, otherwise known as its sparsity pattern, encodes the conditional independence structure of $\boldsymbol{X}$ and hence defines the graph $\mathcal{G}$ (Dempster, 1972). In other words, $\mathcal{G}_{jk} = I(\boldsymbol{\Omega}_{jk} \neq 0)$, where $I(\cdot)$ is an indicator function and $\mathcal{G}_{jk}$ indicates whether an edge connects nodes j and k in the graph. Estimation of the concentration matrix becomes problematic in high dimensional settings where $d > n$, thus leading to an active collection of research utilizing sparsity constraints to obtain identifiability (see Friedman et al., 2007; Banerjee et al., 2008; Li and Toh, 2010; Scheinberg et al., 2010; Hsieh et al., 2011; Rothman et al., 2008; Ravikumar et al., 2009; Lam and Fan, 2009; Peng et al., 2009; Meinshausen and Bühlmann, 2006; Yuan, 2010; Cai et al., 2011; Liu et al., 2012a, for example).

However, these papers all assumed the object of inference is a single graph estimated

from a single set of realizations of $\boldsymbol{X}$. In contrast, little work exists on estimation and inference from a population of graphs. Such a setting arises frequently in the sometimes controversial and rapidly evolving arenas of image- and electrophysiologically-based estimates of functional and structural brain connectivity (Friston, 2011; Horwitz et al., 2003; Fingelkurts et al., 2005; Rubinov and Sporns, 2010; Bullmore and Sporns, 2009). Here, each subject-specific graph is an estimate of subject-specific brain connectivity. To date, no theoretically justified definition for population graphs exists.

In addition, frequently the assumption that the data are independently and identically drawn from a Gaussian distribution is too strong. Recently, Gaussian assumptions were relaxed via the so-called *nonparanormal* distribution family (Liu et al., 2009). A random vector, $\boldsymbol{X}$, is said to be nonparanormally distributed if, after an unspecified monotone transformation, it is Gaussian distributed. Moreover, an optimal rate in graph recovery is obtained utilizing the rank-based estimator Kendall's tau (Liu et al., 2012a). On the other hand, however, little has been done in high dimensional graph estimation when the data are actually not identically drawn from a certain distribution.

This work investigates a specific non-i.i.d. setting where the data arise from multiple datasets, each of which is assumed to be distributed under a different distribution. This idea is central in fields, such as epidemiology, where population summaries, are desired over a collections of independently but not identically distributed data sets. A canonical example is the common odds ratio estimate from a collection of individual odds ratios (see Liu and Agresti, 1996, for example). In the motivating application, each dataset is a seed-based or

region of interest summary of functional magnetic resonance imaging (fMRI) scans where a graphical representation of brain connectivity is of interest. The proposed approach does not assume a common underlying graph for each subject. Instead, the population graph defined is a summary, looking at commonalities in graphical structure across a population of heterogeneous graphs. Thus, it is proposed that under the presumption of variation in brain graphical network structure, the investigation of a population graph is of conceptual and practical interest, especially when comparing population graphs across clinical diagnostic categories.

To best summarize the information from aggregated network datasets, the idea of "median graphs" from the pattern recognition field (Bunke and Shearer, 1998; Jiang et al., 2001) is employed. However, it is herein extended to *sparse median graphs*. A sparse median graph is defined as the sparse graph with the smallest sum of Hamming distances to all graphs in a given sample. Combined with the strength of the nonparanormal modeling, a new method for estimating sparse median graphs is proposed. Here we prove the obtained estimator is consistent and the upper bound on the convergence rate with respect to the Hamming distance is established.

In the neuroimaging literature, one relevant paper on summarizing multiple graphical models is Ramsay et al. (2009). There are three main differences between our proposed procedure and the one in Ramsay et al. (2009): (i) On the graph of interest, we focus on the undirected graphical models, while their focus is on the directed graphical models. (ii) On defining the summary graph combining the information from multiple datasets, Ram-

say et al. (2009) proposed a Bayesian information criterion (BIC)-based data aggregation criterion, while we propose a median graph based criterion. Our proposed method is shown to motivate a more robust estimation procedure. (iii) On conducting the algorithm, Ramsay et al. (2009) exploited a greedy search based algorithm (GES), while we exploit a convex optimization based algorithm (CLIME).

The rest of the chapter is organized as follows. In Section 6.2, we introduce the notation and review the nonparanormal distribution and rank-based estimators. In Section 6.3, we introduce the model and give the definition of sparse median graphs. In Section 6.4, we propose the rank-based estimation procedures. Section 6.5 gives the theoretical properties of the proposed procedure for graph recovery. Section 6.6 demonstrates experimental results on both synthetic and real-world datasets to back up our theoretical results. Discussion is provided in the last section.

6.2 Background

Let $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$. Let $\boldsymbol{v}_I$ denote the subvector of $\boldsymbol{v}$ with entries indexed by set I. Similarly, let the submatrix of $\mathbf{M}$ with rows indexed by set I and columns indexed by set J be denoted by $\mathbf{M}_{IJ}$. Let $\mathbf{M}_{I*}$ and $\mathbf{M}_{*J}$ be the submatrix of $\mathbf{M}$ with rows in I, and the submatrix of $\mathbf{M}$ with columns in J. For $0 < q < \infty$, define the

ℓ_q and ℓ_∞ vector norms as:

$$\|\boldsymbol{v}\|_q := \left(\sum_{i=1}^{d} |v_i|^q \right)^{1/q} \text{ and } \|\boldsymbol{v}\|_\infty := \max_{1 \leq i \leq d} |v_i|,$$

and we define $\|\boldsymbol{v}\|_0 := \text{card}\{\text{supp}(\boldsymbol{v})\}$, where $\text{card}\{\cdot\}$ and $\text{supp}(\cdot)$ are the cardinality and support, respectively. Likewise, for matrix norms, we define

$$\|\mathbf{M}\|_q := \max_{\|\boldsymbol{v}\|_q = 1} \|\mathbf{M}\boldsymbol{v}\|_q, \ \|\mathbf{M}\|_{\max} := \max\{|M_{ij}|\}, \text{ and } \|\mathbf{M}\|_H := \sum_{j>k} I(\mathbf{M}_{jk} \neq 0),$$

where $I(\cdot)$ denotes the indicator function. We define $\text{diag}(\mathbf{M})$ to be a diagonal matrix with diagonal values same as $\mathbf{M}$ and all off-diagonal values to be zero.

6.2.1 The Nonparanormal

Liu et al. (2009) and Liu et al. (2012a) showed the Gaussian graphical model can be relaxed to the nonparanormal graphical model without significant loss of inference power when the data are Gaussian distributed and with significant gain of inference power when it is not. This observation plays a role in our proposed model for relaxing the Gaussian assumption. In this section, the nonparanormal distribution family is introduced with the corresponding graphical model, following definitions in Liu et al. (2012a).

Definition 6.2.1 (The nonparanormal)**.** *Let $f = \{f_j\}_{j=1}^d$ be a set of univariate strictly increasing functions. A d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is said to follow*

a nonparanormal distribution, denoted

$NPN_d(\mathbf{\Sigma}, f)$*, if and only if*

$$f(\boldsymbol{X}) := \{f_1(X_1), \ldots, f_d(X_d)\}^T \sim N_d(\mathbf{0}, \mathbf{\Sigma}), \text{ where } \mathrm{diag}(\mathbf{\Sigma}) = \mathbf{I}_d,$$

where $\mathbf{I}_d \in \mathcal{R}^{d \times d}$ *is the identity matrix.* $\mathbf{\Sigma}$ *is referred to as the latent correlation matrix and* $\mathbf{\Omega} := \mathbf{\Sigma}^{-1}$ *as the latent concentration matrix.*

Although the nonparanormal distribution family is strictly larger than the Gaussian distribution family, Liu et al. (2009) showed the conditional independence property of the nonparanormal is still encoded in the latent concentration matrix $\mathbf{\Omega}$. More specifically, they noted if the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is nonparanormally distributed (i.e. $\boldsymbol{X} \sim NPN_d(\mathbf{\Sigma}, f)$) then

$$X_j \perp X_k \,|\, X_{\setminus j,k} \Leftrightarrow \mathbf{\Omega}_{jk} = 0. \tag{6.2.1}$$

6.2.2 Rank-based Estimator

Liu et al. (2012a) and Xue and Zou (2012) exploited the rank-based estimator, Kendall's tau, in inferring the latent concentration matrix $\mathbf{\Omega}$ in the nonparanormal family. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be d-dimensional random vectors, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$ for $i = 1, \ldots, n$, be n ob-

served data points of a random vector $\boldsymbol{X}$. The Kendall's tau statistic is defined as:

$$\widehat{\tau}_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) := \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \operatorname{sign}(x_{ij} - x_{i'j}) \cdot (x_{ik} - x_{i'k}). \tag{6.2.2}$$

The Kendall's tau statistic is monotone transformation-invariant correlation between the empirical realizations of X_j and X_k for any $j, k \in \{1, \ldots, d\}$. Let $\widehat{\mathbf{R}} = [\widehat{\mathbf{R}}_{jk}] \in \mathcal{R}^{d \times d}$, with

$$\widehat{\mathbf{R}}_{jk} = \sin\left(\frac{\pi}{2} \widehat{\tau}_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\right), \tag{6.2.3}$$

be the Kendall's tau matrix. Liu et al. (2012a) showed if $\boldsymbol{X}$ is nonparanormally distributed, $\widehat{\mathbf{R}}$ is a consistent estimator of the latent correlation matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$ (with respect to element-wise sup norm $\|\cdot\|_{\max}$), even when the order of d is nearly exponentially larger than n.

Since the latent concentration matrix, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, fully encodes the nonparanormal graphical model and $\widehat{\mathbf{R}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, Kendal's tau is a good estimate of the nonparanormal graphical model, as it directly estimates the latent concentration matrix. Based on the Kendall's tau matrix, Liu et al. (2012a) proposed the nonparanormal SKEPTIC by directly plugging $\widehat{\mathbf{R}}$ into any statistical methods in calculating the inverse covariance/correlation matrix. In this work, we will focus on one particular statistical method, CLIME (Cai et al., 2011). Further details of the nonparanormal SKEPTIC are given in Section 6.4.

6.3 Models and Concepts

6.3.1 Models

In this section, the proposed approach for modeling the complex aggregated data is given. Assume the data are aggregated from multiple datasets, each of which is distributed according to a different nonparanormal distribution.

More specifically, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ be T random vectors with $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{td})^T$ satisfying

$$\boldsymbol{X}_t \sim NPN_d(\boldsymbol{\Sigma}^t, f^t), \quad \text{for } t = 1, \ldots, T.$$

Let $\boldsymbol{\Theta}^t := [\boldsymbol{\Sigma}^t]^{-1}$. Based on $\boldsymbol{\Theta}^t$, we define $\mathcal{G}^t = [\mathcal{G}^t_{jk}] \in \{0,1\}^{d \times d}$ where

$$\mathcal{G}^t_{jk} = 0 \text{ if and only if } \boldsymbol{\Theta}^t_{jk} = 0.$$

Via Equation 6.2.1, $\mathcal{G}^t$ represents the Markov graph associated with $\boldsymbol{X}_t$. In detail, the pair (j,k) such that $\mathcal{G}^t_{jk} \neq 0$ indicates the conditional independence of X_{tj} and X_{tk} given all the rest in $\boldsymbol{X}_t$.

6.3.2 Sparse Median Graphs

This section introduces the concept of a *sparse median graph*, combining the ideas of median graphs from Jiang et al. (2001) and the sparsity concept commonly adopted in high

dimensional statistics (Bühlmann and van de Geer, 2011).

Let $\boldsymbol{d}(\cdot) : \{0,1\}^{d\times d} \times \{0,1\}^{d\times d} \to [0,\infty)$ be a distance function on the graph space. Jiang et al. (2001) define the median graph (reproduced in Definition 6.3.1) as the graph with the smallest sum of distances to all graphs in a given set.

Definition 6.3.1 (Median Graph)**.** *Let $\mathcal{G}^1, \ldots, \mathcal{G}^T$ be T different binary graphs in $\{0,1\}^{d\times d}$, the median graph $\mathcal{G}^*$ is defined by*

$$\mathcal{G}^* := \underset{\mathcal{G}\in\{0,1\}^{d\times d}}{\arg\min} \sum_{t=1}^{T} \boldsymbol{d}(\mathcal{G}, \mathcal{G}^t). \tag{6.3.1}$$

When T is large, $\mathcal{G}^*$ will not be sparse and therefore the resulting median graph may not be interpretable. To attack this problem, consider the concept of a "sparse median graph". The sparse median graph is the graph with the smallest sum of distances to all graphs in a given set, and the non-zero entries in the graph is less than or equal to some small value $s \ll d^2$. In particular, we use the Hamming distance $\|\cdot\|_H$ in calculating the distance of any two graphs.

Definition 6.3.2 (Sparse Median Graph)**.** *Let $\{\mathcal{G}^1, \ldots, \mathcal{G}^T\}$ be T different binary graphs. The sparse median graph $\mathcal{G}_s^*$ is defined as*

$$\mathcal{G}_s^* := \underset{\mathcal{G}\in\{0,1\}^{d\times d}, \|\mathcal{G}\|_H = s}{\arg\min} \sum_{t=1}^{T} \|\mathcal{G} - \mathcal{G}^t\|_H, \tag{6.3.2}$$

where $\|\cdot\|_H$ represents the number of non-zero entries in the upper triangle of the matrix

of interest.

The next proposition presents an equivalent representation of G_s^* and further discusses identifiablility conditions of the model.

Proposition 6.3.3. *Let $\mathcal{G}^t, t = 1, \ldots, T$ and $\mathcal{G}_s^*$ be the sparse median graph defined as above. Let $\zeta_{jk} = \sum_t \mathcal{G}_{jk}^t$ and r_{jk} be the rank of all values $\{\zeta_{jk}\}_{j<k}$. Then:*

$$[\mathcal{G}_s^*]_{jk} = [\mathcal{G}_s^*]_{kj} = \begin{cases} 1, & \text{if } r_{jk} \leq s, \\ 0, & \text{if } r_{jk} > s. \end{cases}$$

Moreover, the model is identifiable with respect to $\mathcal{G}_s^*$ if and only if there are no ties around rank s for the sequence $\{\zeta_{jk}\}_{j<k}$.

Remark 6.3.4. *The population sparse median graph is defined as the optimum of a specified loss function with regard to the Hamming distance. This is a common approach for representing a summary of multiple, possibly heterogenous, data points. In principle, there are potential issues by aggregation, such as averaging out effects when both positive and negative ones exist. However, since we focus only on undirected graphs taking values $\{0, 1\}$, such issues could be minimized. Actually, the robustness to aggregation issues is one strong advantage motivating the sparse median graph. In a later section (Section 6.2), we will further illustrate the empirical power of using the notion of sparse median graph combined with robust estimation.*

6.4 Methods

For $t = 1, \ldots, T$, let $\boldsymbol{x}_i^t = (x_{i1}^t, \ldots, x_{id}^t)^T, i = 1, \ldots, n_t$ be n_t independent realizations of $\boldsymbol{X}_t$ (defined in Section 6.3.1). The observed data are $\{\boldsymbol{x}_i^t\}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, n_t$ and the target is to estimate the sparse median graph $\mathcal{G}_s^*$ defined in Equation (6.3.2). The proposed method is a two step procedure. In the first step, the nonparanormal SKEPTIC is used to obtain the estimators $\{\widehat{\mathcal{G}}^t\}_{t=1}^T$ of $\{\mathcal{G}^t\}_{t=1}^T$. In the second step, $\mathcal{G}_s^*$ is estimated based on the estimators $\{\widehat{\mathcal{G}}^t\}_{t=1}^T$ obtained in the first step.

More specifically, in the first step, for each $t \in \{1, 2, \ldots, T\}$, let

$$\widehat{\mathbf{R}}_{jk} := \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_{n_t}^t)\right),$$

where $\widehat{\tau}_{jk}(\cdot)$ is defined in Equation (6.2.2). By using $\widehat{\mathbf{R}}^t = [\widehat{\mathbf{R}}_{jk}^t] \in \mathbb{R}^{d\times d}$ to estimate $\boldsymbol{\Sigma}^t$, one can plug $\widehat{\mathbf{R}}^t$ into CLIME to get estimates of $\boldsymbol{\Omega}^t$ and $\mathcal{G}^t$:

$$\widehat{\boldsymbol{\Omega}}^t = \arg\min_{\mathbf{M}} \sum_{j,k} |\mathbf{M}_{jk}| \qquad (6.4.1)$$
$$\text{such that } \|\widehat{\mathbf{R}}^t\mathbf{M} - \mathbf{I}_d\|_{\max} \leq \lambda_t,$$

where $\lambda_t > 0$ is a tuning parameter. Cai et al. (2011) showed this optimization can be decomposed into d vector minimization problems, each of which can be reformulated as a linear program. Thus, it has the potential to scale to very large problems. Once $\widehat{\boldsymbol{\Omega}}^t$ is obtained, one can apply an additional thresholding step to estimate the graph, $\mathcal{G}^t$. For this,

the graph estimator $\widehat{\mathcal{G}}^t \in \{0,1\}^{d\times d}$ is defined, in which a pair (j,k) satisfies that $\widehat{\mathcal{G}}^t_{jk} \neq 0$ if and only if $\widehat{\Omega}^t_{jk} > \gamma_t$. Here γ_t is another tuning parameter. However, in practice, the CLIME algorithm works well without a second step truncation.

In the second step, provided the estimates $\{\widehat{\mathcal{G}}^t, t = 1,\ldots,T\}$ have been obtained, the following equation is optimized to obtain $\widehat{\mathcal{G}}^*_s$

$$\widehat{\mathcal{G}}^*_s = \underset{\mathcal{G}\in\{0,1\}^{d\times d}, \|\mathcal{G}\|_H = s}{\arg\min} \quad \sum_t \|\mathcal{G} - \widehat{\mathcal{G}}^t\|_H, \tag{6.4.2}$$

where the term $\|\mathcal{G}\|_H = s$ controls the sparsity degree of $\mathcal{G}$. In this work, it is assumed s is known. Consider then the following proposition, which states Equation (6.4.2) has a closed-form solution.

Proposition 6.4.1. *Let $\widehat{\zeta}_{jk}$ be defined as $\widehat{\zeta}_{jk} := \sum_t \widehat{\mathcal{G}}^t_{jk}$. Let $(j_1,k_1),(j_2,k_2),\ldots$ be s pairs with the highest values in $\{\widehat{\zeta}_{jk}\}_{j<k}$. Then $\widehat{\mathcal{G}}_{jk} = 1$ if and only if $(j,k) \in \{(j_1,k_1),(j_2,k_2),\ldots\}$.*

Remark 6.4.2. *For simplicity, it is assumed there are no ties around the rank s for the sequence $\{\widehat{\zeta}_{jk}\}$. If the model discussed in Section 6.3 is identifiable and several mild conditions as shown in Section 6.5 hold, then there are no ties with high probability.*

6.5 Theoretical Properties

In this section, the estimators from Section 6.4 are proved to be consistent for the true median graph. Notably, an nonasymptotic bound on the rate of convergence in estimating

the sparse median graph with respect to the Hamming distance is provided.

Additional notation is required. Let M_d be a quantity which may scale with the dimension d. Define

$$\mathcal{S}_d(q, s, M_d) := \left\{ \mathbf{\Omega} : \|\mathbf{\Omega}\|_1 \leq M_d \text{ and } \max_{1 \leq j \leq d} \sum_{k=1}^{d} |\mathbf{\Omega}_{jk}|^q \leq s \right\}.$$

For $q = 0$, the class $\mathcal{S}_d(q, s, M_d)$ contains all the s-sparse matrices. The next theorem provides the parameter estimation and graph estimation consistency results for the non-paranormal SKEPTIC estimator defined in Equation (6.4.1).

Theorem 6.5.1 (Liu et al. (2012a))**.** *Let* $\boldsymbol{X}^t \sim NPN_d(\mathbf{\Sigma}^t, f^t)$ *with* $\mathbf{\Omega}^t := [\mathbf{\Sigma}^t]^{-1} \in \mathbb{S}_d(q, s_t, M_d)$ *with* $0 \leq q < 1$. *Let* $\widehat{\mathbf{\Omega}}^t$ *be defined in Equation* (6.4.1). *There exist constants,* C_0 *and* C_1*, only depending on* q*, such that whenever one chooses the tuning parameter* $\lambda_t = C_0 M_d \sqrt{\frac{\log d}{n_t}}$*, with probability no less than* $1 - d^2$*,*

$$\|\widehat{\mathbf{\Omega}}^t - \mathbf{\Omega}^t\|_2 \leq C_1 M_d^{2-2q} \cdot s \cdot \left(\frac{\log d}{n_t} \right)^{(1-q)/2}.$$

Let $\widehat{\mathcal{G}}^t$ *be the graph estimator defined in Section 6.4 with the second tuning parameter* $\gamma_t := 4 M_d \lambda_t$*. If it is further assumed* $\mathbf{\Omega} \in \mathcal{S}_d(0, s, M_d)$ *and* $\min_{j,k:\mathbf{\Omega}_{jk} \neq 0} |\mathbf{\Omega}_{jk}| \geq 2\gamma_t$*, then*

$$\mathbb{P}(\widehat{\mathcal{G}}^t \neq \mathcal{G}^t) \leq 4 d^{-\epsilon_1},$$

where $\epsilon_1 > 0$ *is a constant that does not depend on* (n_t, d, s_t).

Proof. Combine the Theorems 1 and 7 in Cai et al. (2011) and Theorem 4.2 in Liu et al. (2012a). □

Theorem 6.5.2 (Consistency)**.** *With the above notation, the assumptions from Theorem 6.5.1, λ_t, γ_t fixed and the model in Section 6.3 is identifiable, then*

$$\mathbb{P}(\widehat{\mathcal{G}}_s^* \neq \mathcal{G}_s^*) \leq 4Td^{-\epsilon_1}, \tag{6.5.1}$$

where $\widehat{\mathcal{G}}_s^*$ is defined as in Equation (6.4.2)*.*

Proof. If the model is identifiable, then one only needs to show with high probability, all $\mathcal{G}^t$ can be recovered. Note the union bound in Theorem 6.5.1 yields

$$\mathbb{P}\left(\bigcup_{t=1}^{T}\{\widehat{\mathcal{G}}^t \neq \mathcal{G}^t\}\right) \leq \sum_{t=1}^{T}\mathbb{P}(\widehat{\mathcal{G}}^t \neq \mathcal{G}^t) \leq 4d^{-\epsilon_1} \leq 4Td^{-\epsilon_1}.$$

We thus finish the proof. □

The next theorem provides an upper bound of the rate of convergence with respect to the Hamming distance. Such a result is based on the recent explorations in graph recovery with respect to the Hamming distance (Ke et al., 2012; Jin et al., 2012).

Theorem 6.5.3 (Rate of convergence)**.** *Assume the above assumptions in Theorems 6.5.1 and 6.5.2 hold. Let $\mathcal{A}_t$ be the event that*

$$\mathcal{A}_t := \{\|\widehat{\mathcal{G}}^t - \mathcal{G}^t\|_H \leq \delta_t\}$$

and δ_t *be defined such that* $\mathbb{P}(\mathcal{A}_t) = 1 - o(d^{-\epsilon_2})$. *Moreover, reorder* $\{\zeta_{jk}\}_{j<k}$ *to be* $\zeta^{(1)} \geq \zeta^2 \geq \cdots \geq \zeta^{d(d-1)/2}$ *and let* $u^* = (\zeta^{(s)} - \zeta^{(s+1)})/2$. *Then,*

$$\mathbb{P}\left(\|\mathcal{G}_s^* - \widehat{\mathcal{G}}_s^*\|_H \leq \frac{2\sum_{t=1}^T \delta_t}{u^*} \right) = 1 - o(Td^{-\epsilon_2}). \tag{6.5.2}$$

Remark 6.5.4. *The bound constructed in Equation* (6.5.2) *is to balance the difference of* $\{\mathcal{G}^t\}_{t=1}^T$ *to* $\mathcal{G}_s^*$ *and the estimation error of* $\widehat{\mathcal{G}}^t$ *to* $\mathcal{G}^t$. *In other words, the better it is to differentiate* $\{\mathcal{G}^t\}$ *with* $\mathcal{G}_s^*$ *in the population level and the more accuracy* $\widehat{\mathcal{G}}^t$ *can approach* $\mathcal{G}^t$, *the better the final estimator can converge to the sparse median graph.*

6.6 Empirical Results

In this section, we investigate the performance of the proposed method compared to the performance of alternative methods on synthetic and real-world datasets. Since we aim to estimate a summary graph throughout multiple possibly non-i.i.d. datasets, our estimation procedure in general involves two steps: In the first step, for each specific dataset, we employ a graphical model estimation procedure; In the second step, based on the calculated graph estimates, we obtain a single estimate of the summary graph. We call the former step the "estimation of graphs" part, and the later step the "combination[1] of datasets" part. In the following simulations and experiments, we will compare our methods with multiple candidates using different graph estimation and datasets combination approaches, and

[1] We also use the terms "aggregation" and "summarization" synonymously in this work.

reveal the advantage of our proposed one.

6.6.1 Estimation Methods

In our simulations and experiments, we consider the methods Kendall, Pearson, and LW to estimate graphs (or correlation matrices) on individual datasets. To combine multiple datasets, we employ SMG, Naive, and Average. Therefore, we will compare a total of nine methods, each of which denoted by first stating the aggregation method and then the graph estimation method. For example, our proposed method corresponds to SMG Kendall. We elaborate the details of the competing methods as follows.

6.6.1.1 Estimation of Graphs

For any individual dataset, we consider the following approaches for graph estimation:

- Kendall: This method calculates the Kendall's tau correlation matrix and plugs the matrix into CLIME. Details are in Section 6.4.

- Pearson: This method follows the same steps as Kendall except that we plug the Pearson sample correlation matrix into CLIME instead.

- Ledoit-Wolf (LW): Using the `tawny` package (Rowe, 2014), this method calculates a Ledoit-Wolf shrinkage estimation (Ledoit and Wolf, 2003) of the covariance matrix of the dataset, $\widehat{\mathbf{\Sigma}}$, and a corresponding precision matrix, $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Sigma}}^{-1}$. With $\widehat{\mathbf{\Theta}}$, we estimate a graph $\widehat{\mathcal{G}} \in \{0,1\}^{d \times d}$, in which a pair (j,k) satisfies $\widehat{\mathcal{G}}_{jk} \neq 0$ if and only if $\widehat{\mathbf{\Theta}}_{jk} > (0.001 \times \text{Avg}(\widehat{\mathbf{\Theta}}))$, where $\text{Avg}(\widehat{\mathbf{\Theta}}) := \frac{d(d-1)}{2} \sum_{(j,k)} \widehat{\mathbf{\Theta}}_{jk}$.

We select the tuning parameters $\{\lambda_t\}$ in CLIME[2] using the StARS stability-based approach (Liu et al., 2010). StARs selects a tuning parameter that simultaneously makes a graph sparse and replicable under random sampling. The detailed procedure could be found in Section 3.2 in Liu et al. (2010).

6.6.1.2 Combination of Datasets

Our experiments involve inference on T datasets, where each dataset corresponds to a different subject. We consider the following approaches to estimate one sparse graph across the multiple datasets:

- Sparse Median Graph (SMG): We estimate a graph for each of the T datasets. Then, given some sparsity s, we combine these graphs with the method proposed in Section 6.3.2 to obtain a sparse median graph.

[2]Recall the formal defintion of CLIME also require a set of thresholding parameters $\{\gamma_t\}$. While thresholding by γ_t is indeed a valid option, we choose to use the method of setting every non-zero entry in the precision matrix to correspond to an edge. Therefore, we do not actually use γ_t in our simulations and experiments. In practice, we have found that the use of some threshold γ_t has very little impact on the output of the method.

- Naive: We concatenate the T datasets into one dataset on which we estimate a graph using the techniques from Section 6.6.1.1.

- Average: For each of the T datasets, we calculate an associated correlation or precision matrix. We average these matrices and threshold such that only the s entries in the averaged matrix with the largest magnitudes correspond to edges in the estimated graph.

6.6.2 Synthetic Data Simulations

In this simulation, we examine the estimation performance of the proposed method on synthetically generated data. In particular, we generate $T = 15$ different datasets with 100 samples in each dataset. Each dataset follows a different nonparanormal distribution, corresponding to a different undirected graph $\mathcal{G}^t$. For each method, we utilize a sequence of uniformly spaced sparsity parameters $\widehat{s}$ from 0 to $\binom{d}{2}$ to estimate a sequence of graphs, over which we plot a ROC curve. In addition, we repeat this simulation for $d = 50$, 100, and 250. Our results show SMG Kendall exhibit better estimation performance than the competiting methods.

More specifically, we conduct the simulation with the following procedure:

1. Using the huge package (Zhao et al., 2012), we generate a sparse graph $\mathcal{G}_s^*$ with sparsity s, along with a corresponding covariance matrix $\mathbf{\Sigma}$. We will use this as the

oracle graph of the population. In particular, we adopt the following five models for $\mathcal{G}_s^*$: banded, clustered, hub, random, and scale-free (definitions are provided in Zhao et al. (2012)). We then examine $\mathcal{G}_s^*$ at $d = 50$, 100, and 250.

2. For each subject $t = 1, 2, ..., T$, we construct a perturbed graph $\mathcal{G}^t$ to reflect the difference among different subjects. In particular, we add $\lfloor 0.001 \times s \rfloor$ edges and remove $\lfloor 0.75 \times (\binom{d}{2} - s) \rfloor$ edges from $\mathcal{G}_s^*$. We illustrate a typical run of the generated graphs $\mathcal{G}^t$ for a specific t in Figures 6.1, 6.2, and 6.3. In each figure, the black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.

3. Using each $\mathcal{G}^t$, we generate a corresponding covariance matrix $\mathbf{\Sigma}^t$ with an algorithm identical to the one implemented in the huge package.

4. For $t = 1, \ldots, T$, we generate a $(n_t \times d)$ dataset[3] $\mathcal{D}^t$ from $NPN_d(\mathbf{\Sigma}^t, f)$ where $f_1(x) = \cdots = f_d(x) = x^5$. Thus, the population dataset is:

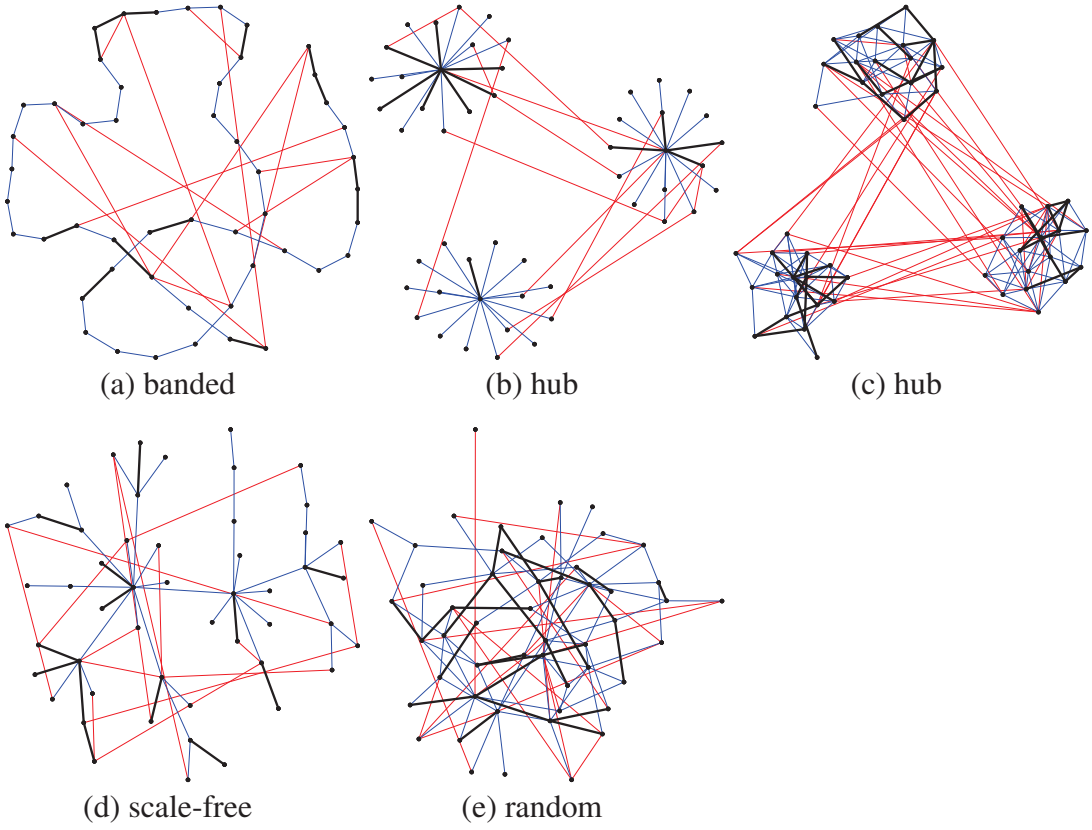
$$\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^t, \ldots, \mathcal{D}^T\}$$

5. Applying the nine methods described in Section 6.6.1 to $\mathcal{D}$, we estimate a sparse graph, $\widehat{\mathcal{G}}_{\widehat{s}}^*$, and calculate the the true positive and true negative rates.

6. We repeat the simulation 100 times and plot an averaged ROC curve over the range

[3]Each $\mathcal{D}^t$ corresponds to a realization $\{\boldsymbol{x}_i^t\}$ for $i = 1, \ldots, n_t$ from Section 6.4.

of $\widehat{s}$. We show the results in Figures 6.4, 6.5, and 6.6 .

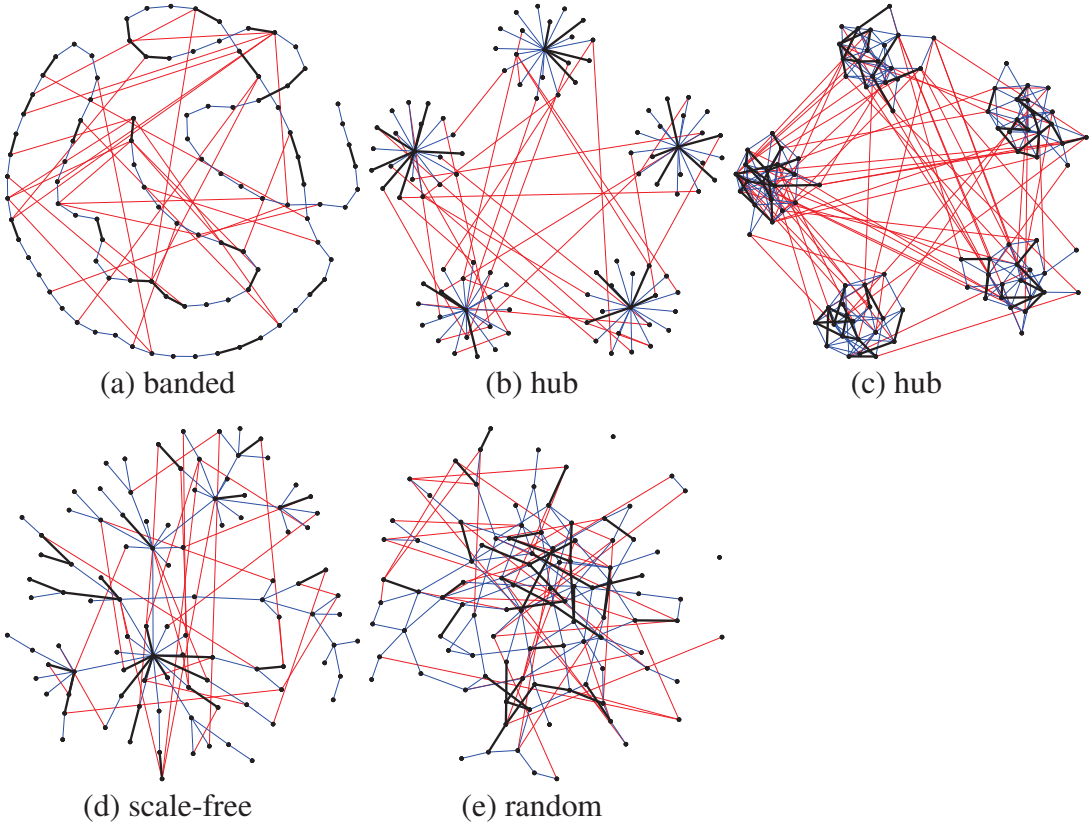


(a) banded (b) hub (c) hub

(d) scale-free (e) random

Figure 6.1: An illustration of the five graph patterns of the sparse graphs $\mathcal{G}_s^*$ and the corresponding one individual dataset's graph $\mathcal{G}^t$ for $d = 50$. Here the black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.

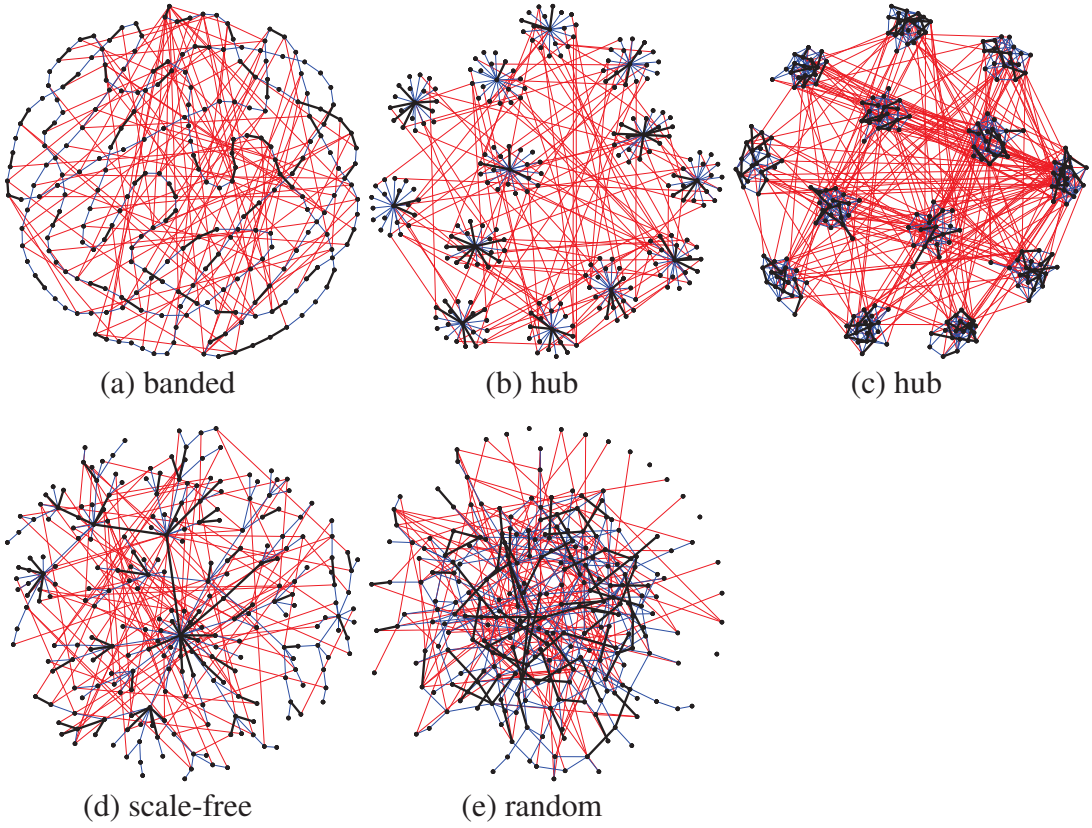
From the curves in Figures 6.4, 6.5, and 6.6, we clearly see our proposed method exhibits a higher estimation performance than competing methods. This is as expected because the proposed method is the only consistent estimator of $\mathcal{G}_s^*$, while all the competing methods deviate from the truth. In addition, note the Kendall-based methods tend to outperform Pearson-based methods – a pattern that becomes more distinct with larger di-



(a) banded (b) hub (c) hub

(d) scale-free (e) random

Figure 6.2: An illustration of the five graph patterns of the sparse graphs $\mathcal{G}_s^*$ and the corresponding one individual dataset's graph $\mathcal{G}^t$ for $d = 100$. Here the black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.



(a) banded (b) hub (c) hub

(d) scale-free (e) random

Figure 6.3: An illustration of the five graph patterns of the sparse graphs $\mathcal{G}_s^*$ and the corresponding one individual dataset's graph $\mathcal{G}^t$ for $d = 250$. Here the black edges represent the ones present in both $\mathcal{G}_s^*$ and $\mathcal{G}^t$, the blue edges represent the ones only present in $\mathcal{G}_s^*$, and the red edges represent the ones only present in $\mathcal{G}^t$.

mensions. This result confirms the claim that utilizing Kendall's tau leads to optimal graph recovery rates (Liu et al., 2012a). Furthermore, the poor performance of the LW-based methods (worse than both Kendall and Pearson) suggests, while covariance shrinkage demonstrate potential in financial applications, their benefits do not carry over to graph estimation.

6.6.3 ADHD Data Experiments

In practice, there exists no golden standard for the structure of the graph of brain imaging data. Therefore, in addition to the above simulation on synthetic data, we investigate the estimation performance, predictive power, and stability of the proposed method on a brain imaging dataset, the ADHD-200 dataset (Milham et al., 2012; Eloyan et al., 2012).

The ADHD-200 dataset is a landmark study compiling over 1,000 functional and structural scans including subjects with and without attention deficit hyperactive disorder (ADHD). The data used in the analysis are from 739 unique subjects: 478 controls and 261 children diagnosed with ADHD of various subtypes. Each subject has at least one blood oxygen level dependent (BOLD) resting state functional MRI scans. The number of scans within an fMRI resting state session varies from 78 to 456, which were measured with different time resolutions (TR) as well as different scan lengths. The varying TR and length of scanning stress the importance of addressing subject-level heteroscedasticity in graph estimates. The data also include demographic variables as predictors. These include age, IQ, gender and handedness. These demographic variables are combined into a matrix with dimen-

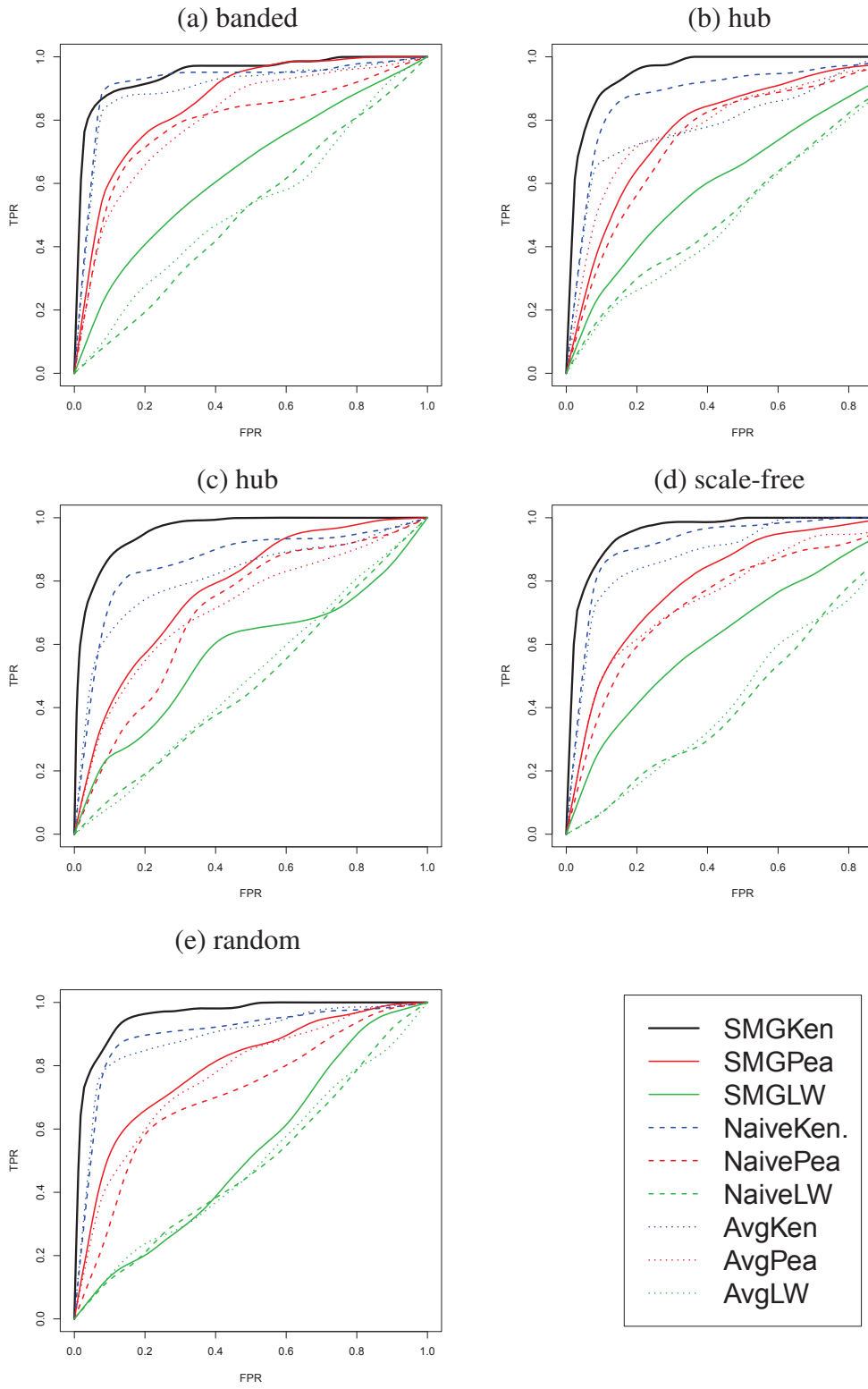


Figure 6.4: ROC curves in estimating the graphical models for different methods in five different graph patterns. Here, $d = 50$ and $n_t = 100$ for all $t = 1, 2, \ldots, 15$.

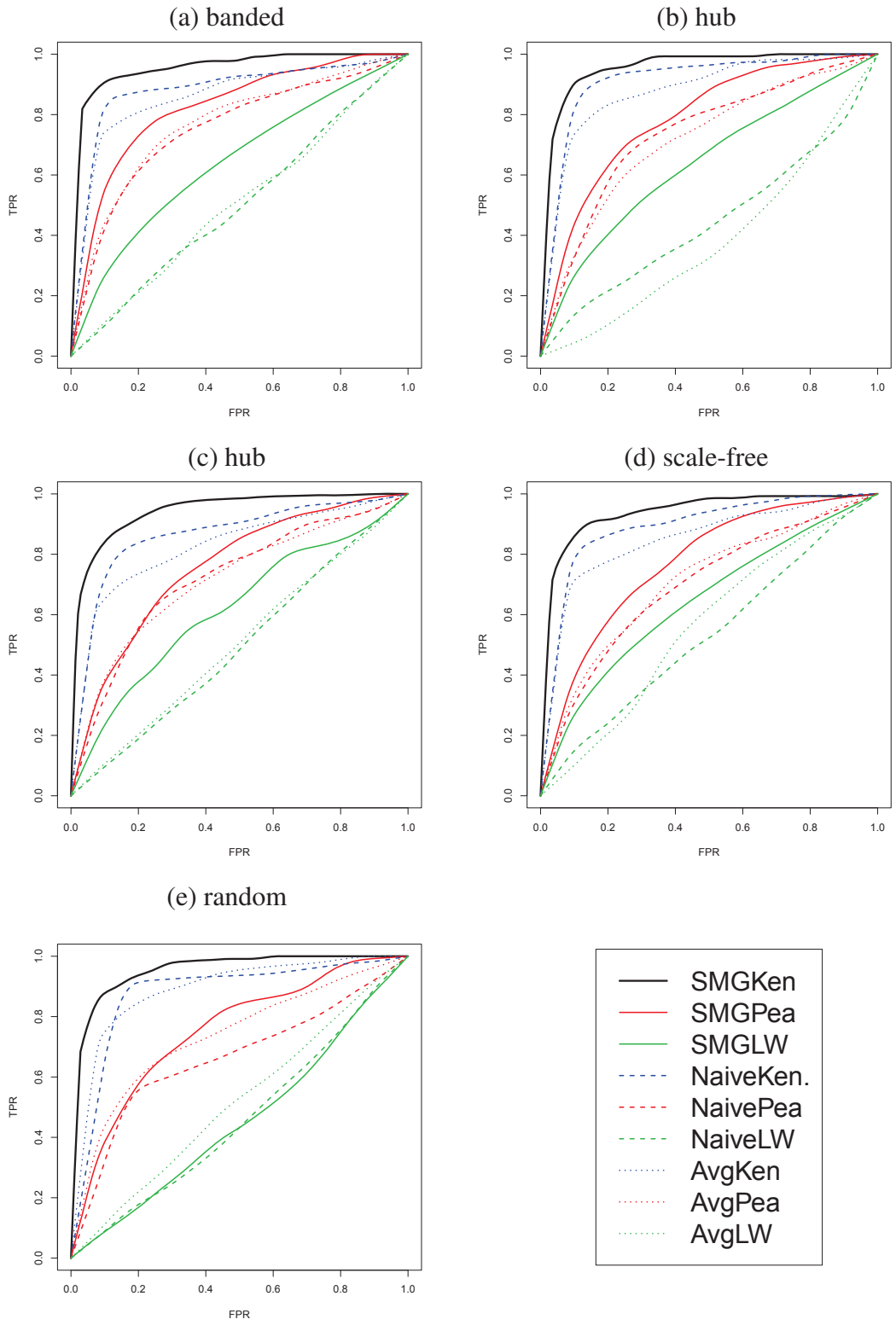


Figure 6.5: ROC curves in estimating the graphical models for different methods in five different graph patterns. Here, $d = 100$ and $n_t = 100$ for all $t = 1, 2, \ldots, 15$.

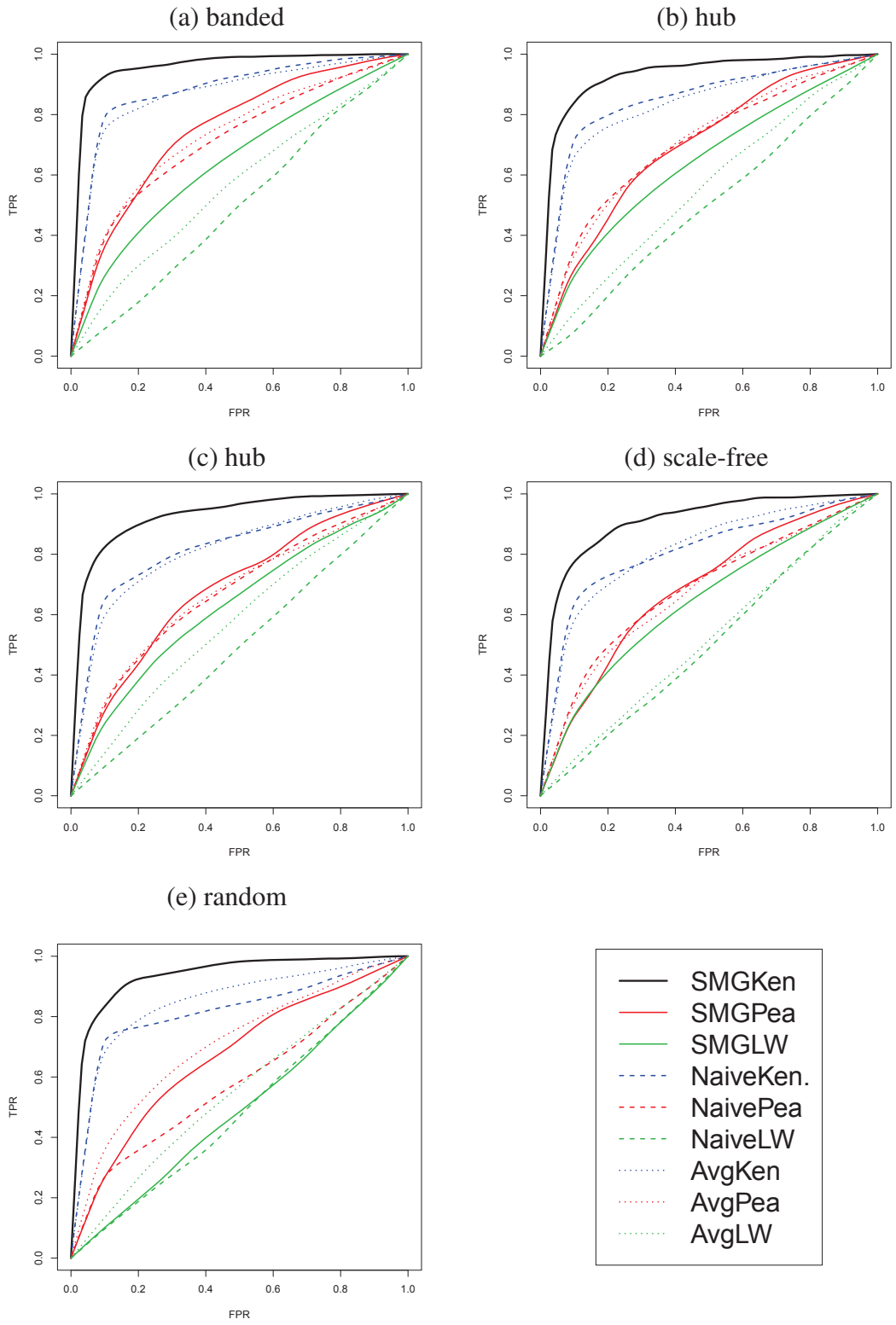


Figure 6.6: ROC curves in estimating the graphical models for different methods in five different graph patterns. Here, $d = 250$ and $n_t = 100$ for all $t = 1, 2, \ldots, 15$.

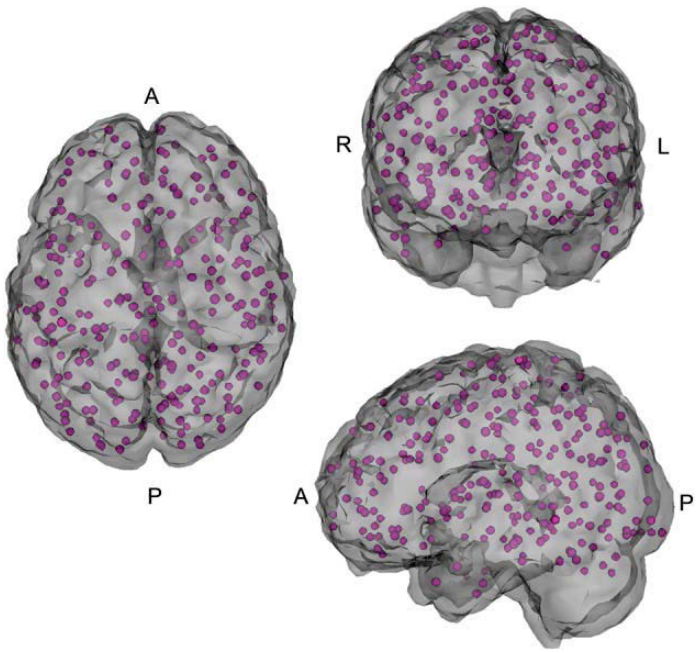


Figure 6.7: The illustration of the locations of the 264 nodes.

sions 4×739. We follow the procedure in Eloyan et al. (2012) for preprocessing with one additional step of concatenating together datasets associated with the same `patientID`.

We constructed our predictors by extracting 264 voxels from each image that broadly cover major functional regions of the cerebral cortex and cerebellum following Power et al. (2011). The locations of these 264 voxels are illustrated in Figure 6.7 and the value of each voxel is calculated as the mean of all data points inside these small seed regions. Therefore, each subject t would correspond to a matrix of size $n_t \times d$ where n_t is the number of images for that subject and $d = 264$.

6.6.3.1 Simulations based on the ADHD Data

Here, we examine the estimation performance of the proposed method on real brain imaging data. This involves three steps: First, we need to generate a "true graph"; Secondly, we simulate multiple datasets with the sparse median graph corresponding to the "true graph"; Thirdly, we examine the estimation performance based on the simulated multiple datsets.

Specifically, we first estimate a sparse graph on a homogenous dataset and use this graph as the "true graph." Then, we simulate non-identically distributed subjects by partitioning the homogenous dataset and adding perturbations to each partition. Using these simulated datasets, we will assess the estimation performance of the nine methods from Section 6.6.1 with a simulation similar to that in Section 6.6.2. Our results confirm SMG Kendall continues to exhibit better estimation performance than the competiting methods when the data originates from real brain imaging data.

In particular, we use the brain imaging data of the subject with the patient ID 15002 in the ADHD dataset. This patient possesses the largest number of scans in the dataset with 456 images. We denote this dataset by $\mathcal{D}$. Then, we implement the following simulation procedure:

1. Using the Kendall method described in Section 6.6.1.1, we estimate an oracle sparse median graph $\mathcal{G}_s^*$ on $\mathcal{D}$ with the s parameter chosen using StARs.

2. To simulate different datasets, we randomly partition $\mathcal{D}$ into $T = 10$ smaller datasets

$\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_t, ..., \mathcal{D}_T\}$. This creates six sets of 46 scans and four sets of 45 scans– each with $d = 264$ corresponding to the number of voxels.

3. For each patient $t = 1, 2, ..., T$, we generate a graph $\mathcal{G}^t$ for the patient by removing edges from $\mathcal{G}_s^*$. More specifically, we select a $p_r := 50\%$ of the d vertices in $\mathcal{G}_s^*$ randomly, and delete all edges incident to these vertices.

4. Let μ and σ denote the mean and standard deviation of the vectorized $\mathcal{D}$. Note that each of the $\lfloor p_r \times d \rfloor$ randomly selected vertices correspond to a column in the datasets. We perturb each $\mathcal{D}_t$ to match $\mathcal{G}^t$ by replacing each entry of the randomly selected columns with a number randomly generated from the distribution $N(\mu, \sigma)$. Let us denote this perturbed dataset as $\widetilde{\mathcal{D}}_t$.

5. To simulate the effects of outliers, we choose 30% of the rows in each $\widetilde{\mathcal{D}}_t$ and apply the following transformation to each entry in a chosen row, i, in the dataset:

$$[\widehat{\mathcal{D}}_t]_{ij} = [\widetilde{\mathcal{D}}_t]_{ij}^5 \times \frac{\sum_{k=1}^d [\widetilde{\mathcal{D}}_t]_{ik}}{\sum_{k=1}^d [\widetilde{\mathcal{D}}_t]_{ik}^5}$$

   In rows that were not chosen, the entries of $\widehat{\mathcal{D}}_t$ and $\widetilde{\mathcal{D}}_t$ are identical. Therefore, $\widehat{\mathcal{D}}^t$ is the final perturbed dataset for one particular subject, and the dataset of all simulated subjects is:

$$\widehat{\mathcal{D}} = \{\widehat{\mathcal{D}}_1, \widehat{\mathcal{D}}_2, ..., \widehat{\mathcal{D}}_t, ..., \widehat{\mathcal{D}}_T\}$$

6. Applying the nine methods described in Section 6.6.1 to $\widehat{\mathcal{D}}$, we estimate a sparse

graph, $\widehat{\mathcal{G}}^*_{\widehat{s}}$, and calculate the the true and false positive rates.

7. We repeat the simulation 100 times and plot an averaged ROC curve over the range of $\widehat{s}$. The results are shown in Figure 6.8.

Comparing the results from Figure 6.8 to those in Section 6.6.2, we see the proposed method continues to demonstrate the best estimation performance, and the LW-based methods continue to perform the worst among these competing methods. However, in this simulation, SMG Kendall and SMG Pearson outperform Naive Kendall and Naive Pearson, where each Kendall-based method still outperforms the corresponding Pearson-based method. This shows the benefits of sparse median graphs tend to dominate when estimating graphs on real brain imaging data – unlike the synthetic setting where the benefits of utilizing Kendall's tau tend to dominate. Nonetheless, the results from this simulation and Section 6.6.2 both agree that the best estimation performance is achieved by the proposed method.

6.6.3.2 Predictive Power Experiment

In this section, we compare the predictive power of our proposed method to that of the competing methods[4]. To this end, we examine the difference of summary graphs between different subpopulations. In the sequel, we focus on SMG Kendall, SMG Pearson, and

[4]We consider the predictive power of the methods in classification. Because the classification power increases with greater separation between different classes, our experiment measures the predictive power by calculating the scaled Hamming distance between the sparse graphs estimated over two classes of data.

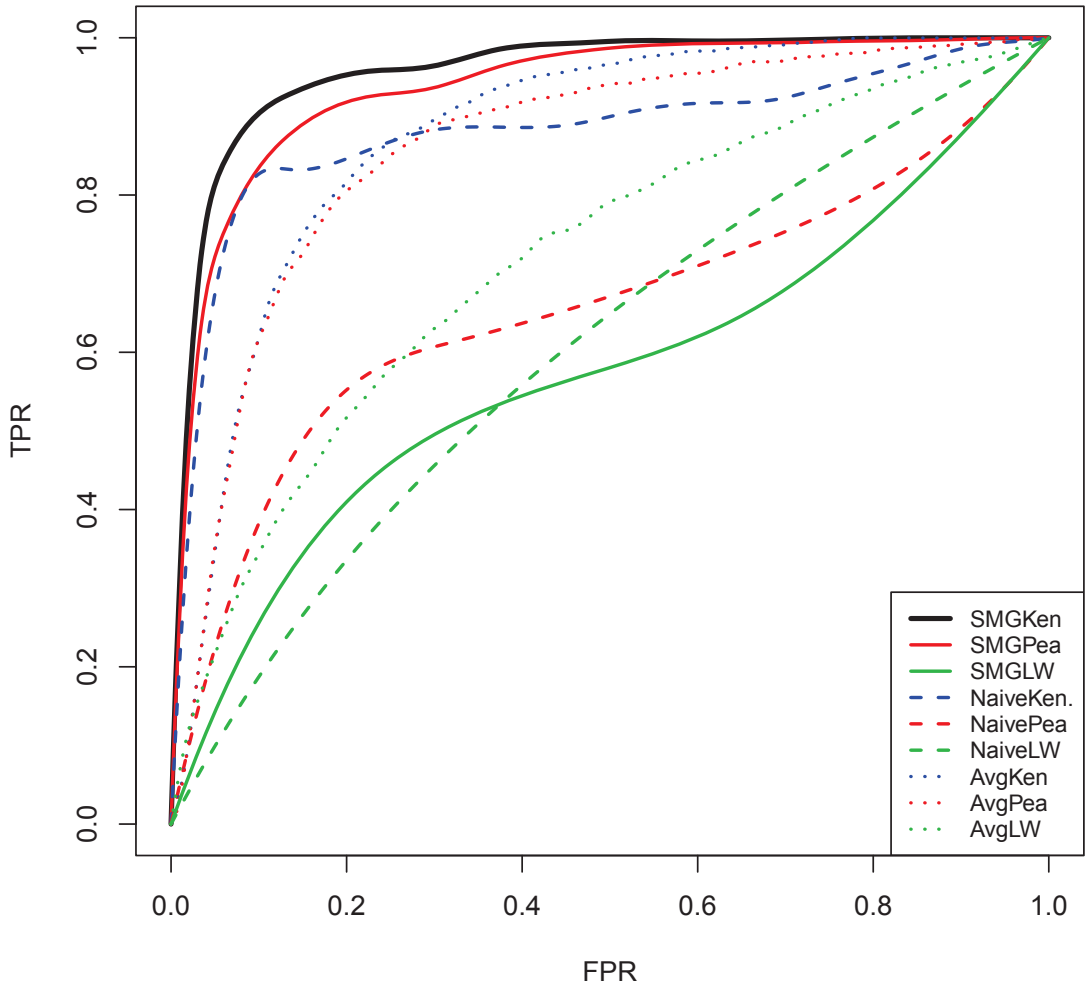


Figure 6.8: ROC curves in estimating the summary graphical models using data based on the data of subject of ID 15002 in the ADHD-200 dataset. Here, $d = 264$ and $T = 10$.

Naive Kendall, which perform the best in simulation[5] (Avg Kendall and Avg Pearson also performed well, but we omit them because they are not robust to outliers).

Several population sparse graph contrasts of interest were investigated and included: ADHD case status (denoted by **Case** and **Control**), gender (denoted by **Female** and **Male**), and age. Given the pediatric population in the ADHD study, this investigates young adults versus children using a cutoff of 12 years. Subjects having ages larger than 12 years are denoted by **Senior** and those less than or equal to 12 years denoted by **Junior**.

Figure 6.9 provides the comparison of the brain connectivity graphs obtained using the three methods on **Case** and **Control** data. We find the **Case** and **Control** graphs show the most edge disagreements when estimated with SMG Kendall. This is consistent to the simulation results, and strongly indicates sparse median graphs coupled with the Kendall's tau estimation procedure works in studying real applications.

A more detailed analysis was subsequently attempted. We applied the three methods on subpopulations to find difference between graphs with different covariate levels. For example, graphs between cases and controls were investigated stratified by gender. Summary statistics for these subpopulations differences are presented in Table 6.1.

In all cases, we again find SMG Kendall estimates the greatest difference between any two classes. In addition, observe while SMG Pearson performs very closely to SMG

[5] All the remaining experiments utilize all patients in the ADHD dataset. For selecting tuning parameters, since we must estimate a graph for each patient, and parameter selection is computationally expensive, we randomly sample 100 subjects from the 739 subjects and apply StARs to estimate the CLIME parameter for each subject. Then, we find the median valued parameter among the selected parameters and use it as the universal parameter for all applications of CLIME in the following experiments. We found that the median parameter from using both Kendall and Pearson is $\lambda = 0.171$.

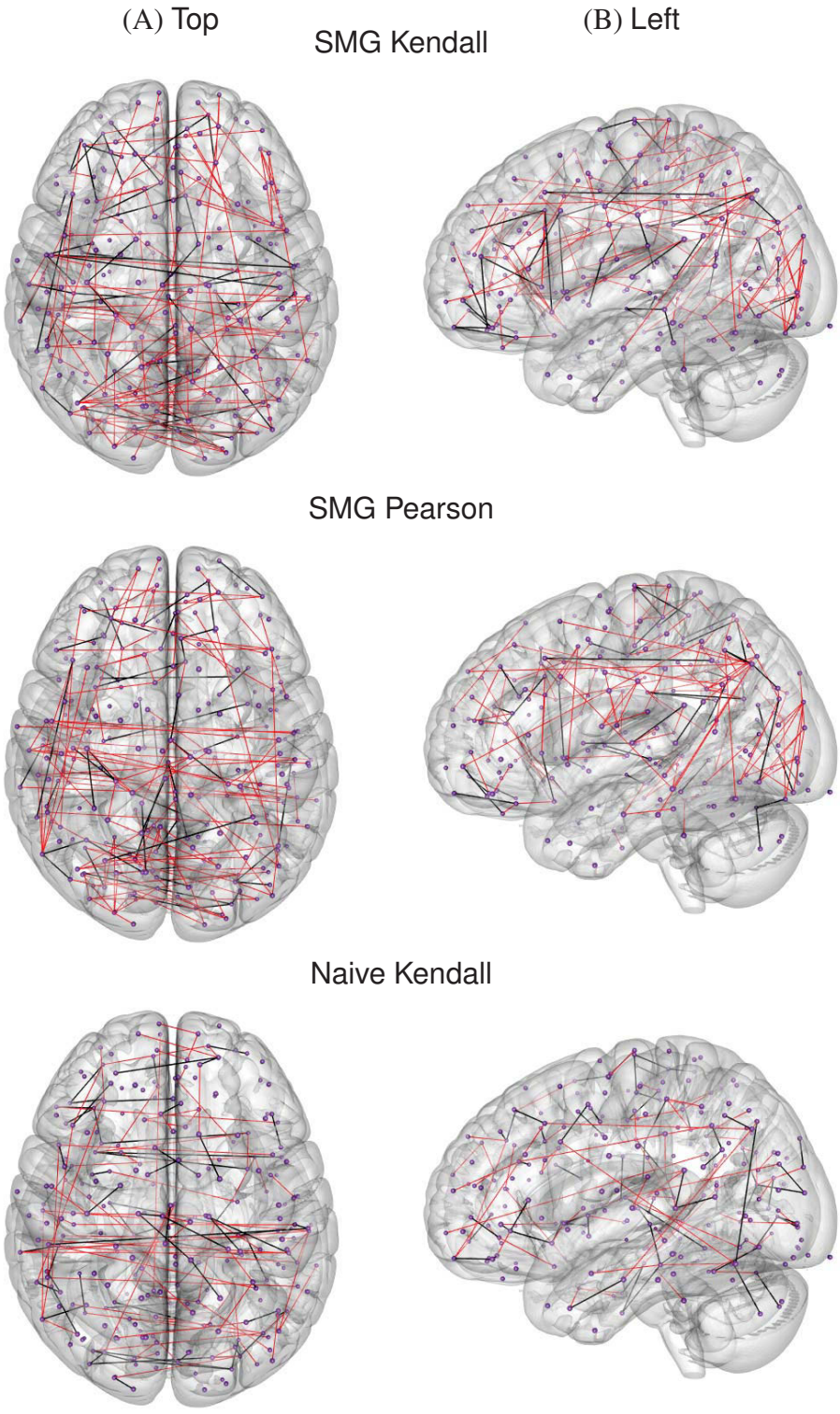


Figure 6.9: The difference between the estimated sparse graphs of the cases and control subjects using SMG Kendall, SMG Pearson, and Naive Kendall. Here, the black color represents the edges only present in the graph for cases but not in controls persons, while the red represents the opposite.

Table 6.1: **Predictive Power.** Predictive power of SMG Kendall and competiting methods. We measure predictive power by the Hamming distance between patients of two classes divided by $\binom{d}{2}$. We use $\lambda = 0.171$ for the CLIME parameter. This table represents values at 10^{-3} scale.

Case and Control Difference			
data	**SMG Ken.**	**SMG Pea.**	**Naive Ken.**
Whole	5.18	5.10	4.35
Male	10.57	10.05	4.90
Female	6.60	5.67	4.90
Junior	6.22	6.31	4.35
Senior	9.02	7.89	5.10
Male and Female Difference			
data	**SMG Ken.**	**SMG Pea.**	**Naive Ken.**
Case	9.07	8.64	5.67
Control	7.81	6.63	4.35
Junior and Senior Difference			
data	**SMG Ken.**	**SMG Pea.**	**Naive Ken.**
Case	9.45	8.93	5.67
Control	9.36	9.25	6.19

Kendall in most cases, there is a larger difference between the two methods in the tests separating or comparing the subjects by gender. This suggests SMG Kendall is more sensitive to the differences between male and female brains, than SMG Pearson. Furthermore, both SMG-based methods show much more predictive power than Naive Kendall. This result demonstrates the predictive advantage of assuming a non-i.i.d population.

6.6.3.3 Stability: CLIME Parameter Perturbations

In this experiment, we compare the stability of the proposed method to that of the competiting methods under parameter perturbations. In particular, we examine the stability by measuring the scaled Hamming distance between a sparse graph estimated with the CLIME parameter $\lambda = 0.171$ and the sparse graph estimated using a perturbed CLIME parameter.

To this end, we conduct the experiment as follows for each of the three methods:

1. Using the CLIME parameter $\lambda = 0.171$, we estimate a population level sparse graph $\widehat{\mathcal{G}}_{\lambda}^{s_\lambda}$. Here we select the s_λ parameter by setting s_λ to be the median number of edges of the graphs estimated for each individual subject. More specifically, recall the algorithm first applies Kendall or Pearson (see Section 6.6.1.1) to each individual graph. Each one of these individual graphs possess some number of edges. We choose s_λ to be the median among that set of numbers.

2. We repeat the procedure but estimate the graph of each individual subject with a perturbed CLIME parameter. In particular, we use $p \times \lambda$ for the values of $p = 0.9$, 0.95, 0.99, 1.01, 1.05, and 1.1.

3. We examine the Hamming distance between each $\widehat{\mathcal{G}}_{(p\times\lambda)}^{s_{(p\times\lambda)}}$ and $\widehat{\mathcal{G}}_{\lambda}^{s_\lambda}$ divided by s_λ. The results are shown in Table 6.2.

We see our proposed method is comparable in stability to SMG Pearson. In addition, observe Naive Kendall tends to display significantly higher instability than the SMG-

Table 6.2: **Stability w.r.t CLIME Parameter.** Stability of SMG Kendall and competiting methods with respect to perturbations to the CLIME parameter. Stability is measured as Hamming distance, divided by s_λ, between the graph estimated with the perturbed parameter and the graph estimated with the unperturbed parameter, λ. Here, s_λ is the number edges in the graph of estimated with the unperturbed parameter. We use $\lambda = 0.171$ for as the CLIME parameter. This table represents values at 10^{-1} scale.

	Variation					
	0.9λ		0.95λ		0.99λ	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
SMG Kendall	3.46	0.137	2.14	0.109	1.68	0.148
SMG Pearson	3.23	0.240	2.05	0.142	1.51	0.209
Naive Kendall	3.83	0.381	2.46	0.279	1.57	0.145
	Variation					
	1.01λ		1.05λ		1.1λ	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
SMG Kendall	1.51	0.128	1.77	0.266	2.34	0.066
SMG Pearson	1.48	0.143	1.80	0.096	2.41	0.097
Naive Kendall	1.49	0.231	2.42	0.182	3.17	0.174

based methods for 1.05λ and 1.1λ. Since CLIME outputs more sparse graphs for larger λ parameters, this supports the claim that sparse median graphs provide a more stable estimator of graphs in sparse settings than models assuming the population data arise from i.i.d settings.

6.6.3.4 Stability: Data Perturbations

In this experiment, we consider the stability of the proposed method when the dataset is perturbed. In particular, we take subsamples of the data from each patient to create a new, subsampled dataset. We repeat this procedure multiple times and measure the instability by examining the differences in the resulting graphs.

More specifically, we apply the following procedure using each of the three methods:

1. We randomly draw $K = 100$ subsamples from the set of subjects at subsampling ratios of p. In other words, each subsampled dataset, $\mathcal{D}_k$, contains data corresponding to $\lfloor p \times T \rfloor$ subjects from the entire ADHD dataset, where $T = 739$. We perform this procedure using subsampling ratios of $p = 0.65$, 0.8, and 0.9.

2. Using the CLIME parameter $\lambda = 0.171$, we estimate a sparse graph $\widehat{\mathcal{G}}_k^s$ from $\mathcal{D}_k$.

3. We measure the instability by averaging the disagreements on the presence of edges in $\widehat{\mathcal{G}}_1^s, \widehat{\mathcal{G}}_2^s, ..., \widehat{\mathcal{G}}_K^s$. We refer to Section 3.2 of Liu et al. (2010) for a detailed description of this measure. The results are summarized in Table 6.3 where larger values correspond to more instability.

We see that our proposed method is comparable in stability to SMG Pearson under data perturbations. In addition, observe Naive Kendall displays significantly more instability than either of the other methods employing the sparse median graph approach. This demonstrates the resistance of the sparse median graph approach to the characteristics of individual subjects when estimating a population-level graph.

Table 6.3: **Stability w.r.t Data.** Stability of SMG Kendall and competiting methods with respect to perturbations to the data via subsampling. Here, we measure the total instability as the mean of the disagreements on the presence each edge (Liu et al., 2010). Here $n = 100$ samples were taken. We use $\lambda = 0.171$ used for the CLIME parameter. This table represents values at 10^{-3} scale.

	Sampling Ratio					
	0.65		0.8		0.9	
	Instability	Std. Err.	Instability	Std. Err.	Instability	Std. Err.
SMG Kendall	2.39	0.152	1.55	0.122	0.941	0.096
SMG Pearson	2.30	0.152	1.54	0.126	1.02	0.102
Naive Kendall	3.00	0.171	2.70	0.162	2.46	0.157

6.7 Discussion

In this chapter, we discuss the concept of the sparse median graphs to estimate a population level graph under nonparanormal assumptions. This new approach combines two new developments in graph estimation literature – namely, (i) employing sparsity constraints in high dimensional settings for identifiability and (ii) increasing graph recovery rates by using nonparanormal assumptions (Liu et al., 2009, 2012a) – with the idea of median graphs

from pattern recognition literature (Bunke and Shearer, 1998; Jiang et al., 2001). The resulting method, which we analyzed both theoretically and empirically, allows us to estimate a graph emphasizing the commonalities within a population and downplays outliers of individuals within the population.

In particular, we theoretically prove the consistency of this method and bound its rate of convergence. Then, in two simulations – one with synthetic and one with real brain imaging data – we demonstrate our proposed method displays higher estimation performance than potential competing methods. In addition, we observe the benefits of the nonparanormal assumptions with Kendall's tau tend to dominate in the synthetic data simulations, but the benefits of the sparse median graph aggregation method tend to dominate in the simulations with real data. One possible explanation is that the "biggest challenge" in estimating the graphs from synthetic data is the data's non-Gaussianity (which is "solved" by utilizing Kendall's tau), while the biggest challenge in estimating the graphs from the real brain imaging data stems from the individual outlier characteristics of patients and scans (which are downplayed by the sparse median graph). However, the consistent optimal performance of the proposed method in both simulations demonstrates its value as an estimator of choice for both highly non-Gaussian data, as well as complex aggregated datasets with large variation in individual characteristics.

We then perform experiments using the ADHD-200 brain imaging dataset to demonstrate the proposed method possesses the highest predictive power for classification tasks among its competitors. Furthermore, stability experiments on the same dataset show the

sparse median graph summarization provides much more stable estimators than the Naive Kendall method assuming homogeneity of the entire dataset.

These results offer compelling evidence that the proposed method possesses the potential to become an unified framework for conducting inference on complex datasets of aggregated data. While the current analysis is primarily illustrative, we have demonstrated its value for applications in arenas of image- and eletrophysiologically-based estimates of function and structural brain connectivity, where interest lies primarily in population characteristics. Therefore, we believe this investigation would justify a more thorough inferential investigation of the median graph properties and network modification with disease in future works.

A key idea in this chapter is still the same as previous: We advocate using semiparametric models (transelliptical graphical model) coupled with nonparametric methods (rank-based Kendall's tau statistic). Via exhaustive numerical studies, we show the proposed method is robust to different types of data contamination, as well as enjoys good predictive powers and stability properties. This adds more merits to our unified framework in tackling complex high dimensional data.

Bibliography

Amini, A. and Wainwright, M. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):2877–2921.

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis (3rd edition)*. Wiley, New York.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, volume 2. Wiley.

Arcones, M. and Gine, E. (1993). Limit theorems for U-processes. *The Annals of Probability*, 21(3):1494–1542.

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics*, 37(6B):3822–3840.

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294.

BIBLIOGRAPHY

Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408.

Bakirov, N., Rizzo, M., and Székely, G. (2006). A multivariate nonparametric test of independence. *Journal of Multivariate Analysis*, 97(8):1742–1756.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9:485–516.

Bao, Z., Lin, L.-C., Pan, G., and Zhou, W. (2013). Spectral statistics of large dimensional Spearman's rank correlation matrix and its application. *arXiv preprint arXiv:1312.5119*.

Bao, Z., Pan, G., and Zhou, W. (2012). Tracy-widom law for the extreme eigenvalues of sample correlation matrices. *Electronic Journal of Probability*, 17(88):1–32.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Berthet, Q. and Rigollet, P. (2012). Optimal detection of sparse principal components in high dimension. *forthcoming in the Annals of Statistics*.

BIBLIOGRAPHY

Berthet, Q. and Rigollet, P. (2013). Computational lower bounds for sparse PCA. *arXiv preprint arXiv:1304.0828*.

Bickel, P. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics*. Springer.

Birke, M. and Dette, H. (2005). A note on testing the covariance matrix for large dimension. *Statistics and Probability Letters*, 74(3):281–289.

Boente, G., Barrerab, M. S., and Tylerc, D. E. (2012). A characterization of elliptical distributions and some optimality properties of principal components for functional data. Technical report, Technical report. Available at http://www. stat. ubc. ca/˜ matias/Property_FPCA. pdf.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Bunea, F. and Xiao, L. (2014). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *Bernoulli (in press)*.

BIBLIOGRAPHY

Bunke, H. and Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern recognition letters*, 19(3):255–259.

Cai, T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39(3):1496–1525.

Cai, T. and Jiang, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.

Cai, T., Liu, W., and Xia, Y. (2014a). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372.

Cai, T., Ma, Z., and Wu, Y. (2014b). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields (in press)*.

BIBLIOGRAPHY

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Cai, T. T. and Zhou, H. H. (2012). Minimax estimation of large covariance matrices under ℓ_1 norm. *Statist. Sinica*, 22:1319–1378.

Chatfield, C. and Collins, A. (1980). *Introduction to Multivariate Analysis*, volume 166. Chapman & Hall.

Chen, S., Zhang, L., and Zhong, P. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Choi, K. and Marden, J. (1998). A multivariate version of Kendall's τ. *Journal of Nonparametric Statistics*, 9(3):261–293.

Čižek, P., Härdle, W., and Weron, R. (2005). *Statistical Tools for Finance and Insurance*. Springer.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314.

Croux, C. and Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, 19(4):497–515.

BIBLIOGRAPHY

Croux, C., Filzmoser, P., and Fritz, H. (2013). Robust sparse principal component analysis. *Technometrics*, 55(2):202–214.

Croux, C. and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618.

Croux, C., Ollila, E., and Oja, H. (2002). Sign and rank covariance matrices: statistical properties and application to principal components analysis. *Statistics in Industry and Technology*, pages 257–269.

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I., and Lanckriet, G. R. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49(3):434–448.

Dattorro, J. (2005). *Convex Optimization and Euclidean Distance Geometry*. Meboo Publishing.

Davies, P. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):1269–1292.

BIBLIOGRAPHY

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Di Martino, A., Yan, C., Li, Q., Denio, E., Castellanos, F., Alaerts, K., Anderson, J., Assaf, M., Bookheimer, S., Dapretto, M., et al. (2013). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *to appear Molecular psychiatry*.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(2):642–669.

Eloyan, A., Muschelli, J., Nebel, M., Liu, H., Han, F., Zhao, T., Barber, A., Joel, S., Pekar, J., Mostofsky, S., et al. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6:61.

Embrechts, P., Lindskog, F., and McNeil, A. (2003). Modelling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance*, 8(1):329–384.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605–2637.

BIBLIOGRAPHY

Fang, H., Fang, K., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1):1–16.

Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric Multivariate and Related Distributions*. Chapman&Hall.

Fingelkurts, A., Kähkönen, S., et al. (2005). Functional connectivity in the brain–is it an elusive concept? *Neuroscience and biobehavioral reviews*, 28(8):827–836.

Fisher, T. (2012). On testing for an identity covariance matrix when the dimensionality equals or exceeds the sample size. *Journal of Statistical Planning and Inference*, 142(1):312–326.

Fisher, T., Sun, X., and Gallagher, C. (2010). A new test for sphericity of the covariance matrix for high dimensional data. *Journal of Multivariate Analysis*, 101(10):2554–2570.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friston, K. (2011). Functional and effective connectivity: a review. *Brain Connectivity*, 1(1):13–36.

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference*, volume 168. CRC press.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124.

BIBLIOGRAPHY

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, 39(2):325–346.

Hájek, J., Sidak, Z., and Sen, P. K. (1999). *Theory of Rank Tests (2nd edition)*. Academic Press, New York.

Hallin, M., Oja, H., and Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. II. Optimal R-estimation of shape. *The Annals of Statistics*, 34(6):2757–2789.

Hallin, M. and Paindaveine, D. (2002a). Optimal procedures based on interdirections and pseudo-Mahalanobis ranks for testing multivariate elliptic white noise against ARMA dependence. *Bernoulli*, 8(6):787–815.

Hallin, M. and Paindaveine, D. (2002b). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30(4):1103–1133.

Hallin, M. and Paindaveine, D. (2004). Rank-based optimal tests of the adequacy of an elliptic VARMA model. *The Annals of Statistics*, 32(6):2642–2678.

Hallin, M. and Paindaveine, D. (2005). Affine-invariant aligned rank tests for the mul-

tivariate general linear model with VARMA errors. *Journal of Multivariate Analysis*, 93(1):122–163.

Hallin, M. and Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *The Annals of Statistics*, 34(6):2707–2756.

Hallin, M., Paindaveine, D., and Verdebout, T. (2010). Optimal rank-based testing for principal components. *The Annals of Statistics*, 38(6):3245–3299.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Han, F. and Liu, H. (2013a). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv preprint arXiv:1305.6916*.

Han, F. and Liu, H. (2013b). Principal component analysis on non-Gaussian dependent data. *Proceedings of the Thirtieth International Conference on Machine Learning*.

Han, F. and Liu, H. (2014a). High dimensional semiparametric scale-invariant principal component analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (in press)*.

Han, F. and Liu, H. (2014b). Scale-invariant sparse PCA on high dimensional meta-elliptical data. *Journal of the American Statistical Association (In press)*.

BIBLIOGRAPHY

Han, F., Zhao, T., and Liu, H. (2013). CODA: High dimensional copula discriminant analysis. *Journal of Machine Learning Research*, 14:629–671.

Hoeffding, W. (1948a). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Hoeffding, W. (1948b). A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4):546–557.

Hogg, R. V. and Craig, A. (2012). *Introduction to Mathematical Statistics (7th Edition)*. Pearson.

Horwitz, B. et al. (2003). The elusive concept of brain connectivity. *Neuroimage*, 19(2):466–470.

Hotelling, H. and Pabst, M. (1936). Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics*, 7(1):29–43.

Hsieh, C.-J., Sustik, M. A., Ravikumar, P., and Dhillon, I. S. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics, 2nd edition*. Wiley.

Hubert, M., Rousseeuw, P. J., and Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1):101–111.

BIBLIOGRAPHY

Huffer, F. W. and Park, C. (2007). A test for elliptical symmetry. *Journal of Multivariate Analysis*, 98(2):256–281.

Hult, H. and Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied probability*, 34:587–608.

Inglot, T. (2012). Asymptotic behaviour of linear rank statistics for the two-sample problem. *Probability and Mathematical Statististics*, 32(1):93–116.

Jackson, D. and Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics*, 15(2):129–139.

Jiang, D., Bai, Z., and Zheng, S. (2013). Testing the independence of sets of large-dimensional variables. *Science China Mathematics*, 56(1):135–147.

Jiang, D., Jiang, T., and Yang, F. (2012). Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference*, 142(8):2241–2256.

Jiang, T. (2004). The asymptotic distributions of the largest entries of sample correlation matrices. *The Annals of Applied Probability*, 14(2):865–880.

Jiang, T. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *The Annals of Statistics*, 41(4):2029–2074.

Jiang, X., Munger, A., and Bunke, H. (2001). An median graphs: properties, algorithms,

and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(10):1144–1151.

Jin, J., Zhang, C., and Zhang, Q. (2012). Optimality of graphlet screening in high dimensional variable selection. *arXiv preprint arXiv:1204.6452*.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Johnstone, I. and Lu, A. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531–547.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553.

Jung, S. and Marron, J. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130.

Kallenberg, W. (1982). Cramér type large deviations for simple linear rank statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 60(3):403–409.

BIBLIOGRAPHY

Kang, J. (2013). ABIDE data preprocessing. *personal communication*.

Ke, T., Jin, J., and Fan, J. (2012). Covariance assisted screening and estimation. *arXiv preprint arXiv:1205.4645*.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics. Vols. II*. Griffin, London.

Kendall, M. G. (1948). *Rank Correlation Methods.* Griffin.

Kruskal, W. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):42–54.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, 30(4):1081–1102.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621.

Li, D., Liu, W., and Rosalsky, A. (2010). Necessary and sufficient conditions for the

asymptotic distribution of the largest entry of a sample correlation matrix. *Probability Theory and Related Fields*, 148(1-2):5–35.

Li, D. and Rosalsky, A. (2006). Some strong limit theorems for the largest entries of sample correlation matrices. *The Annals of Applied Probability*, 16(1):423–447.

Li, J. and Chen, S. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940.

Li, L. and Toh, K.-C. (2010). An inexact interior point method for ℓ_1-reguarlized sparse covariance selection. *Mathematical Programming Computation*, 2(3):291–315.

Li, R., Fang, K., and Zhu, L. (1997). Some Q-Q probability plots to test spherical and elliptical symmetry. *Journal of Computational and Graphical Statistics*, 6(4):435–450.

Lindskog, F., McNeil, A., and Schmock, U. (2003). *Kendall's tau for elliptical distributions*. Springer.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012a). High dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*, 40(4):2293–2326.

Liu, H., Han, F., and Zhang, C.-H. (2012b). Transelliptical graphical modeling under a hierarchical latent variable framework. *Technical Report*.

Liu, H., Han, F., and Zhang, C.-H. (2012c). Transelliptical graphical models. In *Proceed-*

ings of the Twenty-fifthy Annual Conference on Neural Information Processing Systems, pages 809–817.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, volume 23.

Liu, I.-M. and Agresti, A. (1996). Mantel-haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, 52(4):1223–1234.

Liu, L., Hawkins, D. M., Ghosh, S., and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23):13167–13172.

Liu, W., Lin, Z., and Shao, Q. (2008). The asymptotic distribution and Berry-Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *The Annals of Applied Probability*, 18(6):2337–2366.

Lounici, K. (2013a). High-dimensional covariance matrix estimation with missing observations. *Bernoulli (In press)*.

Lounici, K. (2013b). Sparse principal component analysis with missing observations. In *High Dimensional Probability VI*, pages 327–356. Springer.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *forthcoming in the Annals of Statistics*.

Ma, Z. and Wu, Y. (2013). Computational barriers in minimax submatrix detection. *arXiv preprint arXiv:1309.5914*.

Mackey, L. (2008). Deflation methods for sparse PCA. In *Advances in Neural Information Processing Systems*, volume 21, pages 1017–1024.

Mackey, L. (2009). Deflation methods for sparse PCA. *Advances in Neural Information Processing Systems*, 21:1017–1024.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

Mao, G. (2014). A new test of independence for high-dimensional data. *Statistics and Probability Letters*, 93:14–18.

Marden, J. (1999). Some robust estimates of principal components. *Statistics & Probability Letters*, 43(4):349–359.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317.

BIBLIOGRAPHY

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Milham, M. P., Fair, D., Mennes, M., and Mostofsky, S. H. (2012). The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62.

Morrison, D. (2004). *Multivariate Statistical Methods (4th edition)*. Cengage Learning, Stamford, CT.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5(2):201–213.

Nagao, H. (1973). On some test criteria for covariance matrix. *The Annals of Statistics*, 1(4):700–709.

Nelsen, R. (1999). *An Introduction to Copulas*. Springer, New York.

Neter, J., Kutner, M., Wasserman, W., and Nachtsheim, C. (1996). *Applied Linear Statistical Models*, volume 4. Irwin Chicago.

Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*, volume 199. Springer.

BIBLIOGRAPHY

Oja, H. and Paindaveine, D. (2005). Optimal signed-rank tests based on hyperplanes. *Journal of Statistical Planning and Inference*, 135(2):300–323.

Oja, H. and Randles, R. H. (2004). Multivariate nonparametric tests. *Statistical Science*, 19(4):598–605.

Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35(2):175–189.

Overton, M. L. and Womersley, R. S. (1992). On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45.

Paul, D. and Johnstone, I. (2012). Augmented sparse principal component analysis for high dimensional data. *Arxiv preprint arXiv:1202.1242*.

Péché, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probability Theory and Related Fields*, 143(3-4):481–516.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746.

Pillai, N. S. and Yin, J. (2012). Edge universality of correlation matrices. *The Annals of Statistics*, 40(3):1737–1763.

Posekany, A., Felsenstein, K., and Sykacek, P. (2011). Biological assessment of robust noise models in microarray data analysis. *Bioinformatics*, 27(6):807–814.

BIBLIOGRAPHY

Power, J., Cohen, A., Nelson, S., Wig, G., Barnes, K., Church, J., Vogel, A., Laumann, T., Miezin, F., Schlaggar, B., et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley.

Rachev, S. T. (2003). *Handbook of Heavy Tailed Distributions in Finance*, volume 1. Elsevier.

Ramsay, J., Hanson, S., Hanson, C., et al. (2009). Six problems for causal inference from fmri. *NeuroImage*, 49(2):1545–58.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2009). Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems*, volume 22.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., and Maechler, M. (2009). Robustbase: basic robust statistics. *R package, URL http://CRAN. R-project. org/package= robustbase*.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

BIBLIOGRAPHY

Rowe, B. L. Y. (2014). *tawny: Provides various portfolio optimization strategies including random matrix theory and shrinkage estimators*. R package version 2.1.2.

Roy, S. N. (1957). *Some Aspects of Multivariate Analysis*. Wiley, New York.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069.

Ruttimann, U. E., Unser, M., Rawlings, R. R., Rio, D., Ramsey, N. F., Mattay, V. S., Hommer, D. W., Frank, J. A., and Weinberger, D. R. (1998). Statistical analysis of functional MRI data in the wavelet domain. *IEEE Transactions on Medical Imaging*, 17(2):142–154.

Sakhanenko, L. (2008). Testing for ellipsoidal symmetry: A comparison study. *Computational Statistics & Data Analysis*, 53(2):565–581.

Scheinberg, K., Ma, S., and Glodfarb, D. (2010). Sparse inverse covariance selection via alternating linearization methods. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23.

Schott, J. (2005). Testing for complete independence in high dimensions. *Biometrika*, 92(4):951–956.

Seoh, M., Ralescu, S., and Puri, M. (1985). Cramér type large deviations for generalized rank statistics. *The Annals of Probability*, 13(1):115–125.

BIBLIOGRAPHY

Serfling, R. (2002). *Approximation Theorems of Mathematical Statistics*, volume 162. Wiley, New York.

Shao, Q. and Zhou, W. (2014). Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *The Annals of Probability*, 42(2):623–648.

Shen, H. and Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034.

Sirkiä, S., Taskinen, S., and Oja, H. (2007). Symmetrised M-estimators of multivariate scatter. *Journal of Multivariate Analysis*, 98(8):1611–1629.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Srivastava, M. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society*, 35(2):251–272.

Srivastava, M. (2006). Some tests criteria for the covariance matrix with fewer observations than the dimension. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 10:77–93.

Srivastava, M. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, 101(6):1319–1329.

BIBLIOGRAPHY

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402.

Székely, G. and Rizzo, M. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Taskinen, S., Kankainen, A., and Oja, H. (2003). Sign test of independence between two random vectors. *Statistics and Probability Letters*, 62(1):9–21.

Taskinen, S., Koch, I., and Oja, H. (2012). Robustifying principal component analysis with spatial sign vectors. *Statistics and Probability Letters*, 82(4):765–774.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Tyler, D. E. (1982). Radial estimates and the test for sphericity. *Biometrika*, 69(2):429–436.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 15(1):234–251.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in

SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289.

van de Geer, S. and Lederer, J. (2013). The Bernstein-Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1–2):225–250.

Vandemaele, M. and Veraverbeke, N. (1982). Cramér type large deviations for linear combinations of order statistics. *The Annals of Probability*, 10(2):423–434.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge University Press.

Visuri, S., Koivunen, V., and Oja, H. (2000). Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91(2):557–575.

Vu, V. and Lei, J. (2012). Minimax rates of estimation for sparse PCA in high dimensions. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 15(1278–1286).

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems*, volume 26, pages 2670–2678.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947.

Waerden, B. L. (1957). *Mathematische Statistik*. Springer-Verlag, Berlin, Germany.

BIBLIOGRAPHY

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.

Wang, Z., Han, F., and Liu, H. (2013). Sparse principal component analysis for high dimensional multivariate time series. *Journal of Machine Learning Research (AISTATS Track)*.

Wedin, P.-A. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Wegkamp, M. and Zhao, Y. (2013). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. arxiv preprint. *forthcoming in Bernoulli*.

Witten, D., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Woodworth, G. (1970). Large deviations and Bahadur efficiency of linear rank statistics. *The Annals of Mathematical Statistics*, 41(1):251–283.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286.

BIBLIOGRAPHY

Yuan, X. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14:899–925.

Zaïtsev, A. Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of S.N. Bernstein's inequality conditions. *Probability Theory and Related Fields*, 74(4):535–566.

Zhang, R., Peng, L., and Wang, R. (2013). Tests for covariance matrix with fixed or divergent dimension. *The Annals of Statistics*, 41(4):2075–2096.

Zhang, Y. and El Ghaoui, L. (2011). Large-scale sparse principal component analysis with application to text data. *Advances in Neural Information Processing Systems*, 24.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 13:1059–1062.

Zhou, W. (2007). Asymptotic distribution of the largest off-diagonal entry of correlation matrices. *Transactions of the American Mathematical Society*, 359(11):5345–5363.

Zou, C., Peng, L., Feng, L., and Wang, Z. (2014). Multivariate sign-based high-dimensional tests for sphericity. *Biometrika*, 101(1):229–236.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

CURRICULUM VITAE

FANG HAN

fhan@jhu.edu

615 N. Wolfe St. E3039

Baltimore, MD 21205

http://www.biostat.jhsph.edu/∼fhan

EDUCATION

2010 - 2015 **Johns Hopkins Bloomberg School of Public Health**, Baltimore, MD

Ph.D. in Biostatistics

Advisor: Han Liu and Brian Caffo

2008 - 2010 **University of Minnesota**, Minneapolis, MN

M.S. in Biostatistics

2004 - 2008 **Peking University**, Beijing, China

B.S. in Probability and Statistics

AWARDS

2013-2015 Google Ph.D. Fellowship in Statistics.

2015 Margaret Merrell Award.

2014 ASA Biometrics Section David P. Byar Young Investigator Travel Award.

2013 ICSA Distinguished Student Paper Award.

CURRICULUM VITAE

2013 AISTATS Notable Paper Award.

2013 ENAR Distinguished Student Paper Award.

TEACHING EXPERIENCE

2014 Instructor, Theory for Big Data. SLAM Group (led by Mei-Cheng Wang and Chiung-Yu Huang), Johns Hopkins University.

2013 Guest lecturer, Advanced Statistical Theory (taught by Daniel Scharfstein). Johns Hopkins University.

2014 TA, Essentials of Probability and Statistical Inference (taught by Charles Rohde and Mei-Cheng Wang). Johns Hopkins University.

2014 TA, Statistical Methods in Public Health (taught by James Tonascia). Johns Hopkins University.

2013 TA, Advanced Statistical Theory (taught by Daniel Scharfstein). Johns Hopkins University.

2013 TA, Probability Theory (taught by James Fill). Johns Hopkins University.

2012 TA, Statistics for Laboratory Scientists (taught by Ingo Ruczinski). Johns Hopkins University.

2011 TA, Public Health Biostatistics (for undergraduate students, taught by Scott Zeger). Johns Hopkins University.

CURRICULUM VITAE

PUBLICATIONS

Peer-Reviewed Journal Publications

Han, F, Lu, H., and Liu, H. (2015+). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, to appear.

Han, F. and Liu, H. (2015+). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli*, to appear.

Qiu, H., Han, F., Liu, H., and Caffo, B. (2015+). Joint estimation of multiple graphical models from high dimensional dependent data. *Journal of the Royal Statistical Society: Series B*, to appear.

Han, F. and Liu, H. (2014). High dimensional semiparametric scale-invariant principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):2016–2032.

Fan, J., Han, F., and Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, 1(3):293–314.

Han, F. and Liu, H. (2014). Scale-invariant sparse PCA on high dimensional meta-elliptical data. *Journal of the American Statistical Association*, 109(505):275–287.

Han, F., Zhao, T., and Liu, H. (2013). CODA: High dimensional copula discriminant analysis. *Journal of Machine Learning Research*, 14:629–671.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional

semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Han, F. and Pan, W. (2011). A composite likelihood approach to latent multivariate Gaussian modeling of SNP data with application to genetic association testing. *Biometrics*, 68(1): 307–315.

Han, F. and Pan, W. (2010). Powerful multi-marker association tests: Unifying genomic distance-based regression and logistic regression. *Genetic Epidemiology*, 34(7): 680–688.

Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1): 42–54.

Pan, W., Han, F., and Shen, X. (2010). Test selection with application to detecting disease association with multiple SNPs. *Human Heredity*, 69(2): 120–130.

Han, F., Wu, J., Xu, J., and Deng, M. (2009). Searching for differentially expressed genes by PLS-VIP method. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 45(1): 1-5, 2009

Peer-Reviewed Journal Publications (Collaborative Work)

Kano, S., Colantuoni, C., Han, F., Zhou, Z., Yuan, Q., Wilson, A., Takayanagi, Y., Lee, Y., Rapoport, J., Eaton, W., Cascella, N., Ji, H., Goldman, D., and Sawa, A. (2013). Genome-wide profiling of multiple histone methylations in olfactory cells: Further implications for cellular susceptibility to oxidative stress in schizophrenia. *Nature: Molecular Psychiatry*, 18(7):740–742.

CURRICULUM VITAE

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., Barber, A. D., Joel, S., Pekar, J. J., Mostofsky, S. H., and Caffo, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6(61):1–9.

Peer-Reviewed Conference Publications

Qiu, H., Xu, S., Han, F., Liu, H., and Caffo, B. (2015). Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *Proceedings of the 32th International Conference on Machine Learning (ICML)*.

Yang, J., Han, F., Irizarry, R., and Liu, H. (2014). Context aware group nearest shrunken centroids in large-scale genomic studies. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Han, F. and Liu, H. (2013). Robust sparse principal component regression. In *Proceedings of the 26th Neural Information Processing Systems Conference (NIPS)*.

Han, F. and Liu, H. (2013). Transition matrix estimation in high dimensional vector autoregressive models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

Han, F. and Liu, H. (2013). Principal component analysis on non-Gaussian dependent data. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

Wang, Z., Han, F., and Liu, H. (2013). Sparse principal component analysis for high dimensional multivariate time series. In *Proceedings of the 16th International Conference on*

Artificial Intelligence and Statistics (AISTATS).

Han, F. and Liu, H. (2012). Transelliptical component analysis. In *Proceedings of the 25th Neural Information Processing Systems Conference (NIPS)*.

Han, F. and Liu, H. (2012). Semiparametric principal component analysis. In *Proceedings of the 25th Neural Information Processing Systems Conference (NIPS)*.

Liu, H., Han, F., and C-H, Zhang. (2012). Transelliptical graphical models. In *Proceedings of the 25th Neural Information Processing Systems Conference (NIPS)*.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). The Nonparanormal SKEPTIC. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

TALKS

2015 Department of Biostatistics, University of California, Berkeley

2015 Department of Biostatistics, Johns Hopkins University

2015 Department of Statistics, University of California, Los Angeles

2015 Department of Biostatistics, Harvard University

2015 Department of Statistics, University of Washington

2015 Department of Biostatistics, Columbia University

2015 Department of Statistics, Cornell University

2015 Department of Biostatistics, University of North Carolina at Chapel Hill

CURRICULUM VITAE

2015 Department of Statistics, North Carolina State University

2015 Department of Statistics, University of Illinois at Urbana-Champaign

2014 Joint Statistical Meetings, Boston, Maryland.

2014 ICSA/KISS Joint Conference, Portland, Oregon.

2014 New England Statistics Symposium, Boston, Massachusetts.

2013 International Conference on Machine Learning, Atlanta, Georgia.

2013 ICSA/ISBS Joint Conference, Bethesda, Maryland.

2013 International Conference on Artificial Intelligence and Statistics, Scottsdale, Arizona.

2013 ENAR Spring Meeting, Orlando, Florida.

2012 Neural Information Processing Systems, Lake Tahoe, Nevada.

2012 International Conference on Machine Learning, Edinburgh, Scotland.

2010 ENAR Spring Meeting, New Orlean, Louisiana.

PROFESSIONAL SERVICE

Review: Annals of Statistics (AOS), Journal of the American Statistical Association (JASA), Journal of the Royal Statistical Society: Series B (JRSSB), Biometrika, Biometrics, Electronic Journal of Statistics (EJS), Computational Statistics and Data Analysis (CSDA), Journal of Statistical Software (JSS), PLOS ONE, Mathematical Science China, Neural Information Processing Systems (NIPS).