Published in final edited form as:
J Am Stat Assoc. 2014 January 1; 109(505): 275–287. doi:10.1080/01621459.2013.844699.

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

Fang Han[*] and **Han Liu**[†]

[*]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA; fhan@jhsph.edu

[†]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; hanliu@princeton.edu.

Abstract

We propose a semiparametric method for conducting scale-invariant sparse principal component analysis (PCA) on high dimensional non-Gaussian data. Compared with sparse PCA, our method has weaker modeling assumption and is more robust to possible data contamination. Theoretically, the proposed method achieves a parametric rate of convergence in estimating the parameter of interests under a flexible semiparametric distribution family; Computationally, the proposed method exploits a rank-based procedure and is as efficient as sparse PCA; Empirically, our method outperforms most competing methods on both synthetic and real-world datasets.

Keywords

High dimensional statistics; Principal component analysis; Elliptical distribution; Robust statistics

1 Introduction

Principal component analysis (PCA) is a powerful tool for dimension reduction and feature selection. 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 Σ. PCA aims at estimating the leading eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of Σ.

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 $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_m$ 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 that the angle between $\hat{\boldsymbol{u}}_1$ 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 eigen-vector $\boldsymbol{u}_1 := (u_{11}, \ldots, u_{1d})^T$, 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. (2004), 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: (i) It is not 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); (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 the 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). The theoretical properties of these elliptical distribution based PCA estimators have been established under the classical asymptotic framework (i.e., 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).

In this article 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 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 0 and scale parameter Σ^0, whose diagonal values are all 1. We call Σ^0 the *latent generalized correlation matrix*. By treating $\{f_j\}_{j=1}^d$ as nuisance parameters, our method estimates the leading eigenvector θ_1 of Σ^0 by exploiting a rank-based estimating procedure and can be viewed as a scale-invariant PCA conducted on $f(X)$. Theoretically we show that 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 that the proposed method outperforms the classical sparse PCA and two robust alternatives on both synthetic and real-world datasets.

The rest of this paper is organized as follows. In the next section, we review the elliptical distribution family and introduce the meta-elliptical distribution. In Section 3, we present the statistical model, introduce the rank-based estimators, and provide computational algorithm for parameter estimation. In Section 4, we provide theoretical analysis. In Section 5, we provide empirical studies on both synthetic and real-world datasets. More discussion and comparison with related methods are put in the last section.

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 $\boldsymbol{v}=(v_1,\ldots,v_d)^T\in\mathbb{R}^d$ be a *d*-dimensional matrix and a *d*-dimensional vector. We denote v_I to be the subvector of *v* whose entries are indexed by a set *I*. We also denote $\mathbf{M}_{I,J}$ to be the submatrix of 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 M with rows in *I*, and the submatrix of M with columns in *J*. Let supp(*v*) := $\{j : v_j \neq = 0\}$. For $0 < q < \infty$, we define the ℓ_0, ℓ_q and ℓ_∞ vector norms as $\|\boldsymbol{v}\|_0$:= card(supp(*v*)), $||\boldsymbol{v}||_q := \left(\Sigma_{i=1}^d |v_i|^q\right)^{1/q}$ and $||\boldsymbol{v}||_\infty := max_{1\le i\le d}|v_i|$. 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 **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 **M**. Let $\|\mathbf{M}\|_2$ be the spectral norm of **M**. We define $(\mathbf{M}) := \left(\mathbf{M}_{*1}^T,\ldots,\mathbf{M}_{*d}^T\right)^T$ and $\mathbb{S}^{d-1} := \left\{\boldsymbol{v}\in\mathbb{R}^d{:}||\boldsymbol{v}||_2{=}1\right\}$ 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 $\boldsymbol{b}$, **A** and **B** by $\langle$a, b$\rangle$:= $\boldsymbol{a}^T\mathbf{b}$ and $\langle\boldsymbol{A},\boldsymbol{B}\rangle$:= Tr($\mathbf{A}^T\mathbf{B}$). For any matrix $\mathbf{M}\in\mathbb{R}^{d\times d}$, we denote diag(**M**) to be the diagonal matrix with the same diagonal entries as **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 **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.

2.1 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 distribution are all continuous. *X* possesses density if it is absolutely continuous with respect to the Lebesgue measure.

Definition 2.1 (Elliptical distribution). A random vector $\mathbf{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_\mathrm{d})^T$ follows an elliptical distribution if and only if **Z** has a stochastic representation: $\boldsymbol{Z}\overset{d}{=}\mu{+}\xi\mathbf{A}\boldsymbol{U}$. *Here* $\mu\in\mathbb{R}^d$, *q* := rank(**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* $\Sigma := \mathbf{A}\mathbf{A}^T$, we denote $\mathbf{Z} \sim EC_d(\boldsymbol{\mu}, \Sigma, \xi)$. *We call Σ the scatter matrix.*

In Definition 2.1, there can be multiple **A**'s corresponding to the same Σ, 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 Σ instead of **A**.

The model family in Definition 2.1 is not identifiable. For example, Σ 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 condition that

$\max_{1 \le i \le d} \Sigma_{ii} = 1$. We define $\Sigma^0 := \mathrm{diag}(\Sigma)^{-1/2} \cdot \Sigma \cdot \mathrm{diag}(\Sigma)^{-1/2}$ to be the *generalized correlation matrix*.

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 1 illustrates the nonGaussianity issue of the stock log-return data. Here we conduct the three marginal normality tests as in Table 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 test. Even with Bonferroni correction there are still over half stocks that fail to pass any normality tests. Figure 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 separately. We see that the log-return values are skewed to the left.

Though the elliptical distribution family has been widely used to model heavy-tail data (Oja, 2010), it assumes that 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:

$$\begin{aligned}\mathscr{R}_d^+ &= \left\{\mathbf{\Sigma} \in \mathbb{R}^{\mathbf{d}\times\mathbf{d}}:\mathbf{\Sigma}^{\mathbf{T}}=\mathbf{\Sigma}, \mathbf{diag}\left(\mathbf{\Sigma}\right)=\mathbf{I_d}, \mathbf{\Sigma} \succ \mathbf{0}\right\},\\ \mathscr{R}_d &= \left\{\mathbf{\Sigma} \in \mathbb{R}^{\mathbf{d}\times\mathbf{d}}:\mathbf{\Sigma}^{\mathbf{T}}=\mathbf{\Sigma}, \mathbf{diag}\left(\mathbf{\Sigma}\right)=\mathbf{I_d}, \mathbf{\Sigma}\succeq\mathbf{0}\right\}.\end{aligned}$$

The meta-elliptical distribution family is defined as follows:

Definition 2.2 (Meta-elliptical distribution). *A continuous random vector* $\boldsymbol{X} = (X_1, \dots, X_d)^T$ follows a meta-elliptical distribution, denoted by $\mathbf{X} \sim \mathrm{ME}_\mathrm{d}(\Sigma^0, \xi; f_1, \dots, f_d)$, if there exist univariate strictly increasing functions $\mathrm{f}_1, \dots, f_d$ *such that*

$$\left(f_1\left(X_1\right), \dots, f_d\left(X_d\right)\right)^T \sim EC_d\left(0, \mathbf{\Sigma}^0, \xi\right), \quad \text{where} \quad \mathbf{\Sigma}^0 \in \mathscr{R}_d. \quad (2.1)$$

Here, Σ^0 is called the latent generalized correlation matrix. When

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

we say that $\mathbf{X}$ follows a nonparanormal distribution, denoted by $\mathbf{X} \sim \mathrm{NPN}_\mathrm{d}(\Sigma^0; f_1, \dots, f_d)$.

The meta-elliptical is a strict extension to the nonparanormal defined in Liu et al. (2012). They both assume that 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 that Definition 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 Σ^0 is strictly enlarged from $\mathscr{R}_d^+$ to $\mathscr{R}_d$; (iii) X does not necessarily possess density. Moreover, even if these two definitions are the same confined in the distribution set with density existing, we define the meta-elliptical in fundamentally different ways by characterizing the transformation functions instead of characterizing the density functions. By exploiting this new definition, we find that several results provided in the later sections can be easier to understand.

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 visualizes the density functions of two meta-elliptical distributions.

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 $\boldsymbol{X} \sim ME_d(\Sigma^0, \xi; f_1, \ldots, f_d)$, the proposed method aims at estimating the leading eigenvector of Σ^0. Since the diagonal entries of Σ^0 are all 1, the proposed method is scale-invariant. From Definition 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.

3.1 Statistical Model

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

Definition 3.1. *We consider the following model, denoted by* $\mathscr{M}_d\left(\boldsymbol{\Sigma}^0, \xi, f; \theta_1, s\right)$, *which is defined to be the set of distributions:*

$$\mathscr{M}_d\left(\boldsymbol{\Sigma}^0, \xi, f; \theta_1, s\right) := \left\{\boldsymbol{X} : \boldsymbol{X} \sim ME_d\left(\boldsymbol{\Sigma}^0, \xi; f_1, \ldots, f_d\right) \quad such \quad that \quad \theta_1, the \quad leading \quad eigenvector \quad of \quad \boldsymbol{\Sigma}^0, \quad satisfies \quad (\right.$$

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.

3.2 Method

We now provide the proposed method that exploits the model (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.

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 $\widetilde{\boldsymbol{X}} := \left(\widetilde{X}_1, \ldots, \widetilde{X}_d\right)^T$ be an independent copy of $\boldsymbol{X}$. The Kendall's tau correlation coefficient between X_j and X_k is defined as

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

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

Theorem 3.2. *Given* $\boldsymbol{X} \sim ME_d\ (\Sigma^0, \xi; f_1, \ldots, f_d)$ *meta-elliptically distributed, we have*

$$\boldsymbol{\Sigma}^0_{jk} = sin\left(\frac{\pi}{2}\tau\left(X_j, X_k\right)\right). \quad (3.2)$$

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 $\boldsymbol{x} := (x_{i1}, \ldots, x_{id})^T$ be n data points of $\boldsymbol{X}$. The sample version Kendall's tau statistic is defined as:

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

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

Definition 3.3 (Kendall's tau correlation matrix). *We define the Kendall's tau correlation matrix* $\hat{\mathbf{R}} = \left[\hat{\mathbf{R}}_{jk}\right]$ *to be a d by d matrix with element entry to be*

$$\hat{\mathbf{R}}_{jk} = sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right). \quad (3.3)$$

3.2.2 Rank-based Estimators—Given the model $\mathscr{M}_d\left(\boldsymbol{\Sigma}^0, \xi, f; \theta_1, s\right)$, Theorem 3.2 provides a natural way to estimate θ_1. In particular, we solve the following optimization problem:

$$\hat{\theta}^*_{1,k} := arg \underset{v \in \mathbb{R}^d}{max} \boldsymbol{v}^T \hat{\mathbf{R}} \boldsymbol{v}, \quad \text{subject to} \quad \boldsymbol{v} \in \mathbb{S}^{d-1} \cap \mathbb{B}_0(k), \tag{3.4}$$

where $\mathbb{B}_0(k) := \left\{ \boldsymbol{v} \in \mathbb{R}^d : ||\boldsymbol{v}||_0 \le k \right\}$, *k* is a sufficiently large tuning parameter, and $\hat{\mathbf{R}}$ is the Kendall's tau correlation matrix. Equation (3.4) is a combinatorial optimization problem and hard to compute. The corresponding global optimum is denoted by $\hat{\theta}^*_{1,k}$.

Because the estimator $\hat{\theta}^*_{1,k}$ 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 $\widetilde{\theta}_{1,k}$.. Here $k := ||\widetilde{\theta}_{1,k}||_0$ 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 x_t to the intersection of the *d*-dimension sphere $\mathbb{S}^{d-1}$ and the ℓ_0 ball with radius $k > 0$. In detail, we sort the absolute values of the elements of 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 ^ {1, …, *d*}, we define the truncation function TRC(·,·) to be

$$TRC(\boldsymbol{v}, J) := (v_1 \cdot I(1 \in J), \ldots, v_d \cdot I(d \in J))^T, \tag{3.5}$$

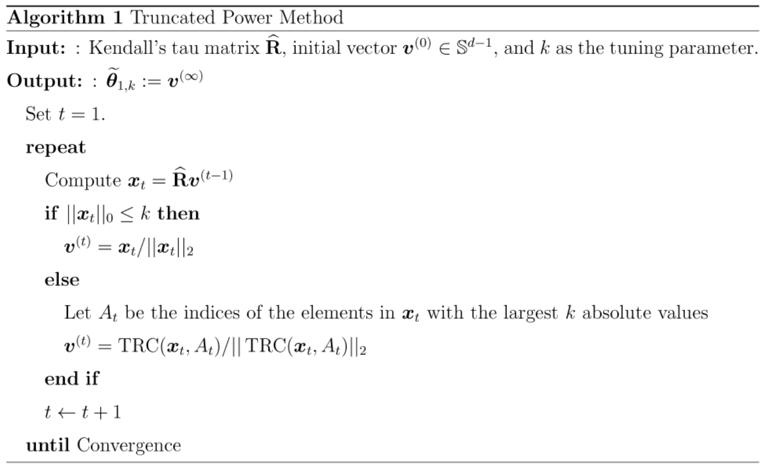
where *I*(·) is the indicator function. The truncated power algorithm is presented in Algorithm 1.

The formulation of the truncated power algorithm is nonconvex and the performance of the estimator relies on the selection of the initial vector $v^{(0)}$. In practice, we use the estimate obtained from the SPCA algorithm (Zou et al., 2006) as the initial vector. We set the termination criteria to be $\|\boldsymbol{v}^{(t)} - \boldsymbol{v}^{(t-1)}\|_2 \le 10^{-4}$.

In Section 4, we show that, by appropriately setting the initial vector $\boldsymbol{v}^{(0)}$, the algorithm converges and the corresponding estimator $\widetilde{\theta}_{1,k}$ is a consistent estimator of $\boldsymbol{\theta}_1$. In practice, we find that this algorithm always converges on all the synthetic and real-world data.

Algorithm 1

Truncated Power Method



```
Algorithm 1 Truncated Power Method
Input: : Kendall's tau matrix $\widehat{\mathbf{R}}$, initial vector $\boldsymbol{v}^{(0)} \in \mathbb{S}^{d-1}$, and $k$ as the tuning parameter.
Output: : $\widetilde{\boldsymbol{\theta}}_{1,k} := \boldsymbol{v}^{(\infty)}$
  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
```

3.3 Estimating the Top *m* Leading Eigenvectors

We exploit the iterative deflation method to estimate the top *m* leading eigenvectors $\theta_1, \ldots, \theta_m$ of Σ^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 semidefinite matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{\mathbf{d}\times\mathbf{d}}$, its deflation with respect to the vector $\boldsymbol{v} \in \mathbb{R}^d$ is defined as:

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

In this way, $\mathbf{D}(\Gamma, \boldsymbol{v})$ is positive semidefinite, left and right orthogonal to $\boldsymbol{v}$, and symmetric. To estimate $\theta_1, \ldots, \theta_m$, we exploit the following approach: (i) The estimate $\hat{\theta}_1$ (can be either $\hat{\theta}^*_{1,k}$ or $\widetilde{\theta}_{1,k}$) of θ_1 is calculated using Equation (3.4) or the truncated power method; (ii) Given $\hat{\theta}_1, \ldots, \hat{\theta}_j$, we estimate $\hat{\theta}_{j+1}$ by plugging $\boldsymbol{\Gamma}^{(j+1)} := \mathbf{D}\left(\boldsymbol{\Gamma}^{(j)}, \hat{\theta}_j\right)$ into Equation (3.4) or the truncated power method ($\Gamma^{(1)} := \Sigma^0$).

4 Theoretical Properties

In this section we provide the theoretical properties of the estimators $\hat{\theta}^*_{1,k}$ and $\widetilde{\theta}_{1,k}$. 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).

4.1 Latent Generalized Correlation Matrix Estimation

In this section we focus on estimating the latent generalized correlation matrix Σ^0. In the next theorem we prove the rate of convergence $O_P\left(\sqrt{log \quad d/n}\right)$ for $|\hat{\mathbf{R}}_{jk} - \mathbf{\Sigma}^0_{jk}|$ uniformly over all indices *j, k*. This is an important result, which indicates that the Gaussian parametric rate in estimating the correlation matrix obtained by Bickel and Levina (2008) can be extended to the meta-elliptical distribution family using the Kendall's tau statistic.

Theorem 4.1. *Let* $\boldsymbol{x}_1, \ldots, x_n$ *be n observations of* $\boldsymbol{X} \sim ME_d(\Sigma^0, \xi; f_1, \ldots, f_d)$ *and let* $\hat{\mathbf{R}}$ be defined as in Equation (3.3). We have, with probability at least $1 - d^{-5/2}$,

$$||\hat{\mathbf{R}} - \mathbf{\Sigma}^0||_{max} \leq 3\pi\sqrt{\frac{log \quad d}{n}}. \quad (4.1)$$

Proof. The result follows from Theorem 4.2 in Liu et al. (2012) but with a slightly different probability bound. A detailed proof is provided in Appendix A.2 for self-completeness.

4.2 Leading Eigenvector Estimation

We analyze the estimation errors of the global optimum $\hat{\theta}^*_{1,k}$ and the estimator $\widetilde{\theta}_{1,k}$ obtained from the truncated algorithm. We say that the model $\mathscr{M}_d\left(\mathbf{\Sigma}^0, \xi, f; \theta_1, s\right)$ holds if the data are drawn from one probability distribution in $\mathscr{M}_d\left(\mathbf{\Sigma}^0, \xi, f; \theta_1, s\right)$. The next theorem provides an upper bound on the angle between $\hat{\theta}^*_{1,k}$ and $\boldsymbol{\theta}_1$.

Theorem 4.2. *Let* $\hat{\theta}^*_{1,k}$ *be the global optimum to Equation (3.4), the model* $\mathscr{M}_d\left(\mathbf{\Sigma}^0, \xi, f; \theta_1, s\right)$ *hold, and k ≥ s. For any two vectors* $\boldsymbol{v}_1 \in \mathbb{S}^{d-1}$ and $\boldsymbol{v}_2 \in \mathbb{S}^{d-1}$, *let* $|sin \quad \angle(\boldsymbol{v}_1, \boldsymbol{v}_2)| := \sqrt{1 - (\boldsymbol{v}_1^T\boldsymbol{v}_2)^2}$. *Then we have, with probability at least* $1 - d^{-5/2}$,

$$|sin \quad \angle\left(\hat{\theta}^*_{1,k}, \theta_1\right)| \leq \frac{6\pi}{\lambda_1 - \lambda_2} \cdot k\sqrt{\frac{log \quad d}{n}}, \quad (4.2)$$

when $\lambda_j := \Lambda_j(\Sigma^0)$ for $j = 1,2$.

Proof. The key idea of the proof is to exploit the results in Theorem 4.1 in bounding the estimation error. Detailed proofs are presented in Appendix A.3.

Remark 4.3. *When s,* λ_1, λ_2 *do not scale with* (*n, d*) *and k* ≥ s is a fixed constant, the rate of convergence in parameter estimation is $O_P\left(\sqrt{log \quad d/n}\right)$, 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. Given that 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 θ_1.

Corollary 4.4 (Feature selection). *Let* $\hat{\theta}^*_{1,k}$ *be the global optimum to* Equation (3.4), *the model* $\mathscr{M}_d\left(\boldsymbol{\Sigma}^0, \xi, f; \theta_1, s\right)$ *hold, and k* ≥ *s. Let* T:= *supp*(θ_1), *and* $\hat{\boldsymbol{\Theta}}^*_k := supp\left(\hat{\theta}^*_{1,k}\right)$. *If we further have* $min_{j\in\Theta}|\theta_{1j}| \geq \frac{6\sqrt{2\pi}}{\lambda_1-\lambda_2}\cdot k\sqrt{\frac{log \quad d}{n}}$, then $\mathbb{P}\left(\Theta \subset \hat{\Theta}^*_k\right) \geq 1-d^{-5/2}$.

Proof. The key of the proof is to construct a contradiction given Theorem 4.2 and the condition on the minimum value of $\{|\theta_{1j}|\}_{j=1}^d$. Detailed proof is shown in Appendix A.4

In the next theorem, we provide a result on the convergence rate of the estimator $\widetilde{\theta}_{1,k}$ obtained by exploiting the truncated power algorithm. This theorem, coming from Yuan and Zhang (2013), indicates that under sufficient conditions $\widetilde{\theta}_{1,k}$ converges to θ_1 in a $s\sqrt{log \quad d/n}$ rate.

Theorem 4.5. *If the model* $\mathscr{M}_d\left(\boldsymbol{\Sigma}^0, \xi, f; \theta_1, s\right)$ holds, the conditions in Theorem 1 in Yuan and Zhang (2013) hold, and k ≥ *s, we have, with probability at least* $1 - d^{-5/2}$,

$$|sin \quad \angle\left(\widetilde{\theta}_{1,k}, \theta_1\right)| \leq C\cdot(s+2k)\sqrt{\frac{log \quad d}{n}},$$

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

The result in Theorem 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 4.4, it can be shown that under certain conditions, $supp\left(\theta_1\right) \subset supp\left(\widetilde{\theta}_{1,k}\right)$ with high probability.

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 $\boldsymbol{X} \sim ME_d(\Sigma^0, \xi; f_1, \ldots, f_d)$ follow a meta-elliptical distribution and $x_1, \ldots, x_n$ be n observations of $\boldsymbol{X}$ with $x_i := (x_{i1}, \ldots, x_{id})^T$. Let $Z := (f_1(X_1), \ldots, f_d(X_d))^T$ be the transformed random vector. By definition, $\boldsymbol{Z} \sim EC_d(0, \Sigma^0, \xi)$ is elliptically distributed. Let Q_g be the marginal distribution function of $\boldsymbol{Z}$ (From Proposition A.2, 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 $\hat{F}_j\left(t; \delta_n\right)$ be defined as

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

We define

$$\hat{f}_j\left(t;\delta_n\right) := Q_g^{-1}\left(\hat{F}_j\left(t;\delta_n\right)\right) \quad (4.3)$$

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 that $\hat{f}_j\left(\cdot;\delta_n\right)$ converges to $f_j(\cdot)$ uniformly over an expanding interval with high probability.

Theorem 4.6 (Han et al. (2013)). *Suppose that* $\boldsymbol{X} \sim NPN_d(\Sigma^0; f_1, \ldots, f_d)$ *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*

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

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

Using Theorem 4.6, we have the following theorem, which shows that, under appropriate conditions, we can recover the first principal component of any data point *x*.

Theorem 4.7. *For any observation* $\boldsymbol{x} := (x_1, \ldots, x_d)^T$ *of* $\boldsymbol{X} \sim NPN_d(\Sigma; f_1, \ldots, f_d)$, *under the conditions of Theorem 4.2, letting*

$$\hat{f}\left(\boldsymbol{x}\right) := \left(\hat{f}_1\left(x_1;(2n)^{-1}\right), \ldots, \hat{f}_d\left(x_d;(2n)^{-1}\right)\right)^T \quad \text{and} \quad f\left(\boldsymbol{x}\right) := \left(f_1\left(x_1\right), \ldots, f_d\left(x_d\right)\right)^T,$$

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

$$|\hat{f}(\boldsymbol{x})^T\hat{\theta}_{1,k}^{*} - f(\boldsymbol{x})^T\theta_1^{*}| = O_P\left(\sqrt{(s+k)\cdot\frac{log \quad log \quad n}{n^{1-b/2}}} + \frac{k}{\lambda1-\lambda_2}\sqrt{\frac{(s+k)\, log \quad d \quad log \quad n}{n}}\right),$$

where $\theta_1^{*} := sign\left(\theta_1^T\hat{\theta}_{1,k}^{*}\right)\cdot\theta_1$.

Proof. Theorem 4.7 is proved by combining the results in Theorems 4.2 and 4.6. More details are presented in Appendix A.5.

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 that 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.

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 Σ^0 using a similar idea as in Yuan and Zhang (2013): First a covariance matrix 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

$$\boldsymbol{\Sigma} := \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{2}} (\omega_{\mathbf{j}} - \mathbf{1})\, \mu_{\mathbf{j}} \mu_{\mathbf{j}}^{\mathbf{T}} + \mathbf{I}_{\mathbf{d}}, \quad \text{where} \quad \omega_{\mathbf{1}} = \mathbf{6}, \omega_{\mathbf{2}} = \mathbf{3}.$$

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

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

The latent generalized correlation matrix Σ^0 is $\Sigma^0 = \text{diag}\Sigma()^{-1/2} \cdot \Sigma \cdot \text{diag}(\Sigma)^{-1/2}$. 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 $x_1, \ldots, x_n$ be n observations of $\boldsymbol{X} \sim N_d(0, \Sigma^0)$.

Scheme 2: Let $x_1, \ldots, x_n$ be n observations of $\boldsymbol{X} \sim N_d(0, \Sigma^0)$, but with 5% entries in each x_i randomly picked up and replaced by −5 or 5.

Scheme 3: Let $x_1, \ldots, x_n$ be n observations of $\boldsymbol{X} \sim NPN_d(\Sigma^0; f_1, \ldots, f_1)$ with $f_1(x) = x^3$.

Scheme 4: Let $x_1, \ldots, x_n$ be n observations of $\boldsymbol{X} \sim ME_d(\Sigma^0, \xi_1; f_0, \ldots, f_0)$ with $f_0(x) = x$ and $\xi_1 \overset{d}{=} \sqrt{\kappa}\xi_1^*/\xi_2^*$. Here $\xi_1^* \overset{d}{=} \chi_d$ and $\xi_2^* \overset{d}{=} \chi_\kappa$ 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 $x_1, \ldots, x_n$ be n observations of $\boldsymbol{X} \sim ME_d(\Sigma^0, \xi_2; f_0, \ldots, f_0)$ 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 $x_1, \ldots, x_n$ be n observations of $\boldsymbol{X} \sim ME_d(\Sigma^0, \xi_3; f_0, \ldots, f_0)$ with ξ_3 follows an exponential distribution with the rate parameter 1.

Here Schemes 1 to 3 represent three different versions of the 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 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 θ_1, Figure 3 plots the false positive rates against the true positive rates for the four different estimators under different schemes.

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 shows the averaged distances between the estimated leading eigenvector and θ_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 $|sin \quad \angle(\boldsymbol{v}_1, \boldsymbol{v}_2)|$.

Both Figure 3 and Table 2 show that when the data are non-Gaussian but follow 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 that 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.

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. 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 $X_{tj} := \log(St_{t,j}/St_{t-1,j})$.

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 X and obtain the top four leading eigenvectors. Here the iterative deflation method discussed in Section 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 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 successful matches of the market trend using the stocks in A_k as:

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

We visualize the result by plotting (card(A_k), ρ_{A_k}) in Figure 4, which shows that Kendall summarizes the trend of the whole stock market better than the other three methods.

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), Telecommunications Services (6 stocks), Materials (29 stocks), and Utilities (32 stocks). Table 3 provides a more detailed description of these ten categories with their numbers and abbreviations provided.

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 4. We see that, in general, Kendall has the best ability in grouping the stocks of the same category together. Therefore, Kendall provides a more interpretable result.

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 fundamentally 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 paper 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. (2012) and Xue and Zou (2012) proposed a procedure called the nonparanormal SKEPTIC, which exploits the nonparanormal family for graph estimation. The non-paranormal 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 paper focuses on principal component analysis. In a preliminary version of this work (Han and Liu, 2012), they mainly focused on estimating the first leading eigenvector of the latent generalized correlation matrix by directly solving Equation (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 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 $\hat{f}_j(t) = \left(t - \hat{\mu}_j\right)/\hat{\sigma}_j$. Here $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the marginal sample mean and standard deviation for the *j*-th entry.
2. If not, we construct a set of marginal distribution functions:

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

3. For any $Q_g \in \Pi$, we calculate $\hat{f} = \left\{\hat{f}_1, \ldots, \hat{f}_d\right\}$ using Equation (4.3).
4. We transform the data using $\hat{f}$.
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.

Acknowledgments

We thank the associate editor and two anonymous reviewers for their very helpful and constructive comments and suggestions. Fang Han's research is supported by Google fellowship in statistics. Han Liu's research is supported by NSF Grants III-1116730 and NSF III-1332109, an NIH sub-award and a FDA sub-award from Johns Hopkins University.

A Appendix

A.1 Properties of the Elliptical Distribution

The next proposition provides two alternative ways to characterize an elliptical distribution and their proofs can be found in Fang et al. (1990).

Proposition A.1 (Fang et al. (1990)). *A random vector* $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$ *satisfies that* $\mathbf{Z} \sim EC_d(\boldsymbol{\mu}, \Sigma, \xi)$ *if and only if* $\mathbf{Z}$ *has the characteristic function* $\exp(i\mathbf{t}'\boldsymbol{\mu})\phi(\mathbf{t}'\Sigma\mathbf{t})$, *where* i: $\sqrt{-1}$ and ϕ is a properly-defined characteristic function. We denote $\mathbf{Z} \sim \mathrm{EC}_d(\boldsymbol{\mu}, \Sigma, \phi)$ *in this setting. If* ξ *is absolutely continuous and* Σ is non-singular, then the density of $\mathbf{Z}$ exists and is of the form:

$$p_Z(z) = |\Sigma|^{-1/2} g\left((z-\mu)^T \Sigma^{-1} (z-\mu)\right),$$

where $g : [0, \infty) \to [0,\infty)$. *We denote* $\mathbf{Z} \sim EC_d(\boldsymbol{\mu},\Sigma, g)$. Here ξ, φ and g uniquely determine one of the other.

The next proposition provides three important properties of the elliptical distribution.

Proposition A.2. If a random vector Z is elliptically distributed, we have:

- *For any* $q \in \mathbb{N}$,$\boldsymbol{v} \in \mathbb{R}^q$, *and any matrix* $\mathbf{B} \in \mathbb{R}^{d\times q}$, $\boldsymbol{v} + \mathbf{B}^T\mathbf{Z}$ *is elliptically distributed. In particular, if* $\mathbf{Z} \sim EC_d(\boldsymbol{\mu},\Sigma,\xi)$, *then* $\boldsymbol{v} + \mathbf{B}^T\mathbf{Z} \sim EC_q(\boldsymbol{v} + \mathbf{B}^T\boldsymbol{\mu}, \mathbf{B}^T\Sigma\mathbf{B},\xi)$.
- *Let* $\mathbf{Z} \sim EC_d(\boldsymbol{\mu},\Sigma,\xi)$ *and* Σ^0 *be the generalized correlation matrix of* $\mathbf{Z}$. *If rank*$(\Sigma) = q$ *and* $\mathbb{E}\xi^2<\infty$, *then* $\mathbb{E}(\boldsymbol{Z}) = \mu$, $Cov(\boldsymbol{Z}) = \frac{\mathbb{E}(\xi^2)}{q}\boldsymbol{\Sigma}$, *and* $\mathrm{Cor}(\mathbf{Z}) = \Sigma^0$.
- *If* $\mathbf{Z} \sim EC_d(0,\Sigma,\phi)$ *with* $\mathrm{diag}(\Sigma) = \mathbf{I}_d$, *then the marginal distributions of* $\mathbf{Z}$ *are the same.*

Proof. The proof of the first two assertions can be found in Fang et al. (1990). To prove the third assertion, we use Proposition A.1 to obtain the characteristic function of Z_j for any $1 \le j \le d$:$\mathbb{E}\quad exp(itZ_j) = \mathbb{E}\quad exp\left(ite_j^T\boldsymbol{Z}\right) = \psi\left(t^2\boldsymbol{e}_j^T\boldsymbol{\Sigma}\boldsymbol{e}_\mathbf{j}\right) = \psi\left(t^2\right)$, where $t \in \mathbb{R}$ and e_j is the *j*-th canonical basis in $\mathbb{R}^d$, i.e., $e_j = (\underbrace{0,\ldots,0}_{j-1},1,0,\ldots,0)$ for $1 \le j \le d$. The result follows from the one-to-one map between the characteristic functions and the random variables.

A.2 Proof of Theorem 4.1

Proof. Realizing that $\hat{\tau}_{jk}$ is a 2nd order U-statistic and sign $(x_{ij} - x_{i'k})$ $(x_{ik} - x_{i'k})$ is bounded in [−1; 1], using Equation (5.7) in Hoeffding (1963), we have

$$\mathbb{P}\left(|\hat{\tau}_{jk} - \tau(X_j,X_k)|>t\right) = \mathbb{P}\left(|\hat{\tau}_{jk} - \mathbb{E}(\hat{\tau}_{jk})|>t\right) \le 2\,exp\left(-\frac{nt^2}{8}\right).$$

Therefore, we have

$$\begin{aligned}\mathbb{P}\left(|\hat{\mathrm{R}}_{jk} - \Sigma^0_{jk}|>t\right) &= \mathbb{P}\left(|sin\left(\tfrac{\pi}{2}\hat{\tau}_{jk}\right) - sin\left(\tfrac{\pi}{2}\tau(X_j,X_k)\right)|>t\right)\\ &\le \mathbb{P}\left(|\hat{\tau}_{jk} - \tau(X_j,X_k)|>\tfrac{2}{\pi}t\right)\\ &\le 2\,exp\left(-\tfrac{nt^2}{2\pi^2}\right).\end{aligned}$$

Taking the union bound, we have

$$\mathbb{P}\left(||\hat{\mathrm{R}}-\Sigma^0||_{max}>t\right)\le d^2\,exp\left(-\frac{nt^2}{2\pi^2}\right).$$

This completes the proof.

A.3 Proof of Theorem 4.2

Proof. Under the model $\mathscr{M}_d\left(\mathbf{\Sigma}^0,\xi,f;\theta_1,s\right)$, we define $\lambda_j := \Lambda_j(\Sigma^0)$ and θ_j to be the corresponding eigenvector for $j = 1, \ldots, d$. We then define $\mathbf{\Psi}_0{:}\Sigma_{j=2}^{d}\lambda_j\theta_j\theta_j^T$ and let

$$\varepsilon := |sin\angle\left(\theta_1,\hat{\theta}^*_{1,k}\right)| \quad \text{and} \quad \mathbf{\Sigma}^0=\lambda_1\theta_1\theta_1^T+\mathbf{\Psi}_0.$$

For all $\theta\in\mathbb{S}^{d-1}$, we have

$$\begin{aligned}\left\langle\mathbf{\Sigma}^0,\theta_1\theta_1^T-\theta\theta^T\right\rangle &=\left\langle\mathbf{\Sigma}^0,\theta_1\theta_1^T\right\rangle-\left\langle\lambda_1\theta_1\theta_1^T+\mathbf{\Psi}_0,\theta\theta^T\right\rangle\\ &=\lambda_1-\lambda_1\langle\theta_1,\theta\rangle^2-\left\langle\mathbf{\Psi}_0,\theta\theta^T\right\rangle,\end{aligned} \qquad \text{(A.1)}$$

and

$$\begin{aligned}\left\langle\mathbf{\Psi}_0,\theta\theta^T\right\rangle &=\theta^T\mathbf{\Psi}_0\theta=\theta^T\left(\mathbf{I}_d-\theta_1\theta_1^T\right)\mathbf{\Sigma}^0\left(\mathbf{I}_d-\theta_1\theta_1^T\right)\theta\\ &\le\lambda_2||\left(\mathbf{I}_d-\theta_1\theta_1^T\right)\theta||_2^2=\lambda_2-\lambda_2\langle\theta_{\mathbf{1}},\theta\rangle^2.\end{aligned}$$

Moreover, by definition,

$$sin^2\angle\left(\theta_1,\theta\right)=1-\left(\theta_1^T\theta\right)^2=1-\langle\theta_{\mathbf{1}},\theta\rangle^2. \quad \text{(A.2)}$$

Combining Equation (A.1) with Equation (A.2), we have

$$\left\langle\mathbf{\Sigma}^0,\theta_1\theta_1^T-\theta\theta^T\right\rangle\ge\left(\lambda_1-\lambda_2\right)\left(1-\langle\theta_1,\theta\rangle^2\right)=\left(\lambda_1-\lambda_2\right)\,sin^2\angle\left(\theta_1,\theta\right).$$

Therefore, letting $\hat{\theta}^*_{1,k}$ be the global optimum to Equation (3.4), we have

$$\begin{aligned}\varepsilon^2 &\le\tfrac{1}{\lambda_1-\lambda_2}\left\langle\mathbf{\Sigma}^0,\theta_1\theta_1^T-\hat{\theta}^*_{1,k}\hat{\theta}^{*T}_{1,k}\right\rangle\\ &\le\tfrac{1}{\lambda_1-\lambda_2}\left(\left\langle\mathbf{\Sigma}^0-\hat{\mathbf{R}},\theta_1\theta_1^T-\hat{\theta}^*_{1,k}\hat{\theta}^{*T}_{1,k}\right\rangle+\left\langle\hat{\mathbf{R}},\theta_1\theta_1^T-\hat{\theta}^*_{1,k}\hat{\theta}^{*T}_{1,k}\right\rangle\right)\\ &\le\tfrac{1}{\lambda_1-\lambda_2}\left\langle\mathbf{\Sigma}^0-\hat{\mathbf{R}},\theta_1\theta_1^T-\hat{\theta}^*_{1,k}\hat{\theta}^{*T}_{1,k}\right\rangle.\end{aligned} \qquad \text{(A.3)}$$

The last inequality holds because θ_1 is feasible in the optimization constraint in (3.4), implying that

$$\left\langle \hat{\mathbf{R}}, \theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T} \right\rangle = Tr\left(\hat{\mathbf{R}}\theta_1\theta_1^T\right) - Tr\left(\hat{\mathbf{R}}\hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T}\right) = \theta_1^T\hat{\mathbf{R}}\theta_1 - \hat{\theta}_{1,k}^{*T}\hat{\mathbf{R}}\hat{\theta}_{1,k}^{*} \le 0.$$

Therefore, using Equation (A.3),

$$\begin{aligned} \varepsilon^2 &\le \frac{1}{\lambda_1-\lambda_2}\left\langle \mathbf{\Sigma}^0, -\hat{\mathbf{R}}, \theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T} \right\rangle \\ &= \frac{1}{\lambda_1-\lambda_2}\left\langle vec\left(\mathbf{\Sigma}^0 - \hat{\mathbf{R}}\right),\ vec\left(\theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T}\right) \right\rangle \\ &\le \frac{1}{\lambda_1-\lambda_2}\|vec\left(\hat{\mathbf{R}} - \mathbf{\Sigma}^0\right)\|_\infty \cdot \|vec\left(\theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T}\right)\|_1, \end{aligned} \qquad \text{(A.4)}$$

where the last inequality is by the Hölder inequality. Letting $\vartheta = vec\left(\theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T}\right.$, we have

$$\begin{aligned} \|\vartheta\|_2^2 &= \|\theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T}\|_F^2 = Tr\left(\left(\theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T}\right)\left(\theta_1\theta_1^T - \hat{\theta}_{1,k}^{*}\hat{\theta}_{1,k}^{*T}\right)\right) \\ &= 2\left(1 - \left(\theta_1^T\hat{\theta}_{1,k}^{*}\right)^2\right) = 2\varepsilon^2, \end{aligned}$$

implying that

$$\|\vartheta\|_1 \le \sqrt{\|\boldsymbol{\vartheta}\|_0}\|\boldsymbol{\vartheta}\|_2 \le \sqrt{2k^2}\cdot\sqrt{2\varepsilon^2} = 2k\varepsilon. \qquad \text{(A.5)}$$

Therefore, combining Equation (A.4) with Equation (A.5), we get

$$\varepsilon^2 \le \frac{\|vec\left(\hat{\mathbf{R}} - \mathbf{\Sigma}^0\right)\|_\infty}{\lambda_1 - \lambda_2}.2k\varepsilon,$$

which is equivalent to saying that

$$\varepsilon \le \frac{2k\|vec\left(\hat{\mathbf{R}} - \mathbf{\Sigma}^0\right)\|_\infty}{\lambda_1 - \lambda_2}.$$

Using Theorem 4.1, we have, with probability at least $1 - d^{-5/2}$,

$$\epsilon \le \frac{2k\|vec\left(\hat{\mathbf{R}} - \mathbf{\Sigma}^0\right)\|_\infty}{\lambda_1 - \lambda_2} \le \frac{6\pi}{\lambda_1 - \lambda_2}\cdot k\sqrt{\frac{log \quad d}{n}}.$$

This completes the proof.

A.4 Proof of Corollary 4.4

Proof. Without loss of generality, we may assume that $\theta_1^T\hat{\theta}_{1,k}^* \ge 0$ because otherwise we can simply conduct appropriate sign changes in the proof. We first note that card $\left(\hat{\boldsymbol{\Theta}}_k^*\right)=k \ge \text{card}\left(\Theta\right)$. if $\Theta \not\subset \hat{\Theta}_k^*$, then $\Theta/\hat{\Theta}_k^* \neq 0$. This implies that,

$$||\hat{\theta}_{1,k}^* - \theta_1||_2 \ge ||\left(\hat{\theta}_{1,k}^* - \theta_1\right)_{\Theta/\hat{\Theta}_k^*}||_2 \ge \underset{j\in\Theta}{min}|\theta_{1_j}| \ge \frac{6\sqrt{2\pi}}{\lambda_1-\lambda_2}\cdot k\sqrt{\frac{log \quad d}{n}}.$$

We then have

$$sin^2 \quad \angle\left(\hat{\theta}_{1,k}^*, \theta_1\right) = 1-\left(\theta_1^T\hat{\theta}_{1,k}^*\right)^2 \ge 1-\theta_1^T\hat{\theta}_{1,k}^* = \frac{||\theta_1-\hat{\theta}_{1,k}^*||_2^2}{2},$$

implying that

$$|sin \quad \angle\left(\hat{\theta}_{1,k}^*, \theta_1\right)| \ge \frac{||\hat{\theta}_{1,k}^*-\theta_1||_2}{\sqrt{2}} \ge \frac{min_{j\in\Theta}|\theta_{1_j}|}{\sqrt{2}} = \frac{6\pi}{\lambda_1-\lambda_2}\cdot k\sqrt{\frac{log \quad d}{n}}. \quad \text{(A.6)}$$

Therefore, applying Theorem 4.2, we get

$$\mathbb{P}\left(\Theta \not\subset \hat{\Theta}_k^*\right) \le \mathbb{P}\left(|sin\angle\left(\hat{\theta}_{1,k}^*, \theta_1\right)| \ge \frac{6\pi}{\lambda_1-\lambda_2}\cdot k\sqrt{\frac{log \quad d}{n}} \le d^{-5/2}\right). \quad \text{(A.7)}$$

This completes the proof.

A.5 Proof of Theorem 4.7

Proof. Without loss of generality, we assume that $\theta_1^T\hat{\theta}_{1,k}^* \ge 0$. We define $\mathscr{S}=\{j_1,\ldots,j_v\}=supp\left(\theta_1\right)\cup supp\left(\hat{\theta}_{1,k}^*\right)$, where $v=\text{card}\left(\mathscr{S}\right)\le s+k$. we further define

$$T_n=\left[g_{j_1}\left(-\sqrt{b \quad log \quad n}\right), g_{j_1}\left(\sqrt{b \quad log \quad n}\right]\times\ldots\times\left[g_{j_v}\left(-\sqrt{b \quad log \quad n}\right), g_{j_v}\left(\sqrt{b \quad log \quad n}\right)\right],$$

for some 0 < *b* < 1. Moreover, we define the event M_n as

$$M_n := \left\{\boldsymbol{v}\in\mathbb{R}^d{:}\boldsymbol{v}_{\mathscr{S}}\in T_n\right\}.$$

Thus, conditioning on M_n, using Theorem 4.6, we have

$$\begin{aligned}
&|\hat{f}(\boldsymbol{x})^T\hat{\theta}^*_{1,k}\\
&\quad -f(\boldsymbol{x})^T\theta^*_1| \le |\left(\hat{f}(\boldsymbol{x})-f(\boldsymbol{x})\right)^T\hat{\theta}^*_{1,k}|\\
&\quad +|f(\boldsymbol{x})^T\left(\hat{\theta}^*_{1,.k}-\theta^*_1\right)| \le ||\left(\hat{f}(x)-f(x)\right)_{\mathscr{S}}||_2\\
&\quad +||(f(x))_{\mathscr{S}}||_2||\hat{\theta}^*_{1,k}-\theta^*_1||_2 O_P\left(\sqrt{(s+k)\cdot\frac{log\quad log\quad n}{n^{1-b/2}}}\right)\\
&\quad +O_P\left(\frac{k}{\lambda_1-\lambda_2}\sqrt{\frac{log\quad d}{n}}\right)\cdot||(f(\boldsymbol{x}))_{\mathscr{S}}||_2.
\end{aligned} \tag{A.8}$$

Since $f_j(x_j) \sim N_d(0, 1)$, using Mill’s inequality, we have

$$|f_j(x_j)|=O_P\left(\sqrt{log\quad n}\Rightarrow||(f(\boldsymbol{x}))_{\mathscr{S}}||_2=O_P\left(\sqrt{(s+k)\,log\quad n}\right).\right. \tag{A.9}$$

Finally, using Mill’s inequality again, we have

$$\mathbb{P}=\left(f_j(x_j)\ge\sqrt{n\quad log\quad n}\right)=O\left(n^{-b/2}\right)\Rightarrow\mathbb{P}(\boldsymbol{x}\in M_n^c)=O\left((s+k)\,n^{-b/2}\right)=o(1). \tag{A.10}$$

Combining Equations (A.8), (A.9), and (A.10), we have the desired result.

References

Amini A, Wainwright M. High-dimensional analysis of semidefinite relaxations for sparse principal components. The Annals of Statistics. 2009; 37(5B):2877–2921.

Anderson, TW. An Introduction to Multivariate Statistical Analysis. Vol. 2. Wiley; 1958.

Berthet Q, Rigollet P. Optimal detection of sparse principal components in high dimension. forthcoming in the Annals of Statistics. 2012

Bickel P, Levina E. Regularized estimation of large covariance matrices. The Annals of Statistics. 2008; 36(1):199–227.

Bühlmann, P.; van de Geer, S. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer; 2011.

Chatfield, C.; Collins, A. Introduction to Multivariate Analysis. Vol. 166. Chapman & Hall; 1980.

Choi K, Marden J. A multivariate version of Kendall’s τ. Journal of Non-parametric Statistics. 1998; 9(3):261–293.

Croux C, Dehon C. Influence functions of the Spearman and Kendall correlation measures. Statistical Methods & Applications. 2010; 19(4):497–515.

Croux C, Filzmoser P, Fritz H. Robust sparse principal component analysis. Technometrics. 2013; 55(2):202–214.

Croux C, Haesbroeck G. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika. 2000; 87(3):603–618.

Croux C, Ollila E, Oja H. Sign and rank covariance matrices: statistical properties and application to principal components analysis. Statistics in Industry and Technology. 2002:257–269.

Croux C, Ruiz-Gazen A. High breakdown estimators for principal components: the projection-pursuit approach revisited. Journal of Multivariate Analysis. 2005; 95(1):206–226.

d’Aspremont A, El Ghaoui L, Jordan M, Lanckriet G. A direct formulation for sparse pca using semidefinite programming. SIAM Review. 2004; 49(3):434–448.

Davies P. Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. The Annals of Statistics. 1987; 15(3):1269–1292.

Fang H, Fang K, Kotz S. The meta-elliptical distributions with given marginals. Journal of Multivariate Analysis. 2002; 82(1):1–16.

Fang, K.; Kotz, S.; Ng, K. Symmetric Multivariate and Related Distributions. Chapman&Hall; 1990.

Gibbons, JD.; Chakraborti, S. Nonparametric Statistical Inference. Vol. 168. CRC press; 2003.

Gnanadesikan R, Kettenring JR. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics. 1972; 28(1):81–124.

Hallin M, Paindaveine D, Verdebout T. Optimal rank-based testing for principal components. The Annals of Statistics. 2010; 38(6):3245–3299.

Hampel FR. The influence curve and its role in robust estimation. Journal of the American Statistical Association. 1974; 69(346):383–393.

Han F, Liu H. Transelliptical component analysis. Advances in Neural Information Processing Systems. 2012; 25:368–376.

Han F, Zhao T, Liu H. CODA: High dimensional copula discriminant analysis. Journal of Machine Learning Research. 2013; 14:629–671.

Hoeffding W. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association. 1963; 58(301):13–30.

Huber, PJ.; Ronchetti, E. Robust Statistics. 2nd edition. Wiley; 2009.

Hubert M, Rousseeuw PJ, Verboven S. A fast method for robust principal components with applications to chemometrics. Chemometrics and Intelligent Laboratory Systems. 2002; 60(1): 101–111.

Huffer FW, Park C. A test for elliptical symmetry. Journal of Multivariate Analysis. 2007; 98(2):256–281.

Jackson D, Chen Y. Robust principal component analysis and outlier detection with ecological data. Environmetrics. 2004; 15(2):129–139.

Johnstone I, Lu A. On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association. 2009; 104(486):682–693. [PubMed: 20617121]

Journée M, Nesterov Y, Richtárik P, Sepulchre R. Generalized power method for sparse principal component analysis. Journal of Machine Learning Research. 2010; 11:517–553.

Kendall, MG. Rank Correlation Methods. Griffin; 1948.

Kruskal W. Ordinal measures of association. Journal of the American Statistical Association. 1958; 53(284):814–861.

Li R, Fang K, Zhu L. Some Q-Q probability plots to test spherical and elliptical symmetry. Journal of Computational and Graphical Statistics. 1997; 6(4):435–450.

Lindskog, F.; McNeil, A.; Schmock, U. Kendall’s tau for elliptical distributions. Springer; 2003.

Liu H, Han F, Yuan M, Lafferty J, Wasserman L. High dimensional semiparametric gaussian copula graphical models. The Annals of Statistics. 2012; 40(4):2293–2326.

Ma Z. Sparse principal component analysis and iterative thresholding. forthcoming in the Annals of Statistics. 2013

Mackey L. Deflation methods for sparse PCA. Advances in Neural Information Processing Systems. 2009; 21:1017–1024.

Marden J. Some robust estimates of principal components. Statistics & Probability Letters. 1999; 43(4):349–359.

Maronna RA. Robust *M*-estimators of multivariate location and scatter. The Annals of Statistics. 1976; 4(1):51–67.

Maronna RA, Zamar RH. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics. 2002; 44(4):307–317.

Möttönen J, Oja H. Multivariate spatial sign and rank methods. Journal of Nonparametric Statistics. 1995; 5(2):201–213.

Oja, H. Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Vol. 199. Springer; 2010.

Paul D, Johnstone I. Augmented sparse principal component analysis for high dimensional data. Arxiv preprint arXiv:1202.1242. 2012

Puri, ML.; Sen, PK. Nonparametric Methods in Multivariate Analysis. Wiley; 1971.

Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Maechler M. Robustbase: basic robust statistics. R package. 2009 URL http://CRAN.R-project.org/package=robustbase.

Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. Journal of the American Statistical Association. 1993; 88(424):1273–1283.

Sakhanenko L. Testing for ellipsoidal symmetry: A comparison study. Computational Statistics & Data Analysis. 2008; 53(2):565–581.

Shen H, Huang J. Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis. 2008; 99(6):1015–1034.

Visuri S, Koivunen V, Oja H. Sign and rank covariance matrices. Journal of Statistical Planning and Inference. 2000; 91(2):557–575.

Vu V, Lei J. Minimax rates of estimation for sparse PCA in high dimensions. International Conference on Artificial Intelligence and Statistics (AISTATS). 2012; 15:1278–1286.

Witten D, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10(3):515–534. [PubMed: 19377034]

Xue L, Zou H. Regularized rank-based estimation of high-dimensional non-paranormal graphical models. The Annals of Statistics. 2012; 40(5):2541–2571.

Yuan X, Zhang T. Truncated power method for sparse eigenvalue problems. Journal of Machine Learning Research. 2013; 14:899–925.

Zhang Y, El Ghaoui L. Large-scale sparse principal component analysis with application to text data. Advances in Neural Information Processing Systems. 2011; 24

Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics. 2006; 15(2):265–286.

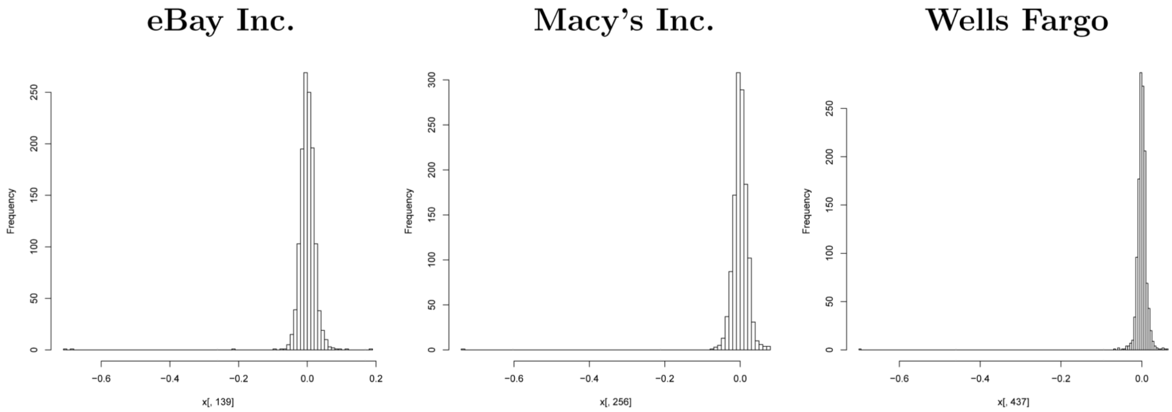


Figure 1.

Illustration of the asymmetry issue of the log-return stock data.

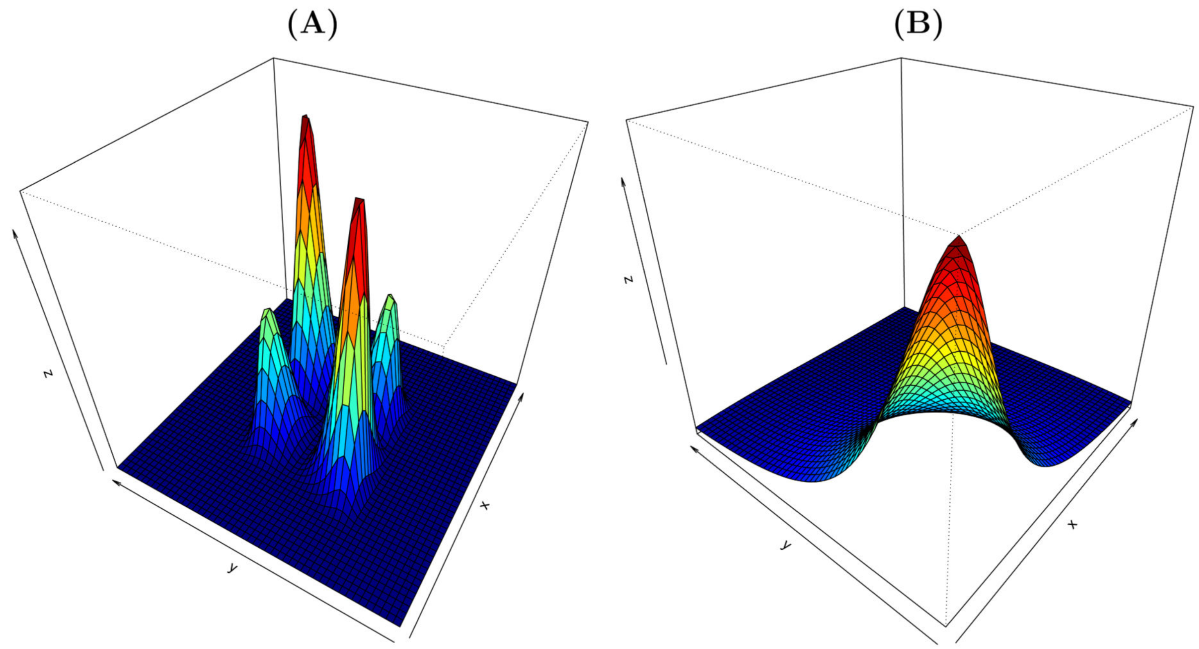


Figure 2.

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.

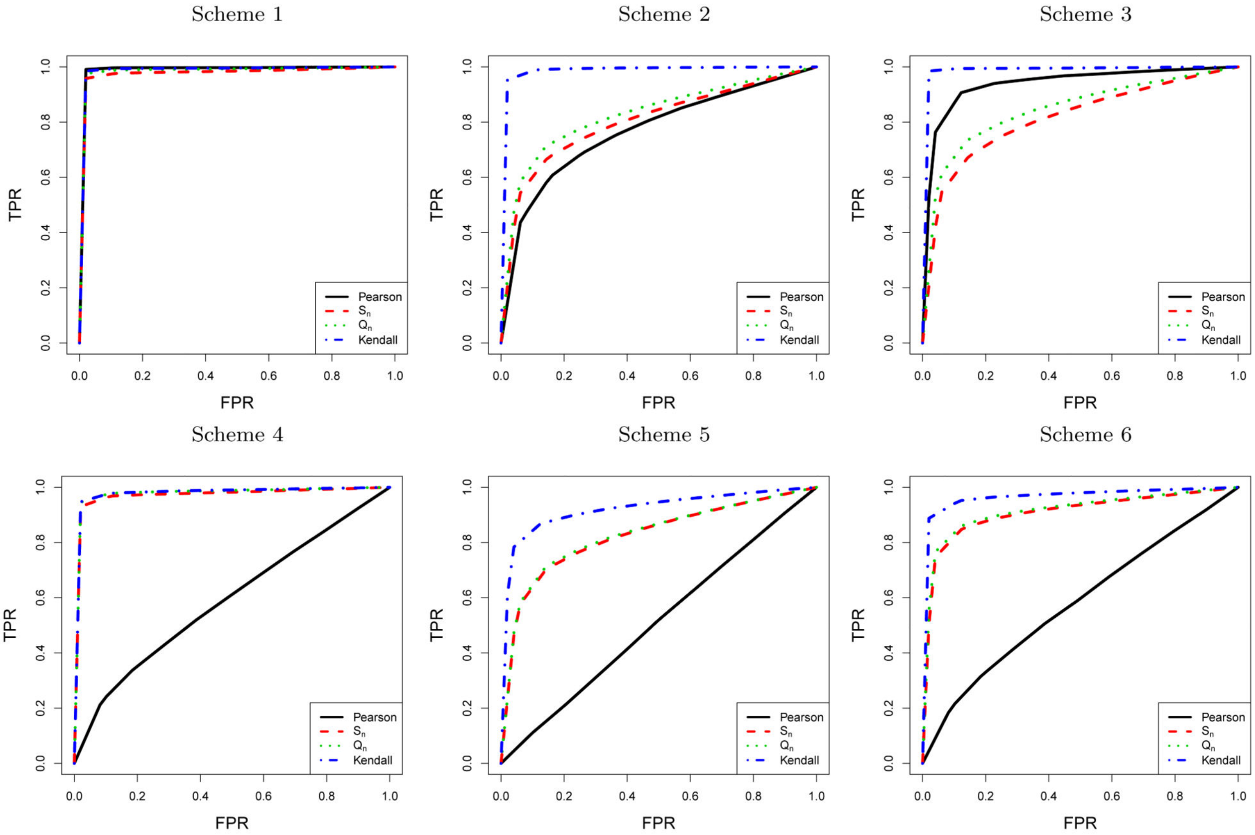


Figure 3.

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

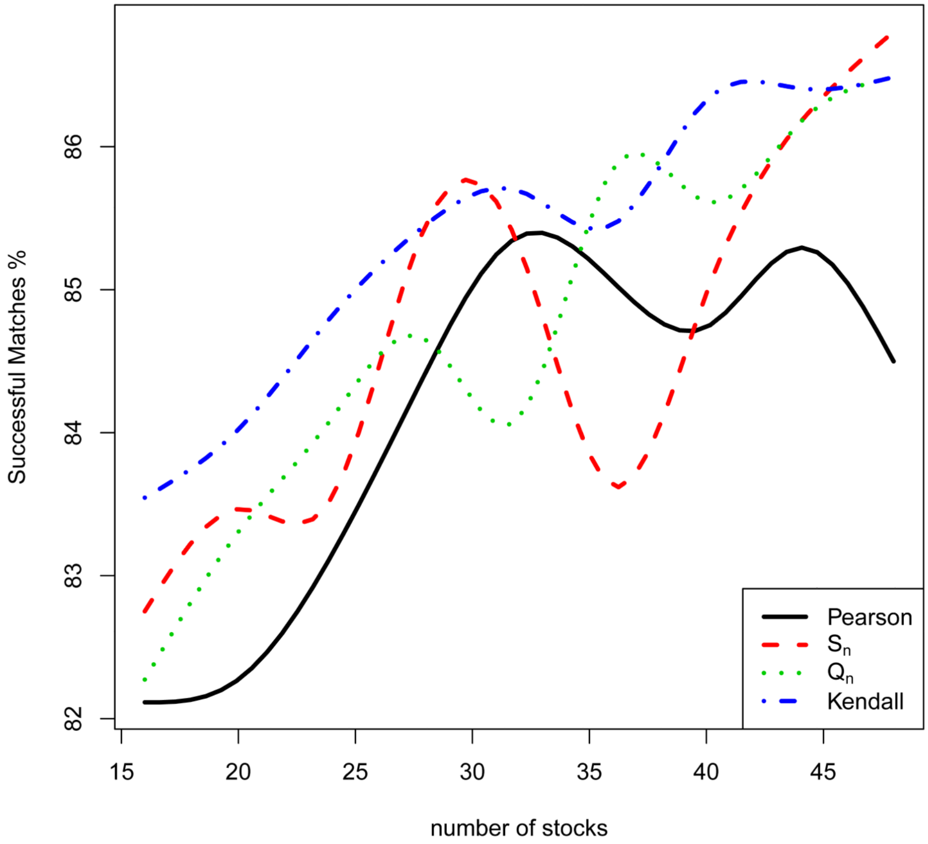


Figure 4.

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.

Table 1

Normality test of the stock 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

Table 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)

Table 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

Table 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 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