Principal Component Analysis on non-Gaussian Dependent Data

Fang Han FHAN@JHSPH.EDU
Johns Hopkins University, 615 N.Wolfe Street, Baltimore, MD 21205 USA

Han Liu HANLIU@PRINCETON.EDU
Princeton University, 98 Charlton Street, Princeton, NJ 08544 USA

Abstract

In this paper, we analyze the performance of a semiparametric principal component analysis named Copula Component Analysis (COCA) (Han & Liu, 2012) when the data are dependent. The semiparametric model assumes that, after unspecified marginally monotone transformations, the distributions are multivariate Gaussian. We study the scenario where the observations are drawn from non-i.i.d. processes (m-dependency or a more general ϕ-mixing case). We show that COCA can allow weak dependence. In particular, we provide the generalization bounds of convergence for both support recovery and parameter estimation of COCA for the dependent data. We provide explicit sufficient conditions on the degree of dependence, under which the parametric rate can be maintained. To our knowledge, this is the first work analyzing the theoretical performance of PCA for the dependent data in high dimensional settings. Our results strictly generalize the analysis in Han & Liu (2012) and the techniques we used have the separate interest for analyzing a variety of other multivariate statistical methods.

1. Introduction

In much of studies on Principal Component Analysis (PCA) it is assumed that the n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ are independent. Moreover, in high dimensions, it is commonly assumed that $\boldsymbol{X}$ follows a multivariate Gaussian or sub-Gaussian distribution such that the estimators are consistent.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

In this paper we focus on a semiparametric method built on the *nonparanormal* model. A continuous random vector $\boldsymbol{X} := (X_1, \ldots, X_d)^T$ follows a nonparanormal distribution if there exists a set of univariate monotone functions $f := \{f_j\}_{j=1}^d$ such that $f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T$ follows a Gaussian distribution. In this paper we show that the proposed method can loosen both the data independence and Gaussian/sub-Gaussian assumptions.

Let $\boldsymbol{\Sigma}$ be the covariance matrix of $\boldsymbol{X}$. PCA aims at recovering the top m leading vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of $\boldsymbol{\Sigma}$. The usual procedures are to estimate the top m leading eigenvectors $\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_m$ of the sample covariance matrix $\mathbf{S}$. However, there are two drawbacks: (i) when $d > n$, Johnstone & Lu (2009) show that PCA is inconsistent. More specifically, let $\boldsymbol{u}_1$ and $\widehat{\boldsymbol{u}}_1$ be the leading eigenvectors of $\boldsymbol{\Sigma}$ and $\mathbf{S}$. For two vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$, we denote the angle between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ by $\angle(\boldsymbol{v}_1, \boldsymbol{v}_2)$. Johnstone & Lu (2009) prove that $\angle(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1)$ does not converge to 0. (ii) The performances of the estimators rely on independence of the n observations. Their performance is unknown if dependence is present among $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

A remedy for the inconsistency problem when $d > n$ is to assume that $\boldsymbol{u}_1 = (u_{11}, \ldots, u_{1d})^T$ is sparse, i.e.,

$$\mathrm{card}(\mathrm{supp}(\boldsymbol{u}_1)) := \mathrm{card}(\{j : u_{1j} \neq 0\}) = s < n,$$

where $\mathrm{card}(\cdot)$ represents the cardinality of a given set. Different sparse PCA procedures have been developed to exploit the sparsity structure: greedy algorithms (d'Aspremont et al., 2008), lasso-type methods including SCoTLASS (Jolliffe et al., 2003), SPCA (Zou et al., 2006) and sPCA-rSVD (Shen & Huang, 2008), a number of power methods (Journée et al., 2010; Yuan & Zhang, 2011; Ma, 2011), the biconvex algorithm PMD (Witten et al., 2009) and the semidefinite relaxation DSPCA (d'Aspremont et al., 2004).

For data dependence, in low dimensions, Skinner et al. (1986) study the behavior of PCA when the selection

of a sample of observations depends on a vector of latent covariates as, for example, in stratified sampling. Their analysis is based on the normality assumption and that the knowledge of the survey design is known. There are even fewer literatures in high dimensions for the dependent data. Loh & Wainwright (2011) study the high dimensional regression for Gaussian random vectors following a stationary vector regressive process. Very recently, Fan et al. (2012) analyze the penalized least square estimators, taking a weakly dependence structure, called α-mixing, of the noisy term into consideration.

There are several drawbacks of PCA (or sparse PCA): (i) It is not scale-invariant, i.e., changing the measurement scale of variables makes the estimates different; (ii) Most estimating procedures require the data to be either Gaussian or sub-Gaussian so that the sample covariance matrix $\mathbf{S}$ converges to $\mathbf{\Sigma}$ in a fast rate; (iii) It cannot handle dependent data. Compared with PCA and sparse PCA, Han & Liu (2012) exploit a nonparametric Kendall's tau based regularization procedure, named Copula Component Analysis (COCA), for parameter estimation. They show that COCA is scale-invariant, able to deal with continuous data with arbitrary margins. In this paper, we further generalize their results, showing that COCA can allow weak data dependence. In particular, we provide the generalization bounds of convergence for both support recovery and parameter estimation for dependent data using our method. We provide explicit sufficient conditions on the the degree of dependence, under which the same parametric rate can be achieved. To our knowledge, this is the first work of analyzing the theoretical performance of PCA for dependent data in high dimensional settings.

The rest of the paper is organized as follows. In the next section, we briefly review the nonparanormal family (Liu et al., 2009; 2012) and the data dependence structure. In Section 3, we introduce the models and rank-based estimators proposed by Han & Liu (2012). We provide our main theoretical analysis of the rank-based estimators for the dependent data in Section 4. In Section 5, we employ the on synthetic data to illustrate the robustness of COCA to data dependence.

2. Background

We start with notations: Let $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$. Let $\boldsymbol{v}$'s subvector with entries indexed by I be denoted by $\boldsymbol{v}_I$. Let $\mathbf{M}$'s submatrix with rows indexed by I and columns indexed by 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 $0 < q < \infty$, we define the ℓ_q and ℓ_∞ vector norms as

$$||\boldsymbol{v}||_q := \left(\sum_{i=1}^d |v_i|^q\right)^{1/q} \quad \text{and} \quad ||\boldsymbol{v}||_\infty := \max_{1\le i\le d} |v_i|,$$

and we define $||\boldsymbol{v}||_0 := \text{card}(\text{supp}(\boldsymbol{v}))\cdot||v||_2$. We define the matrix $\ell_{\max}$ norm as the elementwise maximum value: $||\mathbf{M}||_{\max} := \max\{|M_{ij}|\}$. Let $\Lambda_j(\mathbf{M})$ be the j-th largest eigenvalue of $\mathbf{M}$. In particular,

$$\Lambda_{\min}(\mathbf{M}) := \Lambda_d(\mathbf{M}) \text{ and } \Lambda_{\max}(\mathbf{M}) := \Lambda_1(\mathbf{M})$$

are the smallest and largest eigenvalues of $\mathbf{M}$. The vectorized matrix of $\mathbf{M}$, denoted by $\text{vec}(\mathbf{M})$, is defined as:

$$\text{vec}(\mathbf{M}) := (\mathbf{M}_{*1}^T, \ldots, \mathbf{M}_{*d}^T)^T.$$

Let $\mathbb{S}^{d-1} := \{\boldsymbol{v} \in \mathbb{R}^d : ||\boldsymbol{v}||_2 = 1\}$ be the d-dimensional ℓ_2 sphere. For any two vectors $\boldsymbol{a}, \mathbf{b} \in \mathbb{R}^d$ and any two squared matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, denote the inner product of $\boldsymbol{a}$ and $\mathbf{b}$, $\mathbf{A}$ and $\mathbf{B}$ by

$$\langle \boldsymbol{a}, \mathbf{b}\rangle := \boldsymbol{a}^T\mathbf{b} \quad \text{and} \quad \langle \mathbf{A}, \mathbf{B}\rangle := \text{Tr}(\mathbf{A}^T\mathbf{B}).$$

The sign $\overset{d}{=}$ denotes that the two sides of the equality have the same distributions. For a sequence of random vector $\{\boldsymbol{X}_t\}_{t=-\infty}^{\infty}$ and two integers $L < U$, we denote $\boldsymbol{X}_L^U := \{\boldsymbol{X}_t\}_{t=L}^U$. The notation $\mathbb{P}(\cdot)$ represents the probability if it is a set inside the brackets and the law (distribution) if it is a random vector inside the brackets.

2.1. The Nonparanormal

We first provide the definition of the nonparanormal following Liu et al. (2012).

Definition 2.1 (The nonparanormal). *Let $f = \{f_j\}_{j=1}^d$ be a set of monotone univariate functions. We say that a d dimensional random variable $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ follows a nonparanormal distribution $NPN_d(\mathbf{\Sigma}, f)$, if*

$$f(\boldsymbol{X}) := (f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(\mathbf{0}, \mathbf{\Sigma}), \ \text{diag}(\mathbf{\Sigma}) = \mathbf{1}.$$

We call $\mathbf{\Sigma}$ the latent correlation matrix.

We next proceed to the invariance property of the rank-based estimator Kendall's tau in the nonparanormal family. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ be a random vector. Let $\widetilde{X}_j$ and $\widetilde{X}_k$ be two independent copies of X_j and X_k. The population version of the Kendall's tau statistic is:

$$\tau(X_j, X_k) := \text{Corr}\left(\text{sign}(X_j - \widetilde{X}_j), \text{sign}(X_k - \widetilde{X}_k)\right).$$

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be n observed data points. The sample version Kendall's tau statistic is defined as:

$$\widehat{\tau}_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) := \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \operatorname{sign}(x_{ij} - x_{i'j}) \cdot (x_{ik} - x_{i'k}), \quad (2.1)$$

which is monotone transformation-invariant correlation between the empirical realizations of two random variables X_j and X_k. For $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ independent, it is easy to verify that $\mathbb{E}\widehat{\tau}_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \tau(X_j, X_k)$. We denote $\widehat{\mathbf{R}} = [\widehat{R}_{jk}] \in \mathbb{R}^{d \times d}$ with

$$\widehat{R}_{jk} = \sin(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))$$

the Kendall's tau matrix.

Another interpretation of the Kendall's tau statistic is that it is an association measure based on the notion of concordance. We call two pairs of real numbers (s,t) and (u,v) concordant if $(s-t)(u-v) > 0$ and disconcordant if $(s-t)(u-v) < 0$. Kruskal (1958) show that

$$\tau(X_j, X_k) = \mathbb{P}\left((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0\right) - \mathbb{P}\left((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) < 0\right). \quad (2.2)$$

The following theorem, coming from Kruskal (1958), states the invariance property of the relationship between the population Kendall's tau statistic $\tau(X_j, X_k)$ and the latent correlation coefficient $\boldsymbol{\Sigma}_{jk}$ in the nonparanormal family.

Theorem 2.2. *Let $\boldsymbol{X} := (X_1, \ldots, X_d)^T \sim NPN_d(\boldsymbol{\Sigma}, f)$. We denote $\tau(X_j, X_k)$ to be the population Kendall's tau statistic between X_j and X_k. Then $\boldsymbol{\Sigma}_{jk} = \sin\left(\frac{\pi}{2}\tau(X_j, X_k)\right)$.*

Proof. To prove Theorem 2.2, we actually have

$$\begin{aligned}
\tau(X_j, X_k) &= \mathbb{P}\left((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0\right) \\
&\quad - \mathbb{P}\left((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) < 0\right) \\
&= \mathbb{P}\left((f_j(X_j) - f_j(\widetilde{X}_j))(f_k(X_k) - f_k(\widetilde{X}_k)) > 0\right) \\
&\quad - \mathbb{P}\left((f_j(X_j) - f_j(\widetilde{X}_j))(f_k(X_k) - f_k(\widetilde{X}_k)) < 0\right) \\
&= \frac{2}{\pi}\arcsin(\boldsymbol{\Sigma}_{jk}).
\end{aligned}$$

The last equality is coming from Kruskal (1958)'s result for Gaussian distribution. □

2.2. Mixing Conditions

In this section we provide definitions of several models of non-independent data. In particular, we will introduce the notions of strong mixing conditions: ϕ-mixing and η-mixing. These will be utilized later in analyzing the performance of the proposed method for the dependent data. We first introduce the stationary m-dependence sequences as follows.

Definition 2.3 (m-dependence). *A stationary sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is said to be m-dependence if and only if (i) $\boldsymbol{X}_i \overset{d}{=} \boldsymbol{X}$ for $i \in \{1, \ldots, n\}$ and some random vector $\boldsymbol{X} \in \mathbb{R}^d$; (ii) For any $s, t \in \{1, \ldots, n\}$, $\boldsymbol{X}_s$ is independent of $\boldsymbol{X}_t$ whenever $|s-t| > m$.*

m-dependency is of particular interest in several fields. For example, in genetics and epidemiology, data might contain samples of families and there are correlation between members of the same family. Moreover, m-dependence can be seen as a simplified version of a time series, where $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ will often be highly dependent if $|s-t|$ is small, with decreasing dependence as $|s-t|$ increases.

Next we proceed to some more general weak dependency conditions. In particular, we build the dependence structure on the mixing sequences. To this end, we first introduce the ϕ measure of dependence as follows.

Definition 2.4 (ϕ measure of dependence). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space and $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ be two σ-fields. We define the ϕ measures of dependence as:*

$$\phi(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A}, B \in \mathcal{B}, \mathbb{P}(A) > 0} |\mathbb{P}(B \mid A) - \mathbb{P}(B)|. \quad (2.3)$$

We then describe the strong mixing conditions. Let $\boldsymbol{X} = \{\boldsymbol{X}_t\}_{t=-\infty}^{\infty}$ be a sequence of random vectors. For $-\infty \le L \le U \le \infty$, define the σ-field $\mathcal{F}_L^U$ to be $\mathcal{F}_L^U := \sigma(\boldsymbol{X}_i, L \le i \le U)$. For two probability measures μ_1, μ_2 on the measurable space $(\Omega, \mathcal{F})$, the total variation distance between μ_1 and μ_2 is defined as:

$$||\mu_1 - \mu_2||_{TV} := \sup_{A \in \mathcal{F}} |\mu_1(A) - \mu_2(A)|. \quad (2.4)$$

With the above notations, the ϕ and η dependence coefficients are defined as:

Definition 2.5. *Let $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=0}^{\infty}$ be a sequence of random vectors defined in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\boldsymbol{X}_i^j$ be the subsequence. For $0 \le L \le U \le \infty$, remind that $\mathcal{F}_L^U := \sigma(\boldsymbol{X}_i, L \le i \le U)$. The ϕ*

and η dependence coefficients are defined as:

$$\phi(m) := \sup_{j\in\mathbb{Z}} \phi(\mathcal{F}_0^j, \mathcal{F}_{j+m}^\infty);$$
$$\eta_{ij} := \operatorname*{ess\ sup}_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{x}'} ||\mathbb{P}(\boldsymbol{X}_j^n \mid \boldsymbol{X}_1^{i-1} = \boldsymbol{y}, \boldsymbol{X}_i = \boldsymbol{x}) - \mathbb{P}(\boldsymbol{X}_j^n \mid \boldsymbol{X}_1^{i-1} = \boldsymbol{y}, \boldsymbol{X}_i = \boldsymbol{x}')||_{TV}.$$

The following lemma connects η_{ij} to $\phi(m)$. This is coming from Kontorovich & Ramanan (2008) for discrete case and a stronger version applicable to continuous case can be traced back to Samson (2000). This lemma will play an important role later in analyzing our proposed method. For self-containedness, we include a proof here.

Lemma 2.6 (Samson (2000)). *With the above notations, we have $\eta_{ij} \leq 2\phi(j-i)$.*

Proof. By definition, we have

$$\begin{aligned}
&||\mathbb{P}(\boldsymbol{X}_j^n|\boldsymbol{X}_1^{i-1}=\boldsymbol{y},\boldsymbol{X}_i = \boldsymbol{x})-\mathbb{P}(\boldsymbol{X}_j^n|\boldsymbol{X}_1^{i-1}=\boldsymbol{y},\boldsymbol{X}_i = \boldsymbol{x}')||_{TV}\\
&\leq ||\mathbb{P}(\boldsymbol{X}_j^n \mid \boldsymbol{X}_1^{i-1} = \boldsymbol{y}, \boldsymbol{X}_i = \boldsymbol{x}) - \mathbb{P}(\boldsymbol{X}_j^n)||_{TV}\\
&+ ||\mathbb{P}(\boldsymbol{X}_j^n \mid \boldsymbol{X}_1^{i-1} = \boldsymbol{y}, \boldsymbol{X}_i = \boldsymbol{x}') - \mathbb{P}(\boldsymbol{X}_j^n)||_{TV}\\
&\leq 2\phi(j-i).
\end{aligned}$$

Therefore, by the continuity of the probability, we have $\eta_{ij} \leq 2\phi(j-i)$. This completes the proof. □

3. Methods

In this section, we briefly review the statistical models of Copula Component Analysis (COCA) proposed by Han & Liu (2012). The aim of COCA is to recover the principal components of the latent Gaussian distribution in the nonparanormal family.

3.1. Scale-invariant PCA

PCA is not scale invariant, meaning that variables measured in different scales will result in different estimators (Jolliffe, 2005). To attack this problem, PCA conducted on the sample correlation matrix $\mathbf{S}^0$ instead of the sample covariance matrix $\mathbf{S}$ is commonly used. We call the procedure of conducting PCA on $\mathbf{S}^0$ the scale-invariant PCA. It is realized that a large portion of works claiming doing PCA are actually doing the scale-invariant PCA. It is under debate whether PCA or the scale-invariant PCA are preferred in different circumstances and we refer to Jolliffe (2005) for more discussions on it. In the population level, the scale-invariant PCA aims at recovering the leading eigenvectors of the correlation matrix, which has the same sparsity pattern as the leading eigenvectors of the covariance matrix.

3.2. Models

One of the intuition of PCA is coming from the Gaussian distribution. In geometric, the principal components define the major axes of the contours of constant probability for the multivariate Gaussian (Jolliffe, 2005). However, such a nice interpretation does not exist anymore when the distributions are away from the Gaussian. Balasubramanian & Schwartz (2002) construct examples such that PCA loses in the sense of preserving the structure of the data to the most. However, under the nonparanormal model and considering the monotone transformation f as a type of data contamination, the geometric intuition of PCA comes back.

In particular, for a positive definite matrix $\boldsymbol{\Sigma}$ with $\operatorname{diag}(\boldsymbol{\Sigma}) = \mathbf{1}$, let $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_d$ be eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d$ be the corresponding eigenvectors. For $0 \leq q \leq 1$, the ℓ_q ball $\mathbb{B}_q(R_q)$ is defined as:

$$\text{when } q = 0,\ \mathbb{B}_0(R_0) := \{\boldsymbol{v} \in \mathbb{R}^d : \operatorname{card}(\operatorname{supp}(\boldsymbol{v})) \leq R_0\};$$
$$\text{when } 0 < q \leq 1,\ \mathbb{B}_q(R_q) := \{\boldsymbol{v} \in \mathbb{R}^d : ||\boldsymbol{v}||_q^q \leq R_q\}.$$

Accordingly, the model $\mathcal{M}(q, R_q, \boldsymbol{\Sigma}, f)$ is considered:

$$\mathcal{M}(q, R_q, \boldsymbol{\Sigma}, f) = \big\{\boldsymbol{X} : \boldsymbol{X} \sim NPN_d(\boldsymbol{\Sigma}, f), \boldsymbol{\theta}_1 \in \mathbb{S}^{d-1} \cap \mathbb{B}_q(R_q)\big\}. \quad (3.1)$$

The ℓ_q ball induces a (weak) sparsity pattern when $0 \leq q \leq 1$ and has been analyzed in linear regression (Raskutti et al., 2011) and sparse PCA (Vu & Lei, 2012). Moreover, the data are assumed to come from a nonparanormal distribution, which is a strict extension to the Gaussian distribution.

Inspired by the model $\mathcal{M}(q, R_q, \boldsymbol{\Sigma}, f)$, we consider the following global estimator $\widetilde{\boldsymbol{\theta}}_1$, which maximizes the following equation with the constraint that $\widetilde{\boldsymbol{\theta}}_1 \in \mathbb{B}_q(R_q)$ for some $0 \leq q \leq 1$:

$$\widetilde{\boldsymbol{\theta}}_1 = \operatorname*{argmax}_{\boldsymbol{v}\in\mathbb{R}^d} \boldsymbol{v}^T \widehat{\mathbf{R}} \boldsymbol{v}, \quad \text{subject to} \quad \boldsymbol{v} \in \mathbb{S}^{d-1} \cap \mathbb{B}_q(R_q). \quad (3.2)$$

Here $\widehat{\mathbf{R}}$ is the estimated Kendall's tau matrix. The corresponding estimator $\widetilde{\boldsymbol{\theta}}_1$ can be considered as a nonlinear dimensional reduction procedure and has the potential to gain more flexibility compared with PCA, as shown in the analysis of Han & Liu (2012).

4. Theoretical Properties

In this section we provide the theoretical properties of the proposed COCA estimator $\widetilde{\boldsymbol{\theta}}_1$ as obtained in Equation (3.2) for the dependent data. To our knowledge,

this is the first work analyzing the theoretical performance of PCA for the dependent data in high dimensions. To provide some insights, we first deliver the rate of $\widetilde{\boldsymbol{\theta}}_1$ converging to $\boldsymbol{\theta}_1$ when the data points are independent from each other. This theorem is coming from Han & Liu (2012).

Theorem 4.1 (Independence). *Let $\widetilde{\boldsymbol{\theta}}_1$ be the global optimum in Equation (3.2) and $\boldsymbol{X} \in \mathcal{M}(q, R_q, \boldsymbol{\Sigma}, f)$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be an independent sequence of realizations of $\boldsymbol{X}$ and $\widehat{\mathbf{R}} := [\sin(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))]$ be defined as in Equation (2.1). 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 $1 - 1/d^2$,*

$$\sin^2\angle(\widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1) \leq \gamma_q R_q^2 \left(\frac{2\pi^2}{(\lambda_1-\lambda_2)^2} \cdot \frac{\log d}{n}\right)^{\frac{2-q}{2}}, \quad (4.1)$$

where $\gamma_q = 2 \cdot I(q=1) + 4 \cdot I(q=0) + (1+\sqrt{3})^2 \cdot I(0 < q < 1)$ and $\lambda_j = \Lambda_j(\boldsymbol{\Sigma})$ for $j = 1, 2$.

Proof. The key idea of the proof is to utilize the $\ell_{\max}$ norm convergence result of $\widehat{\mathbf{R}}$ to $\boldsymbol{\Sigma}$. See Han & Liu (2012) for a detailed proof. □

It can be observed that the convergence rate of $\widetilde{\boldsymbol{\theta}}_1$ to θ_1 will be faster when θ_1 lies in a more sparse ball. It makes sense because the effect of "the curse of dimensionality" will be decreasing when the parameters are more and more sparse. Generally, when R_q and λ_1, λ_2 do not scale with (n, d), the rate is $O_P\left((\frac{\log d}{n})^{1-q/2}\right)$, which is the parametric rate Vu & Lei (2012) obtain. In the following, we show that Theorem 4.1 can be applied to derive a support recovery result.

Corollary 4.2 (Independence). *With the settings and notations in Theorem 4.1 held, let*

$$\Theta := \mathit{supp}(\boldsymbol{\theta}_1) \text{ and } \widehat{\Theta} := \mathit{supp}(\widetilde{\boldsymbol{\theta}}_1).$$

If we further have

$$\min_{j\in\Theta} |\theta_{1j}| \geq \frac{2\sqrt{2}R_0\pi}{\lambda_1-\lambda_2}\sqrt{\frac{\log d}{n}},$$

then we have $\mathbb{P}(\widehat{\Theta} = \Theta) \geq 1 - d^{-2}$.

We then generalize Theorem 4.1 to the non-independent cases. Here the notions of m-dependence and ϕ-mixing sequences as defined in Section 2.2 are exploited. We first provide an upper bound for the estimator $\widetilde{\boldsymbol{\theta}}_1$ when the data points form an m-dependence sequence.

Theorem 4.3 (m-dependence). *Let $\boldsymbol{X} \in \mathcal{M}(q, R_q, \boldsymbol{\Sigma}, f)$ and $\{\boldsymbol{X}_t\}_{t=1}^n$ be a m-dependence stationary sequence with $\boldsymbol{X}_t \overset{d}{=} \boldsymbol{X}$. Let $\widetilde{\boldsymbol{\theta}}_1$ be the global optimum in Equation (3.2), where $\widehat{\mathbf{R}} := [\sin(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n))]$ is defined as in Equation (2.1). Let the parameter*

$$\gamma_{m,n} \coloneqq 2(m+1)^2(1-m/n) + m(m+1)(2m+1)/(3n)$$

represent the effect of dependence on the rate of convergence. Then we have, for any $n \geq 4m^2/(\gamma_{m,n}\log d)$, with probability at least $1 - 1/d^2$

$$\sin^2\angle(\widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1) \leq \gamma_q R_q^2 \left(\frac{4\pi^2\gamma_{m,n}}{(\lambda_1-\lambda_2)^2} \cdot \frac{\log d}{n}\right)^{\frac{2-q}{2}}, \quad (4.2)$$

where $\gamma_q = 2 \cdot I(q=1) + 4 \cdot I(q=0) + (1+\sqrt{3})^2 \cdot I(0 < q < 1)$ and $\lambda_j = \Lambda_j(\boldsymbol{\Sigma})$ for $j = 1, 2$.

Proof. The key is to estimate the convergence rate of $\widehat{\mathbf{R}}$ to $\boldsymbol{\Sigma}$ in the m-dependence setting. The detailed proof is presented in Han & Liu (2013). □

Remark 4.4. *It can be observed in Equation (4.2) that $\widetilde{\boldsymbol{\theta}}_1$ converges to $\boldsymbol{\theta}_1$ in a rate related to both (n, d) and m. Generally, when R_q and λ_1, λ_2 do not scale with (n, d), the rate is $O_P\left((\frac{m^2\log d}{n})^{1-q/2}\right)$. When m is fixed, the rate is optimal.*

Using Theorem 4.3, we provide the support recovery result using a similar technique as the proof of Corollary 4.2.

Corollary 4.5 (m-dependency). *With the settings and notations in Theorem 4.3 held, let*

$$\Theta := \mathit{supp}(\boldsymbol{\theta}_1) \text{ and } \widehat{\Theta} := \mathit{supp}(\widetilde{\boldsymbol{\theta}}_1)$$

. If we further have

$$\min_{j\in\Theta} |\theta_{1j}| \geq \frac{4R_0\pi\gamma_{m,n}}{\lambda_1-\lambda_2}\sqrt{\frac{\log d}{n}},$$

then we have $\mathbb{P}(\widehat{\Theta} = \Theta) \geq 1 - d^{-2}$.

We next proceed to bound the angle between $\widetilde{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_1$ in a more general setting of data dependence. Here the dependence is quantified by ϕ measure as defined in Definitions 2.4 and 2.5.

Theorem 4.6 (ϕ-mixing). *Let $\boldsymbol{X} \in \mathcal{M}(q, R_q, \boldsymbol{\Sigma}, f)$ and $\{\boldsymbol{X}_t\}_{t=1}^n$ be a stationary sequence with $\boldsymbol{X}_t \overset{d}{=} \boldsymbol{X}$. We assume that $\{\boldsymbol{X}_t\}_{t=1}^n$ satisfies that for any $j \neq k \in \{1, \ldots, d\}$ and $m \in \mathbb{N}$,*

$$\sup_{i\in\mathbb{Z}} \phi\big(\sigma((\boldsymbol{X}_1)_{\{j,k\}}, \ldots, (\boldsymbol{X}_i)_{\{j,k\}}), \sigma((\boldsymbol{X}_{i+m})_{\{j,k\}}, \ldots, (\boldsymbol{X}_n)_{\{j,k\}})\big) \leq \phi(m).$$

Here $\{\phi(i)\}_{i=1}^{n-1}$ *is a sequence of positive numbers. Let* $\widetilde{\boldsymbol{\theta}}_1$ *be the global optimum in Equation* (3.2)*, where* $\widehat{\mathbf{R}} := [\sin(\frac{\pi}{2}\widehat{\tau}_{jk}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n))]$ *is defined as in Equation* (2.1)*. Let*

$$\gamma_\phi := \left(1 + 2\sum_{i=1}^{n-1}\phi(i)\right)^2 \tag{4.3}$$

represent the effect of dependence on the rate of convergence. Then we have, supposing

$$n \geq \frac{\left[\sum_{i=1}^{n-1}\left(\frac{n-i}{n-1}\phi(i)\right)\right]^2}{\gamma_\phi \log d},$$

with probability at least $1 - 1/d^2$*,*

$$\sin^2\angle(\widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1) \leq \gamma_q R_q^2 \left(\frac{16\pi^2\gamma_\phi}{(\lambda_1 - \lambda_2)^2} \cdot \frac{\log d}{n}\right)^{\frac{2-q}{2}}, \tag{4.4}$$

where $\gamma_q = 2 \cdot I(q = 1) + 4 \cdot I(q = 0) + (1 + \sqrt{3})^2 \cdot I(0 < q < 1)$ *and* $\lambda_j = \Lambda_j(\boldsymbol{\Sigma})$ *for* $j = 1, 2$*.*

Proof. The main proof is to quantify the bias term, and then utilize Lemma 2.6 to build a bridge between the results in Section 5.1 in Kontorovich (2007) and the desired concentration inequality we need. Detailed proof is presented in Han & Liu (2013). □

Remark 4.7. *Sequences satisfying m-dependence is a subset of* ϕ*-mixing sequences. However, Theorem 4.3 provides a faster convergence rate than the result in Theorem 4.6. Moreover, the proof techniques in Theorem 4.3 has interesting points itself. It can be observed that when* $\phi(i)$ *decreases fast (i.e., weak-dependence), the rate is near-optimal. For example, supposing that* R_q *and* λ_1, λ_2 *do not scale with* (n, d)*, when* $\phi(i) = O(i^{-2})$*, the rate is* $O_P\left(\left(\frac{\log d}{n}\right)^{1-q/2}\right)$*, which is optimal. When* $\phi(i) = O(i^{-1})$*, the rate is* $O_P\left(\left(\frac{\log^2 n \log d}{n}\right)^{1-q/2}\right)$*.*

Again, a support recovery result can be provided using a similar technique as the proof of Corollaries 4.2 and 4.5.

Corollary 4.8 (ϕ-mixing). *With the settings and notations in Theorem 4.3 held, let*

$$\Theta := \mathit{supp}(\boldsymbol{\theta}_1) \text{ and } \widehat{\Theta} := \mathit{supp}(\widetilde{\boldsymbol{\theta}}_1)$$

. If we further have

$$\min_{j \in \Theta} |\theta_{1j}| \geq \frac{8R_0\pi\gamma_\phi}{\lambda_1 - \lambda_2}\sqrt{\frac{\log d}{n}},$$

then we have $\mathbb{P}(\widehat{\Theta} = \Theta) \geq 1 - d^{-2}$*.*

5. Experiments

In this section we investigate the robustness of COCA to data dependence on the synthetic data. We use the truncated power method proposed by Yuan & Zhang (2011) to approximate the global estimator $\widetilde{\boldsymbol{\theta}}_1$ obtained in Equation (3.2). Two procedures are considered here:

Pearson: the classic sparse PCA using the Pearson sample correlation matrix;

Kendall: the proposed rank-based scale-invariant PCA method using the Kendall's tau correlation matrix.

In the simulation study we sample n data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a certain random vector $\boldsymbol{X} \in \mathbb{R}^d$ with some type of data dependence. Here we set $d = 100$. We follow a similar generating scheme as in Han & Liu (2012). A covariance matrix $\boldsymbol{\Sigma}$ is firstly synthesized through the eigenvalue decomposition, where the first two eigenvalues are given and the corresponding eigenvectors are pre-specified to be sparse.

In detail, we let $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_d$ be the d eigenvalues of $\boldsymbol{\Sigma}$ and $u_1, u_2, \ldots, u_d$ be the corresponding eigenvectors. Suppose that the first two dominant eigenvectors of $\boldsymbol{\Sigma}$, u_1 and u_2, are sparse in the sense that only the first $s = 10$ entries of u_1 and the second $s = 10$ entries of u_2 nonzero, i.e.,

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

and $\omega_1 = 5$, $\omega_2 = 2$, $\omega_3 = \ldots = \omega_d = 1$. The remaining eigenvectors are chosen arbitrarily. The correlation matrix $\boldsymbol{\Sigma}^0$ is accordingly generated and the leading eigenvector of $\boldsymbol{\Sigma}^0$ is sparse. We aim at recovering the leading eigenvector $\boldsymbol{\theta}_1$.

To sample data from the nonparanormal, we also need the transformation functions: $f = \{f_j\}_{j=1}^d$. Here the following transformation function is considered: There exist five univariate monotone functions $h_1, h_2, \ldots, h_5 : \mathbb{R} \to \mathbb{R}$ and

$$h = \{h_1, h_2, h_3, h_4, h_5, h_1, h_2, h_3, h_4, h_5, \ldots\},$$

where

$$h_1(x) := x,\ h_2(x) := \mathrm{sign}(x)|x|^{1/2}, h_3(x) := x^3,$$

$$h_4(x) := \Phi(x), h_5(x) := \exp(x).$$

Here Φ is defined to be the cumulative distribution functions of the standard Gaussian. We then generate $n = 100$ data points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ such that $\boldsymbol{y}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is defined as above. To evaluate the robustness of different methods for dependent data, we

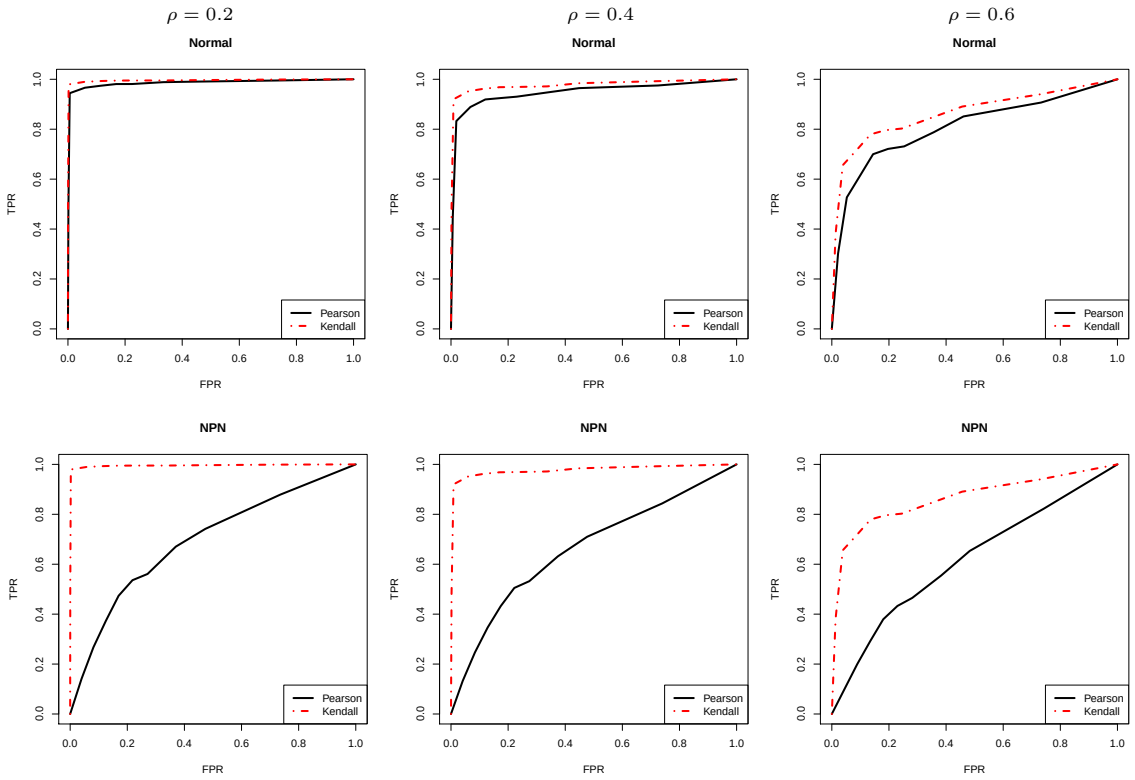

Figure 1. ROC curves for the Gaussian (Scheme 1) and non-Gaussian (Scheme 2) data (above and below) using the truncated power algorithm are presented. Here the data dependence degrees are at different levels ($\rho = 0.2, 0.4, 0.6$). n is fixed to be 100 and $d = 100$.

suppose that $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ follow a stationary vector autoregressive process as defined in Loh & Wainwright (2011). In detail, we assume that $\boldsymbol{z}_1 = \boldsymbol{y}_1$ and for some real number $0 \leq \rho \leq 1$

$$\boldsymbol{z}_{t+1} = \rho \cdot \boldsymbol{z}_t + \sqrt{1-\rho^2}\boldsymbol{y}_{t+1}, \quad \text{for } t = 1, \ldots, n-1.$$

Here we have that $\boldsymbol{z}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ forms a dependent random sequence. Finally, we have the data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$:
[Scheme 1] $\{\boldsymbol{x}_i\}_{i=1}^n = \{\boldsymbol{z}_i\}_{i=1}^n$, with $\boldsymbol{x}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$;
[Scheme 2] $\{\boldsymbol{x}_i\}_{i=1}^n = \{h(\boldsymbol{z}_i)\}_{i=1}^n$ where $h := \{h_1, h_2, h_3, h_4, h_5, \ldots\}$, with $\boldsymbol{x}_i$ follows a non-Gaussian nonparanormal distribution.
The final data matrix we obtained is $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times d}$. The truncated power algorithm is then employed on $\mathbf{X}$ to computer the estimated leading eigenvector $\widetilde{\boldsymbol{\theta}}_1$.

To evaluate the empirical variable selection property of different methods, we define

$$\mathcal{S} := \{1 \leq j \leq d : \theta_{1j} \neq 0\}, \tag{5.1}$$
$$\widehat{\mathcal{S}}_\delta := \{1 \leq j \leq d : \widetilde{\boldsymbol{\theta}}_{1j} \neq 0\}, \tag{5.2}$$

to be the support sets of the true leading eigenvector θ_1 and the estimated leading eigenvector $\widetilde{\boldsymbol{\theta}}_1$ using the tuning parameter δ. In this way, the False Positive Number (FPN) and False Negative Number (FNN) of δ are defined as:

$$\begin{aligned} \text{FPN}(\delta) &:= \text{the number of features in } \widehat{\mathcal{S}}_\delta \text{ not in } \mathcal{S}, \\ \text{FNN}(\delta) &:= \text{the number of features in } \mathcal{S} \text{ not in } \widehat{\mathcal{S}}_\delta. \end{aligned}$$

Then we can further define the False Positive Rate(FPR) and False Negative Rate (FNR) corresponding to the tuning parameter δ to be

$$\text{FPR}(\delta) := \text{FPN}(\delta)/(d-s), \text{FNR}(\delta) := \text{FNN}(\delta)/s.$$

Under the Scheme 1 and Scheme 2 with different levels of dependence ($\rho = 0, 0.2, 0.4, 0.6, 0.8$), we repeatedly generate the data matrix $\boldsymbol{X}$ for 1,000 times and compute the averaged False Positive Rates and False Negative Rates using a path of tuning parameters δ. The feature selection performances of different methods are then evaluated by plotting $(\text{FPR}(\delta), 1-\text{FNR}(\delta))$. The corresponding ROC curves are presented in Figure 1.

There are several observations we can see from Figure 1: (i) With the increase of the data dependence level, both methods' performance decreases. (ii) Compared with the Gaussian case (Scheme 1), the difference between Pearson and Kendall are larger when the data

Table 1. Quantitative comparison on the dataset under the generating Scheme 1 and Scheme 2. The means of the oracle false positive and false negative rates with their standard deviations in parentheses are presented. Here $n = 100$, $d = 100$ and the dependence degree ρ is increasing from 0 to 0.8.

	Gaussian(Scheme 1)				non-Gaussian (Scheme 2)			
	Pearson		Kendall		Pearson		Kendall	
ρ	FPR(%)	FNR	FPR	FNR	FPR	FNR	FPR	FNR
0.0	1.1(0.80)	0.0(0.00)	1.2(0.52)	0.0(0.00)	17.2(3.60)	7.3(3.72)	1.2(0.52)	0.0(0.00)
0.2	1.8(0.89)	0.0(0.14)	1.3(0.69)	0.0(0.00)	17.4(3.42)	7.6(3.66)	1.3(0.69)	0.0(0.00)
0.4	4.4(1.30)	0.2(0.47)	2.7(1.06)	0.0(0.20)	18.5(3.56)	10.5(4.52)	2.7(1.06)	0.0(0.20)
0.6	10.3(2.19)	2.8(1.77)	8.0(1.92)	1.7(1.39)	20.8(4.40)	16.9(5.70)	8.0(1.92)	1.7(1.39)
0.8	20.2(4.30)	20.8(5.87)	18.8(4.17)	18.7(5.52)	24.4(4.93)	27.7(6.38)	18.8(4.17)	18.7(5.52)

are generated from Scheme 2. This coincides with the observations in Han & Liu (2012). (iii) When the data dependence degree ρ increases, Kendall performs better than Pearson in both the Gaussian and Nonparanormal cases, meaning that Kendall is more robust to the data dependence than Pearson.

To explore the empirical performances of difference methods using the truncated power method more, we define an oracle tuning parameter δ^* to be the δ with the lowest $\mathrm{FPR}(\delta) + \mathrm{FNR}(\delta)$:

$$\delta^* := \underset{\delta}{\operatorname{argmin}}(\ \mathrm{FPR}(\delta) + \mathrm{FNR}(\delta)\ ). \tag{5.3}$$

In this way, an estimator $\widetilde{\boldsymbol{\theta}}_1$ using the oracle tuning parameter δ^* can be calculated and we compute the oracle false positive and false negative rates as:

$$\mathrm{FPR}^* = \mathrm{FPR}(\delta^*) \ \text{ and } \ \mathrm{FNR}^* = \mathrm{FNR}(\delta^*). \tag{5.4}$$

We present the means and standard deviations of $(\mathrm{FPR}^*, \mathrm{FNR}^*)$ in Table 1.

There are several observations we can see from Table 1: (i) When ρ is increasing, both methods' oracle positive and negative rates decrease. (ii) In the perfect Gaussian case (Scheme 1) where the data points are independent of each other ($\rho = 0$), there is no statistically significant difference between Kendall and Pearson. (iii) There exist statistically significant differences between Kendall and Pearson in Scheme 2, no matter how large the degree of data dependence (ρ) is. (iv) There is a statistically significant difference between Pearson and Kendall for the Gaussian case when $\rho = 0.4$, and Kendall performs constantly better than Pearson when $\rho > 0$. In all, Kendall is more robust to the data dependence than Pearson.

6. Conclusion

In this paper we analyze both theoretical and empirical performance of a newly proposed high dimensional semiparametric principal component analysis, named Copula Component Analysis (COCA), when the data are dependent. We provide explicit upper bounds of convergence for COCA estimators when the observations are drawn from several different types of non-i.i.d. processes. Our results show that COCA can allow weak dependence. To our knowledge, this is the first work analyzing the theoretical performance of PCA for the dependent data in high dimensions. Our result strictly generalize the analysis in Han & Liu (2012) and the techniques we used have the separate interest for analyzing a variety of other multivariate statistical methods.

7. Acknowledgement

This research was supported by NSF award IIS-1116730.

References

Balasubramanian, M. and Schwartz, E.L. The isomap algorithm and topological stability. *Science*, 295 (5552):7–7, 2002.

d'Aspremont, A., El Ghaoui, L., Jordan, M.I., and Lanckriet, G.R.G. *A direct formulation for sparse PCA using semidefinite programming*. Computer Science Division, University of California, 2004.

d'Aspremont, A., Bach, F., and Ghaoui, L.E. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.

Fan, J., Qi, L., and Tong, X. Penalized least squares estimation with weakly dependent data. 2012.

Han, F. and Liu, H. Semiparametric principal component analysis. In *Advances in Neural Information Processing Systems 25*, pp. 171–179, 2012.

Han, F. and Liu, H. Principal component analysis on high dimensional complex and noisy data. *Technical Report*, 2013.

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

Jolliffe, I. *Principal component analysis*. Wiley Online Library, 2005.

Jolliffe, I.T., Trendafilov, N.T., and Uddin, M. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

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

Kontorovich, L. *Measure concentration of strongly mixing processes with applications*. PhD thesis, Carnegie Mellon University, 2007.

Kontorovich, L.A. and Ramanan, K. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36 (6):2126–2158, 2008.

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

Liu, H., Lafferty, J., and Wasserman, L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. High dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*, 2012.

Loh, P.L. and Wainwright, M.J. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Arxiv preprint arXiv:1109.3714*, 2011.

Ma, Z. Sparse principal component analysis and iterative thresholding. *Arxiv preprint arXiv:1112.2432*, 2011.

McDiarmid, C. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Raskutti, G., Wainwright, M.J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over *ell_q*-balls. *Information Theory, IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

Samson, P.M. Concentration of measure inequalities for markov chains and *phi*-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

Shen, H. and Huang, J.Z. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008.

Skinner, CJ, Holmes, DJ, and Smith, TMF. The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81 (395):789–798, 1986.

Vu, V.Q. and Lei, J. Minimax rates of estimation for sparse pca in high dimensions. *Arxiv preprint arXiv:1202.0786*, 2012.

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

Yuan, X.T. and Zhang, T. Truncated power method for sparse eigenvalue problems. *Arxiv preprint arXiv:1112.2679*, 2011.

Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.