Semiparametric Principal Component Analysis

Fang Han
Department of Biostatistics
Johns Hopkins University
Baltimore, MD 21210
fhan@jhsph.edu

Han Liu
Department of Operations Research
and Financial Engineering
Princeton University, NJ 08544
hanliu@princeton.edu

Abstract

We propose two new principal component analysis methods in this paper utilizing a semiparametric model. The according methods are named Copula Component Analysis (COCA) and Copula PCA. The semiparametric model assumes that, after unspecified marginally monotone transformations, the distributions are multivariate Gaussian. The COCA and Copula PCA accordingly estimate the leading eigenvectors of the correlation and covariance matrices of the latent Gaussian distribution. The robust nonparametric rank-based correlation coefficient estimator, Spearman's rho, is exploited in estimation. We prove that, under suitable conditions, although the marginal distributions can be arbitrarily continuous, the COCA and Copula PCA estimators obtain fast estimation rates and are feature selection consistent in the setting where the dimension is nearly exponentially large relative to the sample size. Careful numerical experiments on the synthetic and real data are conducted to back up the theoretical results. We also discuss the relationship with the transelliptical component analysis proposed by Han and Liu (2012).

1 Introduction

The Principal Component Analysis (PCA) is introduced as follows. Given a random vector $X \in \mathbb{R}^d$ with covariance matrix Σ and n independent observations of X, the PCA reduces the dimension of the data by projecting the data onto a linear subspace spanned by the k leading eigenvectors of Σ, such that the principal modes of variations are preserved. In practice, Σ is unknown and replaced by the sample covariance S. By spectral decomposition, $\Sigma = \sum_{j=1}^d \omega_j u_j u_j^T$ with eigenvalues $\omega_1 \geq \ldots \geq \omega_d$ and the corresponding orthornormal eigenvectors $u_1, \ldots, u_d$. PCA aims at recovering the first k eigenvectors $u_1, \ldots, u_k$.

Although the PCA method as a procedure is model free, its theoretical and empirical performances rely on the distributions. With regard to the empirical concern, the PCA's geometric intuition is coming from the major axes of the contours of constant probability of the Gaussian [10]. [5] show that if X is multivariate Gaussian, then the distribution is centered about the principal component axes and is therefore "self-consistent" [8]. We refer to [10] for more good properties that the PCA enjoys under the Gaussian model, which we wish to preserve while designing its generalization.

With regard to the theoretical concern, firstly, the PCA generally fails to be consistent in high dimensional setting. Given $\widehat{u}_1$ the dominant eigenvector of S, [9] show that the angle between $\widehat{u}_1$ and u_1 will not converge to 0, i.e. $\liminf_{n\to\infty} \mathbb{E}\angle(\widehat{u}_1, u_1) > 0$, where we denote by $\angle(\widehat{u}_1, u_1)$ the angle between the estimated and the true leading eigenvectors. This key observation motivates regularizing Σ, resulting in a series of methods with different formulations and algorithms. The statistical model is generally further specified such that u_1 is sparse, namely $\text{supp}(u_1) := \{j : u_{1j} \neq 0\} \text{ and } \text{card}(\text{supp}(u_1)) = s < n$. The resulting estimator $\widetilde{u}_1$ is:

$$\widetilde{u}_1 = \arg\max_{v \in \mathbb{R}^d} v^T S v \quad \text{subject to} \quad \|v\|_2 = 1, \text{card}(\text{supp}(v)) \leq s. \tag{1.1}$$

To solve Equation (1.1), a variety of algorithms are proposed: greedy algorithms [3], lasso-type methods including SCoTLASS [11], SPCA [25] and sPCA-rSVD [19], a number of power methods [12, 23, 16], the biconvex algorithm PMD [21] and the semidefinite relaxation DSPCA [4]. Secondly, it is realized that the distribution where the data are drawn from needs to be specified, such that the estimator $\widetilde{u}_1$ converges to $\bar{u}_1$ in a fast rate. [9, 1, 16, 18, 20] all establish their results under a strong Gaussian or sub-Gaussian assumption in order to obtain a fast rate under certain conditions.

In this paper, we first explore the use of the PCA conducted on the correlation matrix Σ^0 instead of the covariance matrix Σ, and then propose a high dimensional semiparametric scale-invariant principal component analysis method, named the Copula Component Analysis (COCA). In this paper, the population version of the scale-invariant PCA is built as the estimator of the leading eigenvector of the population correlation matrix Σ^0. Secondly, to handle the non-Gaussian data, we generalize the distribution family from the Gaussian to the larger Nonparanormal family [15]. A random variable $X = (X_1, \ldots, X_d)^T$ belongs to a Nonparanormal family if and only if there exists a set of univariate monotone functions $\{f_j^0\}_{j=1}^d$ such that $(f_1^0(X_1), \ldots, f_d^0(X_d))^T$ is multivariate Gaussian. The Nonparanormal can have arbitrary continuous marginal distributions and can be far away from the sub-Gaussian family. Thirdly, to estimate Σ^0 robustly and efficiently, instead of estimating the normal score transformation functions $\{\widehat{f}_j^0\}_{j=1}^d$ as [15] did, realizing that $\{f_j^0\}_{j=1}^d$ preserve the ranks of the data, we utilize the nonparametric correlation coefficient estimator, Spearman's rho, to estimate Σ^0. [14, 22] prove that the corresponding estimators converge to Σ^0 in a parametric rate. In theory, we analyze the general case that X is following the Nonparanormal and θ_1 is weakly sparse, here θ_1 is the leading eigenvector of Σ^0. We obtain the estimation consistency of the COCA estimator to θ_1 using the Spearman's rho correlation coefficient matrix. We prove that the estimation consistency rates are close to the parametric rate under Gaussian assumption and the feature selection consistency can be achieved when d is nearly exponential to the sample size. In this paper, we also propose a scale variant PCA procedure, named the Copula PCA. The Copula PCA estimates the leading eigenvector of the latent covariance matrix Σ. To estimate the leading eigenvectors of Σ, instead of Σ^0, in a fast rate, we prove that extra conditions are required on the transformation functions.

2 Background

We start with notations: Let $M = [M_{jk}] \in \mathbb{R}^{d \times d}$ and $v = (v_1, ..., v_d)^T \in \mathbb{R}^d$. Let v's subvector with entries indexed by I be denoted by v_I, M's submatrix with rows indexed by I and columns indexed by J be denoted by M_{IJ}. Let $M_{I\cdot}$ and $M_{\cdot J}$ be the submatrix of M with rows in I and all columns, and the submatrix of M with columns in J and all rows. For $0 < q \leq \infty$, we define the ℓ_q and ℓ_∞ vector norm as $\|v\|_q := (\sum_{i=1}^d |v_i|^q)^{1/q}$ and $\|v\|_\infty := \max_{1 \leq i \leq d} |v_i|$, and $\|v\|_0 := \text{card}(\text{supp}(v)) \cdot \|v\|_2$. We define the matrix $\ell_{\max}$ norm as the elementwise maximum value: $\|M\|_{\max} := \max\{|M_{ij}|\}$ and the ℓ_∞ norm as $\|M\|_\infty := \max_{1 \leq i \leq m} \sum_{j=1}^n |M_{ij}|$. Let $\Lambda_j(M)$ be the toppest $j-$th eigenvalue of M. In special, $\Lambda_{\min}(M) := \Lambda_d(M)$ and $\Lambda_{\max}(M) := \Lambda_1(M)$ are the smallest and largest eigenvalues of M. The vectorized matrix of M, denoted by vec(M), is defined as: $\text{vec}(M) := (M_{\cdot 1}^T, \ldots, M_{\cdot d}^T)^T$. Let $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d : \|v\|_2 = 1\}$ be the d-dimensional ℓ_2 sphere. For any two vectors $a, b \in \mathbb{R}^d$ and any two squared matrices $A, B \in \mathbb{R}^{d \times d}$, denote the inner product of a and b, A and B by $\langle a, b \rangle := a^T b \quad \text{and} \quad \langle A, B \rangle := \text{Tr}(A^T B)$.

2.1 The Models of the PCA and Scale-invariant PCA

Let Σ^0 be the correlation matrix of Σ, and by spectral decomposition, $\Sigma = \sum_{j=1}^d \omega_j u_j u_j^T$ and $\Sigma^0 = \sum_{j=1}^d \lambda_j \theta_j \theta_j^T$. Here $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_d > 0$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$ are the eigenvalues of Σ and Σ^0, with $u_1, \ldots, u_d$ and $\theta_1, \ldots, \theta_d$ the corresponding orthonormal eigenvectors. The next proposition claims that the estimators $\{\widehat{u}_1, \ldots, \widehat{u}_d\}$ and $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_d\}$, the eigenvectors of the sample covariance and correlation matrices S and S^0, are the MLEs of $\{u_1, \ldots, u_d\}$ and $\{\theta_1, \ldots, \theta_d\}$:

Proposition 2.1. *Let $x_1 \ldots x_n \sim N(\mu, \Sigma)$ and Σ^0 be the correlation matrix of Σ. Then the estimators of PCA, $\{\widehat{u}_1, \ldots, \widehat{u}_d\}$, and the estimators of the scale-invariant PCA, $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_d\}$, are the MLEs of $\{u_1, \ldots, u_d\}$ and $\{\theta_1, \ldots, \theta_d\}$.*

Proof. Use Theorem 11.3.1 in [2] and the functional invariance property of the MLE. □

Proposition 2.2. *For any $1 \leq i \leq d$, we have $\text{supp}(u_i) = \text{supp}(\theta_i)$ and $\text{sign}(u_{ij}) = \text{sign}(\theta_{ij}), \ \forall\, 1 \leq j \leq d$.*

Proof. For $1 \leq i \leq d$, $u_i = (\theta_{i1}/\sigma_1, \theta_{i2}/\sigma_2, \ldots, \theta_{id}/\sigma_d)$, where $(\sigma_1^2, \ldots, \sigma_d^2)^T := \text{diag}(\Sigma)$. □

It is easy to observe that the scale-invariant PCA is a safe procedure for dimension reduction when variables are measured in different scales. Although there seems no theoretical advantage of scale-invariant PCA over the PCA under the Gaussian model, in this paper we will show that under a more general Nonparanormal (or Gaussian Copula) model, the scale-invariant PCA will pose much less conditions to make the estimator achieve good theoretical performance.

2.2 The Nonparanormal

We first introduce two definitions of the Nonparanormal separately defined in [15] and [14].

Definition 2.1 [15]. A random variable $X = (X_1, ..., X_d)^T$ with population marginal means and standard deviations $\mu = (\mu_1, \dots, \mu_d)^T$ and $\sigma = (\sigma_1. \dots, \sigma_d)^T$ is said to follow a Nonparanormal distribution $NPN_d(\mu, \Sigma, f)$ if and only if there exists a set of univariate monotone transformations $f = \{f_j\}_{j=1}^d$ such that: $f(X) = (f_1(X_1), ..., f_d(X_d))^T \sim N(\mu, \Sigma)$, and $\sigma_j^2 = \Sigma_{jj}, \quad j = 1, \dots, d$.

Definition 2.2 [14]. Let $f^0 = \{f_j^0\}_{j=1}^d$ be a set of monotone univariate functions and $\Sigma^0 \in \mathbb{R}^{d\times d}$ be a positive definite correlation matrix with $\mathrm{diag}(\Sigma^0) = \mathbf{1}$. We say that a d dimensional random variable $X = (X_1, \dots, X_d)^T$ follows a Nonparanormal distribution, i.e. $X \sim NPN_d(\Sigma^0, f^0)$, if $f^0(X) := (f_1^0(X_1), \dots, f_d^0(X_d))^T \sim N(0, \Sigma^0)$.

The following lemma proves that two definitions of the Nonparanormal are equivalent.

Lemma 2.1. *A random variable $X \sim NPN_d(\Sigma^0, f^0)$ if and only if there exist $\mu = (\mu_1, \dots, \mu_d)^T$, $\Sigma = [\Sigma_{jk}] \in \mathbb{R}^{d\times d}$ such that for any $1 \le j, k \le d$, $\mathbb{E}(X_j) = \mu_j,\ \mathrm{Var}(X_j) = \Sigma_{jj}$ and $\Sigma_{jk}^0 = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\cdot\Sigma_{kk}}}$, and a set of monotone univariate functions $f = \{f_j\}_{j=1}^d$ such that $X \sim NPN_d(\mu, \Sigma, f)$.*

Proof. Using the connection that $f_j(x) = \mu_j + \sigma_j f_j^0(x)$, for $j \in \{1, 2 \dots, d\}$. □

Lemma 2.1 guarantees that the Nonparanormal is defined properly. Definition 2.2 is more appealing because it emphasizes the correlation and hence matches the spirit of the Copula. However, Definition 2.1 enjoys notational simplicity in analyzing the Copula-based LDA and PCA approaches.

2.3 Spearman's rho Correlation and Covariance Matrices

Given n data points $x_1, \dots, x_n \in \mathbb{R}^d$, where $x_i = (x_{i1}, \dots, x_{id})^T$, we denote by $\widehat{\mu}_j := \frac{1}{n}\sum_{i=1}^n x_{ij}$ and $\widehat{\sigma}_j = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_{ij} - \widehat{\mu}_j)^2}$, the marginal sample means and standard deviations. Because the Nonparanormal distribution preserves the rank of the data, it is natural to use the nonparametric rank-based correlation coefficient estimator, Spearman's rho, to estimate the latent correlation. In detail, let r_{ij} be the rank of x_{ij} among $x_{1j}, \dots, x_{nj}$ and $\bar{r}_j := \frac{1}{n}\sum_{i=1}^n r_{ij} = \frac{n+1}{2}$, we consider the following statistics: $\widehat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_{ij} - \bar{r}_j)(r_{ik} - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_{ij} - \bar{r}_j)^2 \cdot \sum_{i=1}^n (r_{ik} - \bar{r}_k)^2}}$, and the correlation matrix estimator: $\widehat{R}_{jk} = 2\sin(\frac{\pi}{6}\widehat{\rho}_{jk})$. The Lemma 2.2, coming from [14], claims that the estimation can reach the parametric rate.

Lemma 2.2 ([14]). *When $x_1, \dots, x_n \sim^{i.i.d} NPN_d(\Sigma^0, f^0)$, for any $n \ge \frac{21}{\log d} + 2$,*

$$\mathbb{P}\left(\|\widehat{R} - \Sigma^0\|_{\max} \le 8\pi\sqrt{\frac{\log d}{n}}\right) \ge 1 - 2/d^2. \tag{2.1}$$

We denote by $\widehat{R} := [\widehat{R}_{jk}]$ the Spearman's rho correlation coefficient matrix. In the following let $\widehat{S} := [\widehat{S}_{jk}] = [\widehat{\sigma}_j\widehat{\sigma}_k\widehat{R}_{jk}]$ be the Spearman's rho covariance matrix.

3 Methods

In Figure 1, we randomly generate 10,000 samples from three different types of Nonparanormal distributions. We suppose that $X \sim NPN_2(\Sigma^0, f^0)$. Here we set $\Sigma^0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ and transformation functions as follows: (A) $f_1^0(x) = x^3$ and $f_2^0(x) = x^{1/3}$; (B) $f_1^0(x) = \mathrm{sign}(x)x^2$ and $f_2^0(x) = x^3$; (C) $f_1^0(x) = f_2^0(x) = \Phi^{-1}(x)$. It can be observed that there does not exist a nice geometric explanation now. For example, researchers might wish to conduct PCA separately on different clusters in (A) and (B). For (C), the data look very noisy and a nice major axis might be considered not existing.

However, under the Nonparanormal model and realizing that there is a latent Gaussian distribution behind, the geometric intuition of the PCA naturally comes back. In the next section, we will present the model of the COCA and Copula PCA motivated from this observation.

3.1 COCA Model

We firstly present the model of the Copula Component Analysis (COCA) method, where the idea of scale-invariant PCA is exploited and we wish to estimate the leading eigenvector of the latent correlation matrix. In particular, the following model $\mathcal{M}^0(q, R_q, \Sigma^0, f^0)$ is considered:

$$\mathcal{M}^0(q, R_q, \Sigma^0, f^0): \quad \begin{cases} x_1, \dots, x_n \sim^{i.i.d} NPN_d(\Sigma^0, f^0), \\ \theta_1 \in \mathbb{S}^{d-1} \cap \mathbb{B}_q(R_q), \end{cases} \tag{3.1}$$

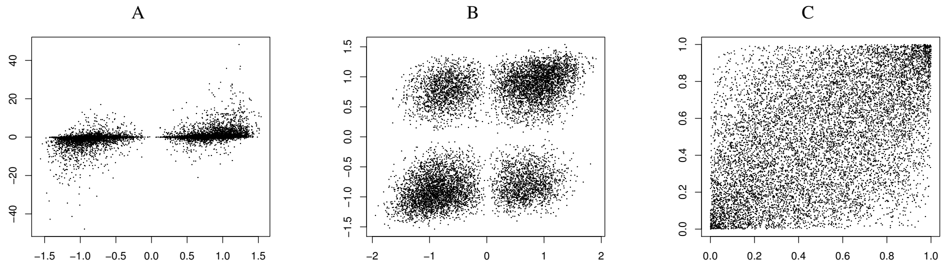

Figure 1: Scatter plots of three Nonparanormals, $X \sim NPN_2(\Sigma^0, f^0)$. Here $\Sigma^0_{12} = 0.5$ and the transformation functions have the form as follows: (A) $f^0_1(x) = x^3$ and $f^0_2(x) = x^{1/3}$; (B) $f^0_1(x) = \text{sign}(x)x^2$ and $f^0_2(x) = x^3$; (C) $f^0_1(x) = f^0_2(x) = \Phi^{-1}(x)$.

where θ_1 is the leading eigenvectors of the latent correlation matrix Σ^0 we are interested in estimating, $0 \leq q \leq 1$ and the ℓ_q ball $\mathbb{B}_q(R_q)$ is defined as:

$$\text{when } \; q = 0, \qquad \mathbb{B}_0(R_0) := \{v \in \mathbb{R}^d : \text{card}(\text{supp}(v)) \leq R_0\}; \tag{3.2}$$

$$\text{when } \; 0 < q \leq 1, \qquad \mathbb{B}_q(R_q) := \{v \in \mathbb{R}^d : \|v\|_q^q \leq R_q\}. \tag{3.3}$$

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

$$\widetilde{\theta}_1 = \arg\max_{v \in \mathbb{R}^d} \; v^T \widehat{R} v, \text{subject to } \; v \in \mathbb{S}^{d-1} \cap \mathbb{B}_q(R_q). \tag{3.4}$$

Here $\widehat{R}$ is the estimated Spearman's rho correlation coefficient matrix. The corresponding COCA estimator $\widetilde{\theta}_1$ can be considered as a nonlinear dimensional reduction procedure and has the potential to gain more flexibility compared with the classical PCA. In Section 4 we will establish the theoretical results on the COCA estimator and will show that it can estimate the latent true dominant eigenvector θ_1 in a fast rate and can achieve feature selection consistency.

3.1.1 Copula PCA Model

In contrast, we provide another model inspired from the classical PCA method, where we wish to estimate the leading eigenvector of the latent covariance matrix. In particular, the following model $\mathcal{M}(q, R_q, \Sigma, f)$ is considered:

$$\mathcal{M}(q, R_q, \Sigma, f) : \quad \begin{cases} x_1, \ldots, x_n \sim^{i.i.d} NPN_d(0, \Sigma, f), \\ u_1 \in \mathbb{S}^{d-1} \cap \mathbb{B}_q(R_q), \end{cases} \tag{3.5}$$

where u_1 is the leading eigenvector of the covariance matrix Σ and it is what we are interested in estimating. The corresponding Copula PCA estimator is:

$$\widetilde{u}_1 = \arg\max_{v \in \mathbb{R}^d} \; v^T \widehat{S} v, \text{subject to } \; v \in \mathbb{S}^{d-1} \cap \mathbb{B}_q(R_q), \tag{3.6}$$

where $\widehat{S}$ is the Spearman's rho covariance coefficient matrix. This procedure is named the Copula PCA. In Section 4, we will show that the Copula PCA requires a much stronger condition than COCA to make $\widetilde{u}_1$ converge to u_1 in a fast rate.

3.2 Algorithms

In this section we provide three sparse PCA algorithms, where the Spearman's rho correlation and covariance matrices $\widehat{R}$ and $\widehat{S}$ can be directly plugged in to obtain sparse estimators.

Penalized Matrix Decomposition (PMD) is proposed by [21]. The main idea of the PMD is a biconvex optimization algorithm to the following problem: $\arg\max_{u,v} u^T \widehat{\Gamma} v, \quad \text{subject to } \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1, \; \|u\|_1 \leq \delta, \; \|v\|_1 \leq \delta$. The COCA with PMD and Copula PCA with PMD are listed in the following: (1) Input: A symmetric matrix $\widehat{\Gamma}$. Initialize $v \in \mathbb{S}^{d-1}$; (2) Iterate until convergence: (a) $u \leftarrow \arg\max_{u \in \mathbb{R}^d} u^T \widehat{\Gamma} v$ subject to $\|u\|_1 \leq \delta$ and $\|u\|_2^2 \leq 1$.(b) $v \leftarrow \arg\max_{v \in \mathbb{R}^d} u^T \widehat{\Gamma} v$ subject to $\|v\|_1 \leq \delta$ and $\|v\|_2^2 \leq 1$; (3) Output: v. Here $\widehat{\Gamma}$ is either $\widehat{R}$ or $\widehat{S}$, corresponding to the COCA with PMD and Copula PCA with PMD. δ is the tuning parameter. [21] suggest using the first leading eigenvector of $\widehat{\Gamma}$ to be the initial value of v. The PMD can be considered as a solver to Equation (3.4) and Equation (3.6) with $q = 1$.

The SPCA algorithm is proposed by [25]. The main idea of the SPCA algorithm is to exploit a regression approach to PCA and then utilize lasso and elastic net [24] to calculate a sparse estimator to the leading eigenvector. The COCA with SPCA and Copula PCA with SPCA are listed as follows: (1) Input: A symmetric matrix $\widehat{\Gamma}$. Initialize $u \in \mathbb{S}^{d-1}$. (2). Iterate until convergence: (a) $v \leftarrow \arg\min_{v\in\mathbb{R}^d}(u-v)^T\widehat{\Gamma}(u-v) + \delta_1\|v\|_2^2 + \delta_2\|v\|_1$; (b) $u \leftarrow \widehat{\Gamma}v/\|\widehat{\Gamma}v\|_2$. (3) Output: $v/\|v\|_2$. Here $\widehat{\Gamma}$ is either $\widehat{R}$ or $\widehat{S}$, corresponding to the COCA with SPCA and Copula PCA with SPCA. $\delta_1 \in \mathbb{R}$ and $\delta_2 \in \mathbb{R}$ are two tuning parameters. [25] suggest using the first leading eigenvector of $\widehat{\Gamma}$ to be the initial value of v. The SPCA can be considered as a solver to Equations (3.4) and (3.6) with $q = 1$.

The Truncated Power method (TPower) is proposed by [23]. The main idea is to utilize the power method, but truncate the vector to a ℓ_0 ball in each iteration. Actually, TPower can be generalized to a family of algorithms to solve Equation (3.4) when $0 \le q \le 1$. We name it the ℓ_q Constraint Truncated Power Method (qTPM). Especially, when $q = 0$, the algorithm qTPM coincides with [23]'s method. The TPower can be considered as a general solver to Equation (3.4) and Equation (3.6) with $q \in [0, 1]$. In detail, we utilize the classical power method, but in 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_q ball with the radius $R_q^{1/q}$. Detailed algorithms are presented in the long version of this paper [6].

4 Theoretical Properties

In this section we provide the theoretical properties of the COCA and Copula PCA methods. Especially, we are interested in the high dimensional case when $d > n$.

4.1 Rank-based Correlation and Covariance Matrices Estimation

This section is devoted to the statement of our result on quantifying the convergence rate of $\widehat{R}$ to Σ^0 and $\widehat{S}$ to Σ. In particular, we establish the results on the $\ell_{\max}$ convergence rates of the Spearman's rho correlation and covariance matrices to Σ and Σ^0. For COCA, Lemma 2.2 is enough. For Copula PCA, however, we still need to quantify the convergence rate of $\widehat{S}$ to Σ.

Definition 4.1 Subgaussian Transformation Function Class. Let $Z \in \mathbb{R}$ be a random variable following the standard Gaussian distribution. The Subgaussian Transformation Function Class $\mathrm{TF}(K)$ is defined as the set of functions $\{g_0 : \mathbb{R} \to \mathbb{R}\}$ which satisfies that: $\mathbb{E}|g_0(Z)|^m \le \frac{m!}{2}K^m, \quad \forall\, m \in \mathbb{Z}^+$.

Here it is easy to see that for any function $g_0 : \mathbb{R} \to \mathbb{R}$, if there exists a constant $L < \infty$ such that $g_0(z) \le L \quad \text{or} \quad g_0'(z) \le L \quad \text{or} \quad g_0''(z) \le L, \ \forall\, z \in \mathbb{R}$, then $g_0 \in \mathrm{TF}(K)$ for some constant K. Then we have the following result, which states that Σ can also be recovered in the parametric rate.

Lemma 4.1. *When $x_1, \ldots, x_n \sim^{i.i.d} NPN_d(\mu, \Sigma, f)$, $0 < 1/c_0 < \min_j\{\sigma_j\} < \max_j\{\sigma_j\} < c_0 < \infty$, for some constant c_0 and $g := \{g_j = f_j^{-1}\}_{j=1}^d$ satisfies for all $j = 1, \ldots, K$, $g_j^2 \in TF(K)$ where $K < \infty$ is some constant, we have for any $1 \le j, k \le d$, for any $n \ge \frac{21}{\log d} + 2$,*

$$\mathbb{P}(|\widehat{S}_{jk} - \Sigma_{jk}| > t) \le 2\exp(-c_1 n t^2), \tag{4.1}$$

where c_1 is a constant only depending on the choice of K.

Remark 4.1. The Lemma 4.1 claims that, under certain constraint on the transformation functions, the latent covariance matrix Σ can be recovered using the Spearman's rho covariance matrix. However, in this case, the marginal distributions of the Nonparanormal are required to be sub-gaussian and cannot be arbitrarily continuous. This makes the Copula PCA a less favored method.

4.2 COCA and Copula PCA

This section is devoted to the statement of our main result on the upper bound of the estimated error of the COCA estimator and Copula PCA estimator.

Theorem 4.1 (Upper bound for the COCA). *Let $\widetilde{\theta}_1$ be the global solution to Equation* (3.4) *and the Model $\mathcal{M}^0(q, R_q, \Sigma^0, f^0)$ holds. For any two vectors $v_1 \in \mathbb{S}^{d-1}$ and $v_2 \in \mathbb{S}^{d-1}$, let $|\sin\angle(v_1, v_2)| = \sqrt{1 - (v_1^T v_2)^2}$, then we have, for any $n \ge \frac{21}{\log d} + 2$,*

$$\mathbb{P}\left(\sin^2\angle(\widetilde{\theta}_1, \theta_1) \le \gamma_q R_q^2 \left(\frac{64\pi^2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{\log d}{n}\right)^{\frac{2-q}{2}}\right) \ge 1 - 1/d^2, \tag{4.2}$$

where $\gamma_q = 2 \cdot I(q = 1) + 4 \cdot I(q = 0) + (1 + \sqrt{3})^2 \cdot I(0 < q < 1)$.

Proof. The key idea of the proof is to utilize the $\ell_{\max}$ norm convergence result of $\widehat{R}$ to Σ^0. Detailed proofs are presented in the long version of this paper [6]. □

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 [16, 20, 18] obtain. When (n, d) goes to infinity, the two dominant eigenvalues λ_1 and λ_2 will typically go to infinity and will at least be away from zero. Hence, our rate shown in Equation (4.2) is better than the seemingly more state-of-art rate: $\gamma_q R_q^2 \left(\frac{64\pi^2\lambda_1^2}{(\lambda_1-\lambda_2)^2} \cdot \frac{\log d}{n}\right)^{\frac{2-q}{2}}$.

The COCA is significantly different from [20] and [18]'s results in the sense that: (1) In theory, the Nonparanormal family can have arbitrary continuous marginal distributions, where a fast rate cannot be obtained using the techniques built for either Gaussian or sub-Gaussian distributions; (2) In methodology, we utilize the Spearman's rho correlation coefficient matrix $\widehat{R}$ to estimate Σ^0, instead of using the sample correlation matrix S^0. This procedure has been shown to lose little in rate and will be much more robust under the Nonparanormal model. Given Theorem 4.1, we can immediately obtain a feature selection consistency result.

Corollary 4.1 (Feature Selection Consistency of the COCA). *Let* $\widetilde{\theta}_1$ *be the global solution to Equation* (3.4) *and the Model* $\mathcal{M}^0(0, R_0, \Sigma^0, f^0)$ *holds. Let* $\Theta^0 := \mathrm{supp}(\theta_1)$ and $\widehat{\Theta}^0 := \mathrm{supp}(\widetilde{\theta}_1)$. *If we further have* $\min_{j\in\Theta^0} |\theta_{1j}| \geq \frac{16\sqrt{2}R_0\pi}{\lambda_1-\lambda_2}\sqrt{\frac{\log d}{n}}$, *then for any* $n \geq 21/\log d + 2$, $\mathbb{P}(\widehat{\Theta}^0 = \Theta^0) \geq 1 - 1/d^2$.

Similarly, we can give an upper bound for the estimation rate of the Copula PCA to the true leading eigenvalue u_1 of the latent covariance matrix Σ. The next theorem provides the detail result.

Theorem 4.2 (Upper bound for Copula PCA). *Let* $\widetilde{u}_1$ *be the global solution to Equation* (3.6) *and the Model* $\mathcal{M}(q, R_q, \Sigma, f)$ *holds. If* $g := \{g_j = f_j^{-1}\}_{j=1}^d$ *satisfies* $g_j^2 \in TF(K)$ *for all* $1 \leq j \leq d$, *and* $0 < 1/c_0 < \min_j\{\sigma_j\} < \max_j\{\sigma_j\} < c_0 < \infty$, *then we have, for any* $n \geq 21/\log d + 2$,

$$\mathbb{P}\left(\sin^2\angle(\widetilde{u}_1, u_1) \leq \gamma_q R_q^2 \left(\frac{4}{c_1(\omega_1-\omega_2)^2} \cdot \frac{\log d}{n}\right)^{\frac{2-q}{2}}\right) \geq 1 - 1/d^2,$$

where $\gamma_q = 2 \cdot I(q = 1) + 4 \cdot I(q = 0) + (1 + \sqrt{3})^2 \cdot I(0 < q < 1)$ *and* c_1 *is a constant defined in Equation* (4.1), *only depending on* K.

Corollary 4.2 (Feature Selection Consistency of the Copula PCA). *Let* $\widetilde{u}_1$ *be the global solution to Equation* (3.6) *and the Model* $\mathcal{M}(0, R_0, \Sigma, f)$ *holds. Let* $\Theta := \mathrm{supp}(u_1)$ and $\widehat{\Theta} := \mathrm{supp}(\widetilde{u}_1)$. *If* $g := \{g_j = f_j^{-1}\}_{j=1}^d$ *satisfies* $g_j^2 \in TF(K)$ *for all* $1 \leq j \leq d$, *and* $0 < 1/c_0 < \min_j\{\sigma_j\} < \max_j\{\sigma_j\} < c_0 < \infty$, *and we further have* $\min_{j\in\Theta} |u_{1j}| \geq \frac{4\sqrt{2}R_0}{\sqrt{c_1}(\omega_1-\omega_2)}\sqrt{\frac{\log d}{n}}$, *then for any* $n \geq \frac{21}{\log d} + 2$, $\mathbb{P}(\widehat{\Theta} = \Theta) \geq 1 - \frac{1}{d^2}$.

5 Experiments

In this section we investigate the empirical usefulness of the COCA method. Three sparse PCA algorithms are considered: PMD proposed by [21], SPCA proposed by [25] and Truncated Power method (TPower) proposed by [23]. The following three methods are considered: (1) Pearson: the classic high dimensional PCA using the Pearson sample correlation matrix; (2) Spearman: the COCA using the Spearman's rho correlation coefficient matrix; (3) Oracle: the classic high dimensional PCA using the Pearson sample correlation matrix of the data from the latent Gaussian (perfect without contaminations).

5.1 Numerical Simulations

In the simulation study we randomly sample n data points $x_1, \ldots, x_n$ from the Nonparanormal distribution $X \sim NPN_d(\Sigma^0, f^0)$. Here we consider the setup of $d = 100$. We follow the same generating scheme as in [19, 23] and [7]. A covariance matrix Σ 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 suppose that the first two dominant eigenvectors of Σ, 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 are nonzero and set to be $1/\sqrt{10}$. $\omega_1 = 5$, $\omega_2 = 2$,

$\omega_3 = \ldots = \omega_d = 1$. The remaining eigenvectors are chosen arbitrarily. The correlation matrix Σ^0 is accordingly generated from Σ, with $\lambda_1 = 4$, $\lambda_2 = 2.5$, $\lambda_3, \ldots, \lambda_d \leq 1$ and the two dominant eigenvectors sparse. To sample data from the Nonparanormal, we also need the transformation functions: $f^0 = \{f_j^0\}_{j=1}^d$. Here two types of transformation functions are considered: (1) **Linear transformation** (or no transformation): $f^0_{\rm linear} = \{h_0, h_0, \ldots, h_0\}$, where $h_0(x) := x$; (2) **Nonlinear transformation**: there exist five univariate monotone functions $h_1, h_2, \ldots, h_5 : \mathbb{R} \to \mathbb{R}$ and $f^0_{\rm nonlinear} = \{h_1, h_2, h_3, h_4, h_5, h_1, h_2, h_3, h_4, h_5, \ldots\}$, where $h_1^{-1}(x) := x$, $h_2^{-1}(x) := \frac{{\rm sign}(x)|x|^{1/2}}{\sqrt{\int |t|\phi(t)dt}}$, $h_3^{-1}(x) := \frac{x^3}{\sqrt{\int t^6\phi(t)dt}}$, $h_4^{-1}(x) := \frac{\Phi(x) - \int \Phi(t)\phi(t)dt}{\sqrt{\int (\Phi(y) - \int \Phi(t)\phi(t)dt)^2\phi(y)dy}}$, $h_5^{-1}(x) := \frac{\exp(x) - \int \exp(t)\phi(t)dt}{\sqrt{\int (\exp(y) - \int \exp(t)\phi(t)dt)^2\phi(y)dy}}$. Here ϕ and Φ are defined to be the probability density and cumulative distribution functions of the standard Gaussian. $h_1, \ldots, h_5$ are defined such that for any $Z \sim N(0,1)$, $\mathbb{E}(h_j^{-1}(Z)) = 0$ and ${\rm Var}(h_j^{-1}(Z)) = 1 \quad \forall\, j \in \{1, \ldots, 5\}$. We then generate $n = 100, 200$ or 500 data points from:

[Scheme 1] $X \sim NPN_d(\Sigma^0, f^0_{\rm linear})$ where $f^0_{\rm linear} = \{h_0, h_0, \ldots, h_0\}$ and Σ_0 is defined as above.

[Scheme 2] $X \sim NPN_d(\Sigma^0, f^0_{\rm nonlinear})$ where $f^0_{\rm nonlinear} = \{h_1, h_2, h_3, h_4, h_5, \ldots\}$.

To evaluate the robustness of different methods, we adopt a similar data contamination procedure as in [14]. Let $r \in [0, 1)$ represents the proportion of samples being contaminated. For each dimension, we randomly select $\lfloor nr \rfloor$ entries and replace them with either 5 or -5 with equal probability. The final data matrix we obtained is $\boldsymbol{X} \in \mathbb{R}^{n \times d}$. The PMD, SPCA and TPower algorithms are then employed on $\boldsymbol{X}$ to computer the estimated leading eigenvector $\widetilde{\theta}_1$.

Under the Scheme 1 and Scheme 2 with different levels of contamination ($r = 0$ or 0.05), 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. The corresponding ROC curves are presented in Figure 2. More quantitative results are provided in the long version of this paper [6]. It can be observed that when $r = 0$ and $\boldsymbol{X}$ is exactly Gaussian, Pearson,Spearman and Oracle can all recover the sparsity pattern perfectly. However, when $r > 0$, the performances of Pearson significantly decrease, while Spearman is still very close to the Oracle. In Scheme 2, even when $r = 0$, Pearson cannot recover the support set of θ_1, while Spearman can still recover the sparsity pattern almost perfectly. When $r > 0$, the performance of Spearman is still very close to the Oracle.

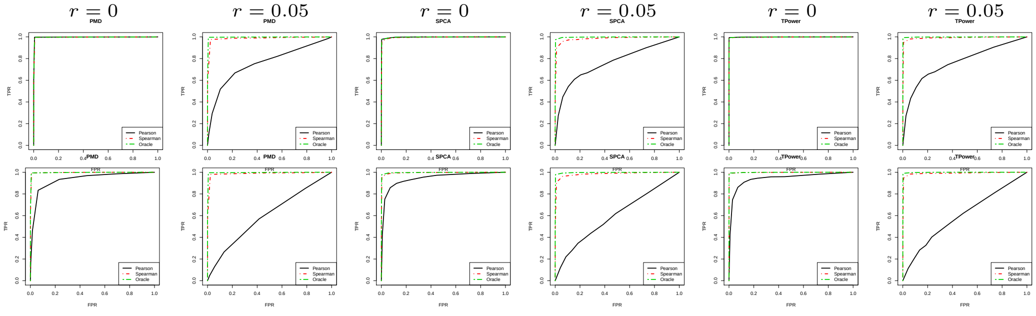

Figure 2: ROC curves for the PMD, SPCA and Truncated Power method (the left two, the middle two, the right two) with linear (no) and nonlinear transformation (top, bottom) and data contamination at different levels ($r = 0, 0.05$). Here $n = 100$ and $d = 100$.

5.2 Large-scale Genomic Data Analysis

In this section we investigate the performance of Spearman compared with the Pearson using one of the largest microarray datasets [17]. In summary, we collect in all 13,182 publicly available microarray samples from Affymetrixs HGU133a platform. The raw data contain 20,248 probes and 13,182 samples belonging to 2,711 tissue types (e.g., lung cancers, prostate cancer, brain tumor etc.). There are at most 1,599 samples and at least 1 sample belonging to each tissue type. We merge the probes corresponding to the same gene. There are remaining 12,713 genes and 13,182 samples. This dataset is non-Gaussian (see the long version of this paper [6]). The main purpose of this experiment is to compare the performance of the COCA with the classical high dimensional PCA. We utilize the Truncated Power method proposed by [23] to achieve the sparse estimated dominant eigenvectors.

We adopt the same idea of data-preprocessing as in [14]. In particular, we firstly remove the batch effect by applying the surrogate variable analysis proposed by [13]. We then extract the top 2,000 genes with the highest marginal standard deviations. There are, accordingly, 2,000 genes left and the data matrix we are focusing is $2,000 \times 13,182$. We then explore several tissue types with the largest sample size: (1) Breast tumor, 1,599 samples; (2) B cell lymphoma, 213 samples; (3) Prostate tumor, 148 samples; (4) Wilms tumor, 143 samples.

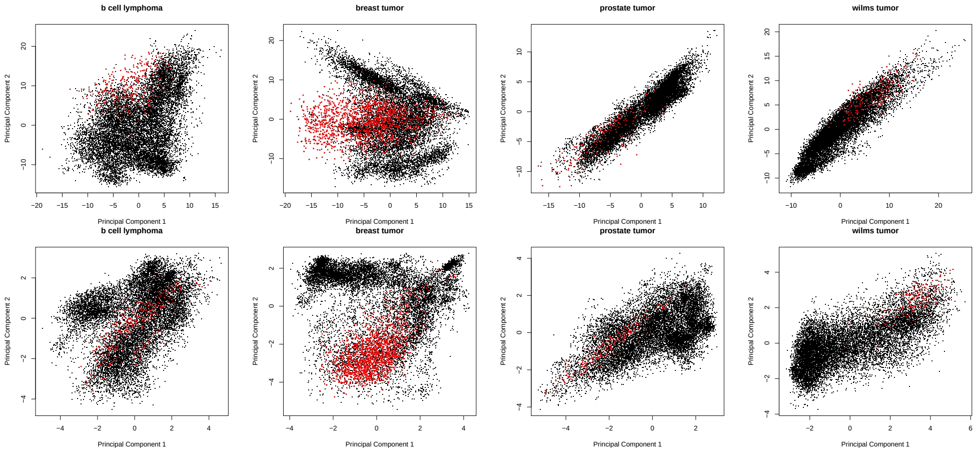

Figure 3: The scatter plots of the first two principal components of the dataset. The Spearman versus Pearson are compared (top to bottom). b cell lymphoma, breast tumor, prostate tumor and Wilms tumor are explored (from left to right). Each black point represents a sample and each red point represents a sample belonging to the corresponding tissue type.

For each tissue type listed above, we apply the COCA (Spearman) and the classic high dimensional PCA (Pearson) on the data belonging to this specific tissue type and obtain the first two dominant sparse eigenvectors. Here we set $R_0 = 100$ for both eigenvectors. For COCA, we do a normal score transformation on the original dataset. We subsequently project the whole dataset to the first two principal components using the obtained eigenvectors. The according 2-dimension visualization is illustrated in Figure 3. In Figure 3 each black point represents a sample and each red point represents a sample belonging to the corresponding tissue type. It can be observed that, in 2D plots learnt by the COCA, the red points are averagely more dense and more close to the border of the sample cluster. The first phenomenon indicates that the COCA has the potential to preserve more common information shared by samples from the same tissue type. The second phenomenon indicates that the COCA has the potential to differentiate samples from different tissue types more efficiently.

6 Discussion and Comparison with Related Work

A similar principal component analysis procedure is proposed by [7], in which they advocate the use of the transformed Kendall's tau correlation matrix (instead of the Spearman's rho correlation matrix as in the current paper) for estimating the sparse leading eigenvectors. Though both papers are working on principal component analysis, the core ideas are quite different: Firstly, the analysis in [7] is based on a different distribution family called transelliptical, while COCA and Copula PCA are based on the Nonparanormal family. Secondly, by improving the modeling flexibility, in [7] there does not exist a scale-variant variant since it is hard to quantify the transformation functions. In contrast, by introducing the subgaussian transformation function family, the current paper provides sufficient conditions for Copula PCA to achieve parametric rates. Thirdly, the method in [7] cannot explicitly conduct data visualization, due to the fact that the latent elliptical distribution is unspecified and accordingly they cannot accurately estimate the marginal transformations. For Copula PCA, we are able to provide the projection visualization such as in the experiment part of this paper. Moreover, via quantifying a sharp convergence rate in estimating the marginal transformations, we can provide the convergence rates in estimating the principal components. Due to space limit, we refer to the longer version of this paper [6] for more details. Finally, we recommend using the Spearman's rho instead of the Kendall's tau in estimating the correlation coefficients provided that the Nonparanormal model holds. This is because Spearman's rho is statistically more efficient than Kendall'tau within the Nonparanormal family. This research was supported by NSF award IIS-1116730.

References

[1] A.A. Amini and M.J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2454–2458. IEEE, 2008.

[2] T.W Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

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

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

[5] B. Flury. *A first course in multivariate statistics*. Springer Verlag, 1997.

[6] F. Han and H. Liu. High dimensional semiparametric scale-invariant principal component analysis. *Technical Report*, 2012.

[7] F. Han and H. Liu. Tca: Transelliptical principal component analysis for high dimensional non-gaussian data. *Technical Report*, 2012.

[8] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, pages 502–516, 1989.

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

[10] I.T. Jolliffe. *Principal component analysis*, volume 2. Wiley Online Library, 2002.

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

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

[13] J.T. Leek and J.D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

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

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

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

[17] Matthew McCall, Benjamin Bolstad, and Rafael Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11:242–253, 2010.

[18] D. Paul and I.M. Johnstone. Augmented sparse principal component analysis for high dimensional data. *Arxiv preprint arXiv:1202.1242*, 2012.

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

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

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

[22] L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics*, 2012.

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

[24] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

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