Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 7874140, 10 pages
https://doi.org/10.1155/2020/7874140

Research Article

Sparse Principal Component Analysis via Fractional Function Regularity

Xuanli Han,[1,2] Jigen Peng,[3] Angang Cui,[1] and Fujun Zhao[1]

[1]*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China*
[2]*College of Sciences, Xi'an University of Science and Technology, Xi'an 710054, China*
[3]*School of Mathematics and Information Science, Guangzhou University, Guangzhou 510006, China*

Correspondence should be addressed to Jigen Peng; jgpengxjtu@126.com

Received 3 May 2020; Revised 3 July 2020; Accepted 27 July 2020; Published 19 August 2020

Guest Editor: Junpeng Shi

In this paper, we describe a novel approach to sparse principal component analysis (SPCA) via a nonconvex sparsity-inducing fraction penalty function SPCA (FP-SPCA). Firstly, SPCA is reformulated as a fraction penalty regression problem model. Secondly, an algorithm corresponding to the model is proposed and the convergence of the algorithm is guaranteed. Finally, numerical experiments were carried out on a synthetic data set, and the experimental results show that the FP-SPCA method is more adaptable and has a better performance in the tradeoff between sparsity and explainable variance than SPCA.

1. Introduction

Principal component analysis (PCA) [1] has become more popular in data analysis, dimension reduction, image processing, and feature extraction [2]. It seeks the linear combinations of the original variables such that the derived variables capture maximal variance of the total original variables. We can obtain PCA(s) of an original variable by the singular value decomposition (SVD) of the data matrix which is composed of the observed variables. Let $\mathbf{X}$ be an $n \times p$ matrix, where n and p are the number of observations and the number of variables, respectively. Suppose the average value of columns of $\mathbf{X}$ are all zero and the singular value decomposition of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\top}. \tag{1}$$

Then, the columns of matrix $\mathbf{U}\mathbf{D}$ are the principal components of $\mathbf{X}$, and the columns of $\mathbf{V}$ are the corresponding loadings of the principal components. The PCA method is popular due to the following two properties: first, principal components capture the maximum variability from the columns of matrix $\mathbf{X}$, which guarantees minimal information loss; second, principal components are uncorrelated, so we can use one of them without considering the others.

At the same time, however, PCA has its own drawback, i.e., each principal component of matrix $\mathbf{X}$ is a linear combination of all p variables and the elements of the loading vectors (the columns of matrix $\mathbf{V}$) are usually nonzero, which make them difficult to interpret the gained PCs.

Based on the abovementioned priority properties and drawback, some scholars consider it will be a wise option to keep the dimension reduction property at the same time to reduce the number of explainable variables. For this purpose, one method is to set the loading whose absolute value is smaller than a threshold to zero.

The same problem also arises in multiple linear regression, where the response variable is explained by a linear combination of the explainable variables. But, which are the important explainable variables that account for response variable information most? To answer this question, researchers have proposed several approaches, and the first type method is the lasso-based approaches [3–6]. Others, for example, Efron et al. [7] proposed the LARS which is sensitive to noise of samples in 2004, and Friedman et al. [8] put forward a pathwise coordinate optimization approach in 2007. The second type is the lasso-bootstrap methods [9, 10]. Other methods including overview type articles [11–13] and

Radchenko et al. [14] proposed a variable inclusion and shrinkage method, Fan and Li [15] proposed a nonconcave penalized likelihood method, and Candès and Tao [16] proposed a dantzig selector method. Among them, the lasso method [5] is a well-known variable selection technique which produces an accurate and sparse model. As time went on, Zou and Hastie [4] proposed the elastic net in 2005, an adaptive method, which has the advantages of both lasso and ridge regression. The elastic net approach can be considered as a generalization of lasso and ridge regression. In 2006, Zou et al. [3] proposed sparse principal analysis (SPCA) using the lasso and ridge regression to produce a modified principal component with sparse loadings. They show that PCA can be formulated as a regression-type optimization problem and sparse loadings are obtained by imposing l_1-norm and l_2-norm regularization on the regression coefficients of the model. In 2009, Leng and Wang [17] proposed a method of simple adaptive sparse principal component analysis, which uses the adaptive lasso penalty instead of the lasso penalty in SPCA. Moreover, they replace the least-squares objective function in SPCA by a general least-squares objective function, which leads to study many related SPCA estimators under a unified theoretical framework and get the method of general adaptive sparse principal component analysis. SPCA can produce a sparse model that has only fewer active coefficients, and the other coefficients are all set to zero. The proposed sparse model SPCA is more interpretable than PCA. In a word, SPCA can specifically identify structure information of the data matrix.

The idea of making the model's coefficients sparse is not a new one. Jollife et al. [18] first proposed a SPCA method using l_1-norm regularization which leads to a variety of sparsity-inducing methods. The success of SPCA in gaining more interpretable model motivates us to propose the method in this paper, where we proposed a fractional function penalty method with respect to sparse PCA. The method is to exploit the fractional function [19] to penalize the model coefficients in order to obtain sparse coefficients, which makes the PCA to have more interpretable ability.

The rest of the paper is organized as follows. In Section 2, definition and derivation of principal components are reviewed. The sparse principal component is presented in Section 3. In Section 4, we propose the fractional functional penalized principal components. Numerical experiments and conclusions are formulated in Sections 5 and 6, respectively.

In this paper, without specification, scalar is denoted as lower case letter x, and vector is denoted as bold lower case letter $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$. The matrix is denoted by bold capital letter $\mathbf{M}$, and $\mathbf{I}$ denotes the identity matrix. The transpose of a real matrix $\mathbf{M}$ is denoted by $\mathbf{M}^\top$. The Frobenius norm of $\cdot$ is denoted by $\|\cdot\|$, where $\cdot$ denotes column vector or matrix, and the spectral norm of a matrix $\mathbf{M}$ is denoted by $\|\mathbf{M}\|_2$.

2. Principal Components and Sparse Principal Components

2.1. Definition of Derivation of Principal Components. Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ be a vector of p random variables, and the variances of the p random variables and the covariance between the p variables will be of importance. If p is large, an alternative key approach is to seek for fewer variables that preserve well most of the information provided by variances and covariances of these p variables. One of the alternative techniques is principle component analysis (PCA) which concerns more variance of variables than covariance. PCA first looks for a linear function $\mathbf{x}\alpha_1$ owning to maximum variance of $\mathbf{x}$, where $\alpha_1 = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p})^\top$, that is,

$$\mathbf{x}\boldsymbol{\alpha}_1 = \alpha_{11}\mathrm{x}_1 + \alpha_{12}\mathrm{x}_2 + \cdots + \alpha_{1\mathrm{p}}\mathrm{x}_p. \tag{2}$$

Then, in the same way, we can find a linear function $\mathbf{x}\alpha_2$, in which $\mathbf{x}\alpha_2$ and $\mathbf{x}\alpha_1$ are linear independence, so that at the k-th step, a linear function $\mathbf{x}\alpha_k$ can be found which has maximum variance and is linear independence with the former $k-1$ linear functions. The $\mathbf{x}\alpha_k$ is the k-th principal component (PC) [1]. Many of the estimation problems in array signal processing can also start with the same issue in (2); however, there may be exist some perturbations [20–22]. Having known what are the PCs, the remaining question is how to find them. If the vector of random variables $\mathbf{x}$ has a known covariance matrix Σ whose (i, j)th entry is the known covariance of random variables x_i and x_j when $i \neq j$, and variance of the element x_i of $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ when $i = j$. When Σ is unknown, the more realistic method is to substitute Σ for sample covariance matrix $\mathbf{S}$.

In fact, because Σ is often unknown, we consider the case $\mathbf{S}$ in the following only.

To obtain the PCs of the sampling matrix $\mathbf{X}_{n\times p}$, where n and p are the number of observations and variables of random vector $\mathbf{x}$, respectively. Consider the coefficient vector $\alpha_1 \in \mathbb{R}^p$, which is forced to maximize the sample variance of $\mathbf{X}\alpha_1$, where α_1 is normalized. Let $\mathbf{x}_i$ denote the i-th column of sampling matrix $\mathbf{X}$, and $\mathbf{x}_i$ is zero-centered, $i = 1, 2, \ldots, p$. The problem can be reformulated as the follows:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}_1} \boldsymbol{\alpha}_1^\top (\mathbf{X}^\top\mathbf{X})\boldsymbol{\alpha}_1 \\ &\text{subject to} \qquad \boldsymbol{\alpha}_1^\top\boldsymbol{\alpha}_1 = 1, \end{aligned} \tag{3}$$

and solving (3), we can obtain the relationship $(\mathbf{X}^\top\mathbf{X})\alpha_1 = \lambda_1\alpha_1$; furthermore, we have

$$\boldsymbol{\alpha}_1^\top (\mathbf{X}^\top\mathbf{X})\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^\top\lambda_1\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1^\top\boldsymbol{\alpha}_1 = \lambda_1, \tag{4}$$

where λ_1 is the largest eigenvalue of matrix $\mathbf{X}^\top\mathbf{X}$ and α_1 is the eigenvector of $\mathbf{X}^\top\mathbf{X}$ corresponding to λ_1. $\mathbf{X}\alpha_1$ is called the 1-th PC of sampling matrix $\mathbf{X}$.

In the same way, the k-th PC of $\mathbf{X}$ is α_k and the sample variance of $\mathbf{X}\alpha_k$ is λ_k, where λ_k is the k-th largest eigenvalue of $\mathbf{X}^\top\mathbf{X}$ and α_k is the corresponding eigenvector.

In general, if $n \times p$ matrix $\mathbf{X}$ is known, $\mathbf{x}_i$, the i-th column of $\mathbf{X}$, is zero-centered. The singular value decomposition of $\mathbf{X}$ is $\mathbf{U}\mathbf{D}\mathbf{V}^\top$. Then, the principal components of matrix $\mathbf{X}$, which are denoted by $\mathbf{Y}$, are $(\mathbf{U}\mathbf{D})_r$, where $(\mathbf{U}\mathbf{D})_r$ is the matrix composed of the first r columns of the matrix $\mathbf{U}\mathbf{D}$ and r is the rank of $\mathbf{X}$. So, we can get $\mathbf{y}_i = \mathbf{X}\mathbf{v}_i$ [1, 3], where $\mathbf{y}_i$ and $\mathbf{v}_i$ are ith columns of matrix $\mathbf{X}$ and $\mathbf{V}$, respectively.

Although principal component analysis is widely used in various technical fields, it has an obvious drawback, that is,

each principal component is a linear combination of all original variables, but it is often difficult to interpret the result. To solve this problem, Zou et al. [3] proposed sparse principal component analysis. They first consider PCA as a linear regression question.

2.2. Preparation of Sparse Principal Components

2.2.1. Linear Regression and LASSO. Consider matrix $\mathbf{X}_{n\times p}$ with n observations of the p variables, and let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ be the response vector and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ the predictor variables, where $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^\top$ is the j-th column of matrix $\mathbf{X}$ $(j = 1, 2, \ldots, p)$. Assume $\mathbf{y}$ and $\mathbf{x}_j$ have zero means. The question of evaluating of PCA of matrix $\mathbf{X}$ can be considered as a linear regression problem:

$$\min_{\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)^\top} \left\| \mathbf{y} - \sum_{j=1}^{p} \beta_j \mathbf{x}_j \right\|^2 = \min_{\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)^\top} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2. \tag{5}$$

This model is simple, and the result is easy to get in some cases; however, the result is difficult to interpret [3]. Based on model (5), Tibshirani [5] proposed the lasso method:

$$\widehat{\beta}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \beta_j \mathbf{x}_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{6}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ and $\lambda \geq 0$.

Model (6) can produce accurate and sparse result at the same time, which strengths the result's interpretable ability. However, the lasso method has some limitations, and the most relevant one is that the number of variables produced by the lasso is limited by the number of observations. Given $\mathbf{X}_{n\times p}$, when the number of rows n and number of columns p of $\mathbf{X}$ satisfy $p > n$, the lasso can select n explainable variables at most.

To overcome the abovementioned drawback, Zou and Hastie [4] proposed the elastic net model:

$$\widehat{\beta}_{\text{en}} = (1 + \lambda_2) \left\{ \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \beta_j \mathbf{x}_j \right\|^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\}, \tag{7}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$, $\lambda_1 \geq 0$, and $\lambda_2 \geq 0$.

The elastic net model shares the properties of lasso and ridge regression at the same time because when $\lambda_2 = 0$ the elastic is the lasso, and when $\lambda_1 = 0$, it becomes the ridge regression. When $p > n$, by choosing proper λ_2, the elastic net model can potentially include all variables.

2.2.2. Ridge Regression and the Naive Elastic Net. In model (6) and (7), the sparse coefficients are all l_1-norm-induced and the sparsity is none of l_2-norm's business. Jolliffe et al. [18] proposed the SCoTLASS method:

$$\boldsymbol{\alpha}_i^\top (\mathbf{X}^\top \mathbf{X}) \boldsymbol{\alpha}_i \quad \text{s.t.} \quad \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_i = 1, \quad \boldsymbol{\alpha}_l^\top \boldsymbol{\alpha}_i = 0 \, (i \geq 2, l < i), \quad \sum_{j=1}^{p} |\alpha_{ij}| \leq t, \tag{8}$$

where $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ip})^\top$ and t is a parameter. This method has the ability to obtain sparse loadings by imposing l_1-norm penalty on principle component analysis (PCA).

The SCoTLASS method has two drawbacks. One is difficulty of choosing parameter t and another problem is the high computational cost. So, Zou et al. [3] consider a modified method. They first show that principle component analysis (PCA) can be expressed as a ridge regression problem.

Suppose the singular value decomposition of $\mathbf{X}_{n\times p}$ is $\mathbf{U}\mathbf{D}\mathbf{V}^\top$. Let $\mathbf{U}_i$ and $\mathbf{D}_{ii}$ be the i-th column of matrix $\mathbf{U}$ and the i-th element of diagonal matrix $\mathbf{D}$, respectively, and then, $\mathbf{D}_{ii}\mathbf{U}_i$ is the i-th principal component of matrix $\mathbf{X}_{n\times p}$, $i = 1, 2, \ldots, r$, $r = \text{rank}(\mathbf{X}_{n\times p})$. The ridge estimation of β is given by the following expression:

$$\widehat{\beta}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{D}_{ii}\mathbf{U}_i - \mathbf{X}_{n\times p}\boldsymbol{\beta} \right\|^2 + \lambda \|\boldsymbol{\beta}\|^2. \tag{9}$$

Let $\widehat{\mathbf{v}} = (\widehat{\beta}_{\text{ridge}} / \|\widehat{\beta}_{\text{ridge}}\|)$, and then, $\widehat{\mathbf{v}} = \mathbf{v}_i$, where $\lambda > 0$ and $\mathbf{v}_i$ is the i-th column of $\mathbf{V}$.

This conclusion shows the relationship between principal component analysis and the regression method. The term $\lambda\|\beta\|^2$ in (9) is to guarantee the unique of $\widehat{\beta}_{\text{ridge}}$ when the inverse of $\mathbf{X}^\top\mathbf{X}$ does not exist. Furthermore, Zou and Hastie [4] proposed a method to add l_1-norm penalty to model (9) and got the following optimization problem:

$$\widehat{\beta}_{\text{nen}} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{D}_{ii}\mathbf{U}_i - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1, \tag{10}$$

They call the $\widehat{\mathbf{v}}_i = (\widehat{\beta}_{\text{nen}} / \|\widehat{\beta}_{\text{nen}}\|)$ is an approximation of $\mathbf{v}_i$, and $\mathbf{X}\widehat{\mathbf{v}}_i$ is the i th approximated principal component. To distinguish this model from model (7), model (10) is called the naive elastic net.

2.3. Sparse Principle Component Based on PCA.

Based on the general conclusion in the end of Section 2.1, Zou et al. [3] proposed a two-stage analysis: perform PCs first, and then, find suitable spare approximation of PCs.

Consider the first r principal components of $\mathbf{X}_{n\times p}$, where $r = \text{rank}(\mathbf{X}_{n\times p})$. Let $\mathbf{A}_{p\times r} = (\alpha_1, \alpha_2, \ldots, \alpha_r)$, $\mathbf{B}_{p\times r} = (\beta_1, \beta_2, \ldots, \beta_r)$, and

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \left\| \mathbf{X}^\top - \mathbf{A}\mathbf{B}^\top\mathbf{X}^\top \right\|^2 + \lambda \|\mathbf{B}\|^2 \quad \text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I}_r, \tag{11}$$

for any $\lambda > 0$. Then, $\widehat{\beta}_i \propto \mathbf{v}_i, i = 1, 2, \ldots, r$, where $\widehat{\beta}_i$ is the estimation of β_i.

Then, we can perform SPCA, which is based on the PCA and regression method, by adding l_1-norm penalty to the regression coefficients to induce sparse loadings. Thus, the following result is obtained:

$$
\begin{aligned}
(\widehat{\mathbf{A}},\widehat{\mathbf{B}}) &= \arg\min_{\mathbf{A},\mathbf{B}} \left\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top}\right\|^{2}+\lambda\|\mathbf{B}\|^{2}+\sum_{j=1}^{r}\lambda_{1,j}\left\|\boldsymbol{\beta}_{j}\right\|_{1},\\
&= \arg\min_{\mathbf{A},\mathbf{B}} \left\|\mathbf{X}^{\top}-\mathbf{A}\mathbf{B}^{\top}\mathbf{X}^{\top}\right\|^{2}+\lambda\|\mathbf{B}\|^{2}+\sum_{j=1}^{r}\lambda_{1,j}\left\|\boldsymbol{\beta}_{j}\right\|_{1},\\
&= \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n}\left\|\mathbf{x}_{i}-\mathbf{A}\mathbf{B}^{\top}\mathbf{x}_{i}\right\|^{2}+\lambda\sum_{j=1}^{r}\left\|\boldsymbol{\beta}_{j}\right\|^{2}+\sum_{j=1}^{r}\lambda_{1,j}\left\|\boldsymbol{\beta}_{j}\right\|_{1},
\end{aligned}
\tag{12}
$$

subject to $\mathbf{A}^{\top}\mathbf{A}=\mathbf{I}_{r}$. Where $\mathbf{x}_{i}$ is the i th column of matrix $\mathbf{X}^{\top}$ and β_{j} is the j th column of matrix $\mathbf{B}$, $r=\operatorname{rank}(\mathbf{X})$.

3. Fractional Function-Penalized Sparse Principal Components

3.1. Fraction Function-Penalized Method. Recently, Li et al. [19] proposed a fraction-penalized function. They study the problem of a nonconvex sparsity promoting penalty function,

$$
P_{a}(\mathbf{x})=\sum_{i=1}^{n}p_{a}(x_{i})=\sum_{i=1}^{n}\frac{a|x_{i}|}{1+a|x_{i}|},
\tag{13}
$$

to relax the l_{0}-norm in compressed sensing [15, 16, 23, 24] to get the sparsity signal, where $a>0$ is a parameter and $\mathbf{x}=(x_{1},x_{2},\ldots,x_{n})^{\top}$. $P_{a}(\mathbf{x})$ is a fraction function which is nonconvex and sparse promoting. They studied the properties of this fractional function and derived a closed form expression of the optimal solution to the fractional function penalty problem and proposed an iterative thresholding algorithm to solve the fraction function penalty problem. A series of experiments have been conducted and the results show, compared with the soft thresholding algorithm and the half thresholding algorithms, that the proposed algorithm has the property of better sparse signal recovery with and without measurement noise.

Based on the good performance of the fraction function in sparse promotion in compressed sensing [23, 25], in this paper, we shall adopt this method to the sparse principal component analysis (SPCA), i.e., replace the l_{1}-norm penalty in SPCA with the fraction function penalty.

Combining fd12(12) and (13)fd13, we can get

$$
\begin{aligned}
(\widehat{\mathbf{A}},\widehat{\mathbf{B}}) &= \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n}\left\|\mathbf{x}_{i}-\mathbf{A}\mathbf{B}^{\top}\mathbf{x}_{i}\right\|^{2}\\
&\quad+\lambda\sum_{j=1}^{r}\left\|\boldsymbol{\beta}_{j}\right\|^{2}+\sum_{j=1}^{r}\lambda_{1,j}P_{a}\left(\boldsymbol{\beta}_{j}\right),\\
&\text{subject to}\quad \mathbf{A}^{\top}\mathbf{A}=I_{r}.
\end{aligned}
\tag{14}
$$

We now introduce the following fact for the purpose of solving the problem (14).

$$
\sum_{i=1}^{n}\left\|\mathbf{x}_{i}-\mathbf{A}\mathbf{B}^{\top}\mathbf{x}_{i}\right\|^{2}=\left\|\mathbf{X}^{\top}-\mathbf{A}\mathbf{B}^{\top}\mathbf{X}^{\top}\right\|^{2}=\left\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top}\right\|^{2}.
\tag{15}
$$

Owing to $\mathbf{A}^{\top}\mathbf{A}=I_{r}$, let $\mathbf{A}_{\perp}$ be the matrix satisfying $(\mathbf{A}\,|\,\mathbf{A}_{\perp})$ a $p\times p$ orthonormal matrix, which $(\mathbf{A}\,|\,\mathbf{A}_{\perp})$ is a block matrix. Let $\mathbf{Q}$: $=(\mathbf{A}\,|\,\mathbf{A}_{\perp})$. Then,

$$
\begin{aligned}
\left\|(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\mathbf{Q}\right\|^{2} &= \operatorname{tr}\left[\left((\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\mathbf{Q}\right)^{\top}(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\mathbf{Q}\right],\\
&= \operatorname{tr}\left[\mathbf{Q}^{\top}(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})^{\top}(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\mathbf{Q}\right],\\
&= \operatorname{tr}\left[(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})^{\top}(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\right],\\
&= \left\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top}\right\|^{2},
\end{aligned}
\tag{16}
$$

while

$$
\begin{aligned}
\left\|(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\mathbf{Q}\right\|^{2} &= \left\|(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})(\mathbf{A}\,|\,\mathbf{A}_{\perp})\right\|^{2},\\
&= \left\|(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\mathbf{A}\,\middle|\,(\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top})\mathbf{A}_{\perp}\right\|^{2},\\
&= \left\|(\mathbf{X}\mathbf{A}-\mathbf{X}\mathbf{B})\,|\,\mathbf{X}\mathbf{A}_{\perp}\right\|^{2},\\
&= \|\mathbf{X}\mathbf{A}-\mathbf{X}\mathbf{B}\|^{2}+\left\|\mathbf{X}\mathbf{A}_{\perp}\right\|^{2},
\end{aligned}
\tag{17}
$$

so

$$
\begin{aligned}
\left\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top}\right\|^{2} &= \|\mathbf{X}\mathbf{A}-\mathbf{X}\mathbf{B}\|^{2}+\left\|\mathbf{X}\mathbf{A}_{\perp}\right\|^{2},\\
&= \sum_{j=1}^{r}\left\|\mathbf{X}\boldsymbol{\alpha}_{j}-\mathbf{X}\boldsymbol{\beta}_{j}\right\|^{2}+\left\|\mathbf{X}\mathbf{A}_{\perp}\right\|^{2}.
\end{aligned}
\tag{18}
$$

So, problem (14) can be transformed into

$$
\begin{aligned}
(\widehat{\mathbf{A}},\widehat{\mathbf{B}}) &= \arg\min_{\mathbf{A},\mathbf{B}} \sum_{j=1}^{r}\left\|\mathbf{X}\boldsymbol{\alpha}_{j}-\mathbf{X}\boldsymbol{\beta}_{j}\right\|^{2}+\left\|\mathbf{X}\mathbf{A}_{\perp}\right\|^{2}\\
&\quad+\lambda\sum_{j=1}^{r}\left\|\boldsymbol{\beta}_{j}\right\|^{2}+\sum_{j=1}^{r}\lambda_{1,j}P_{a}\left(\boldsymbol{\beta}_{j}\right),\\
&\text{subject to}\quad \mathbf{A}^{\top}\mathbf{A}=I_{r}.
\end{aligned}
\tag{19}
$$

Problem (19) can be solved by an alternating algorithm. If $\mathbf{A}$ is given, $\widehat{\mathbf{B}}=(\widehat{\beta}_{1},\widehat{\beta}_{2},\ldots,\widehat{\beta}_{r})$ can be estimated by

$$
\widehat{\boldsymbol{\beta}}_{j}=\arg\min_{\boldsymbol{\beta}_{j}}\left\|\mathbf{X}\boldsymbol{\alpha}_{j}-\mathbf{X}\boldsymbol{\beta}_{j}\right\|^{2}+\lambda\left\|\boldsymbol{\beta}_{j}\right\|^{2}+\lambda_{1,j}P_{a}\left(\boldsymbol{\beta}_{j}\right),
\tag{20}
$$

and if $\mathbf{B}$ is given, problem (14) is to solve

$$
\begin{aligned}
\widehat{\mathbf{A}} &= \arg\min_{\mathbf{A}}\sum_{i=1}^{n}\left\|\mathbf{x}_{i}-\mathbf{A}\mathbf{B}^{\top}\mathbf{x}_{i}\right\|^{2},\\
&= \arg\min_{\mathbf{A}}\left\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^{\top}\right\|^{2}\ \text{subject to}\quad \mathbf{A}^{\top}\mathbf{A}=I_{r},
\end{aligned}
\tag{21}
$$

where $\mathbf{x}_{i}$ is the i-th column of matrix $\mathbf{X}^{\top}$. According to the procrustes rotation [26–28], the estimation of $\mathbf{A}$ can be obtained; suppose the SVD of matrix $(\mathbf{X}^{\top}\mathbf{X})\mathbf{B}$ is $\mathbf{U}_{1}\mathbf{D}_{1}(\mathbf{V}_{1})^{\top}$, by $\mathbf{U}_{1}(\mathbf{V}_{1})^{\top}$ [3], that is, $\widehat{\mathbf{A}}=\mathbf{U}_{1}(\mathbf{V}_{1})^{\top}$.

3.2. Transformation of Problem (16). In (20), let $\mathbf{A}=(\alpha_{1},\alpha_{2},\ldots,\alpha_{r})$ be given by $\alpha_{j}=\mathbf{v}_{j}$ for $j=1,2,\ldots,r$, where $\mathbf{v}_{i}$ is the i-th column of $\mathbf{V}$-the matrix in the $\mathbf{U}\,\mathbf{D}\mathbf{V}^{\top}$

which is the singular decomposition of the matrix $\mathbf{X}$. Let $\mathbf{y}_j = \mathbf{X}\mathbf{v}_j$, and (22) can be reformulated as

$$\widehat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} \left\|\mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j\right\|^2 + \lambda\left\|\boldsymbol{\beta}_j\right\|^2 + \lambda_{1,j}P_a\left(\boldsymbol{\beta}_j\right). \quad (22)$$

Let

$$\mathbf{X}^* = \frac{1}{\sqrt{1+\lambda}}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\,\mathbf{I}\end{pmatrix}_{(n+p)\times p},$$
$$\mathbf{y}_j^* = \begin{pmatrix}\mathbf{y}_j\\ \mathbf{0}\end{pmatrix}_{(n+p)\times 1}, \quad (23)$$

$\gamma_j = (\lambda_{1,j}/\sqrt{1+\lambda})$ and $\beta_j^* = \sqrt{1+\lambda}\beta_j$. Then, (22) can be rewritten as

$$\widehat{\boldsymbol{\beta}}_j^* = \arg\min_{\boldsymbol{\beta}_j^*} \left\|\mathbf{y}_j^* - \mathbf{X}^*\boldsymbol{\beta}_j^*\right\|^2 + \gamma_j P_a\left(\boldsymbol{\beta}_j^*\right). \quad (24)$$

The relationship of $\widehat{\beta_j^*}$ and $\widehat{\beta}_j$ can be explained as the following. Let

$$F\left(\boldsymbol{\beta}_j\right) = \left\|\mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j\right\|^2 + \lambda\left\|\boldsymbol{\beta}_j\right\|^2 + \lambda_{1,j}P_a\left(\boldsymbol{\beta}_j\right),$$
$$H\left(\boldsymbol{\beta}_j\right) = G\left(\boldsymbol{\beta}_j^*\right) = \left\|\mathbf{y}_j^* - \mathbf{X}^*\boldsymbol{\beta}_j^*\right\|^2 + \gamma_j P_a\left(\boldsymbol{\beta}_j^*\right). \quad (25)$$

According to the expressions of $\mathbf{X}^*, \mathbf{y}_j^*, \gamma_j, \beta_j^*$ and the definition of $P_a(\beta_j)$, we can get

$$F\left(\boldsymbol{\beta}_j\right) = \left\|\mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j\right\|^2 + \lambda\left\|\boldsymbol{\beta}_j\right\|^2 + \lambda_{1,j}\sum_{i=1}^{p}\frac{a\left|\beta_{ji}\right|}{1+a\left|\beta_{ji}\right|},$$

$$H\left(\boldsymbol{\beta}_j\right) = \left\|\mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j\right\|^2 + \lambda\left\|\boldsymbol{\beta}_j\right\|^2 + \frac{\lambda_{1,j}}{\sqrt{1+\lambda}}\sum_{i=1}^{p}\frac{a\left|\sqrt{1+\lambda}\,\beta_{ji}\right|}{1+a\left|\sqrt{1+\lambda}\,\beta_{ji}\right|},$$

$$= \left\|\mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j\right\|^2 + \lambda\left\|\boldsymbol{\beta}_j\right\|^2 + \lambda_{1,j}\sum_{i=1}^{p}\frac{a\left|\beta_{ji}\right|}{1+a\sqrt{1+\lambda}\left|\beta_{ji}\right|}. \quad (26)$$

Owing to $\lambda > 0$, it is easy to obtain that $H(\beta_j) < F(\beta_j)$, so that

$$\min_{\boldsymbol{\beta}_j} H\left(\boldsymbol{\beta}_j\right) < \min_{\boldsymbol{\beta}_j} F\left(\boldsymbol{\beta}_j\right), \quad (27)$$

that is,

$$\min_{\boldsymbol{\beta}_j^*} G\left(\boldsymbol{\beta}_j^*\right) < \min_{\boldsymbol{\beta}_j} F\left(\boldsymbol{\beta}_j\right) 0. \quad (28)$$

Both $G(\beta_j^*)$ and $H(\beta_j)$ are continuous functions with respect to β_j^* and β_j, respectively, so as $\lambda \longrightarrow 0$, we have $\beta_j^* \longrightarrow \beta_j$, and further, we get $G(\beta_j^*) \longrightarrow G(\beta_j) \longrightarrow F(\beta_j)$. That is, $G(\beta_j^*) \longrightarrow F(\beta_j)$ as $\lambda \longrightarrow 0$.

So, we conclude that $\widehat{\beta}_j$ can be approximated well by $\widehat{\beta_j^*}$ when positive λ is small enough.

Remark 1. The reason why we substitute $F(\beta_j)$ with $G(\beta_j^*)$ is model (24) is easier to solve than model (22).

Now, we will show that the optimal solution to the minimization problem,

$$\min_{\boldsymbol{\beta}_j^*} \left\|\mathbf{y}_j^* - \mathbf{X}^*\boldsymbol{\beta}_j^*\right\|^2 + \gamma_j P_a\left(\boldsymbol{\beta}_j^*\right), \quad (29)$$

can be expressed as a thresholding operation.

3.3. The Algorithm of the Problem (29). For any $\gamma_j, \mu \in (0, +\infty)$, and $\mathbf{z} \in \mathbb{R}^P$, let

$$C_{\gamma_j}\left(\boldsymbol{\beta}_j^*\right) = \left\|\mathbf{y}_j^* - \mathbf{X}^*\boldsymbol{\beta}_j^*\right\|^2 + \gamma_j P_a\left(\boldsymbol{\beta}_j^*\right),$$
$$C_\mu\left(\boldsymbol{\beta}_j^*, \mathbf{z}\right) = \mu\left(C_{\gamma_j}\left(\boldsymbol{\beta}_j^*\right) - \left\|\mathbf{X}^*\boldsymbol{\beta}_j^* - \mathbf{X}^*\mathbf{z}\right\|^2\right) + \left\|\boldsymbol{\beta}_j^* - \mathbf{z}\right\|^2,$$
$$B_\mu\left(\boldsymbol{\beta}_j^*\right) = \boldsymbol{\beta}_j^* + \mu\left(\mathbf{X}^*\right)^\top\left(\mathbf{y}_j^* - \mathbf{X}^*\boldsymbol{\beta}_j^*\right). \quad (30)$$

We have the following results.

Lemma 1. *For any fixed parameter $\mu, a, \gamma_j \in (0, +\infty)$ and $\mathbf{z} \in \mathbb{R}^P$, if $\overline{\beta}_j^* = (\overline{\beta}_{j1}^*, \overline{\beta}_{j2}^*, \ldots, \overline{\beta}_{jp}^*)^\top$ is a local minimizer to $C_\mu(\beta_j^*, \mathbf{z})$, then*

$$\overline{\beta}_{ji}^* = 0 \Longleftrightarrow \left|\left(B_\mu(\mathbf{z})\right)_i\right| \le t,$$
$$\overline{\beta}_{ji}^* = g_{\gamma_j\mu}\left(\left(B_\mu(\mathbf{z})\right)_i\right) \Longleftrightarrow \left|\left(B_\mu(\mathbf{z})\right)_i\right| > t, \quad (31)$$

where $i = 1, 2, \ldots, p$,

$$t = \begin{cases}\dfrac{\gamma_j}{2}a, & \text{if } \gamma_j \le \dfrac{1}{a^2},\\ \sqrt{\gamma_j} - \dfrac{1}{2a}, & \text{if } \gamma_j > \dfrac{1}{a^2},\end{cases} \quad (32)$$

and according to [19],

$$\phi(s) = \arccos\left(\frac{27\gamma_j\mu a^2}{4\left(1+a|s|\right)^3} - 1\right),$$
$$g_{\gamma_j\mu}(s) = \operatorname{sgn}(s)\left(\frac{(1+a|s|)/3\left(1+2\cos\left((\phi(|s|)/3) - (\pi/3)\right)\right) - 1}{a}\right). \quad (33)$$

Proof. We can see that $C_\mu(\beta_j^*, \mathbf{z})$ can be expressed as

$$\begin{aligned} C_\mu\left(\boldsymbol{\beta}_j^*, \mathbf{z}\right) &= \left\|\boldsymbol{\beta}_j^* - \left(\mathbf{z} - \mu\left(\mathbf{X}^*\right)^\top\mathbf{X}^*\mathbf{z} + \mu\left(\mathbf{X}^*\right)^\top\mathbf{y}_j^*\right)\right\|^2 \\ &\quad + \gamma_j\mu P_a\left(\boldsymbol{\beta}_j^*\right) + \mu\left\|\mathbf{y}_j^*\right\|^2 + \|\mathbf{z}\|^2 - \mu\left\|\mathbf{X}^*\mathbf{z}\right\|^2 \\ &\quad - \left\|\mathbf{z} - \mu\left(\mathbf{X}^*\right)^\top\mathbf{X}^*\mathbf{z} + \mu\left(\mathbf{X}^*\right)^\top\mathbf{y}_j^*\right\|^2, \\ &= \left\|\boldsymbol{\beta}_j^* - B_\mu(\mathbf{z})\right\|^2 + \gamma_j\mu P_a\left(\boldsymbol{\beta}_j^*\right) + \mu\left\|\mathbf{y}_j^*\right\|^2 + \|\mathbf{z}\|^2 \\ &\quad - \mu\left\|\mathbf{X}^*\mathbf{z}\right\|^2 - \left\|B_\mu(\mathbf{z})\right\|^2, \end{aligned} \quad (34)$$

that is to say, minimizing $C_\mu(\beta_j^*, \mathbf{z})$ for $\mathbf{z}$, given μ, γ_j, and a, is equivalent to

$$\min_{\boldsymbol{\beta}_j^* \in R^p} \sum_{i=1}^{p} \left(\beta_{ji}^* - \left(B_\mu(\mathbf{z})\right)_i\right)^2 + \gamma_j \mu \sum_{i=1}^{p} p_a\left(\beta_{ji}^*\right). \quad (35)$$

So, $\overline{\beta}_j^* = (\overline{\beta}_{j1}^*, \overline{\beta}_{j2}^*, \ldots, \overline{\beta}_{jp}^*)^\top$ is a local minimizer of $C_\mu(\beta_j^*, \mathbf{z})$ if and only if $\overline{\beta}_{ji}^*$ solves the following problem:

$$\min_{\beta_{ji}^* \in R} \left(\beta_{ji}^* - \left(B_\mu(\mathbf{z})\right)_i\right)^2 + \gamma_j \mu p_a\left(\beta_{ji}^*\right). \quad (36)$$

According to Lemma 10 in [19], we complete the proof. Furthermore, we can get Theorem 1. □

Theorem 1. *If $\widetilde{\beta}_j^* = (\widetilde{\beta}_{j1}^*, \widetilde{\beta}_{j2}^*, \ldots, \widetilde{\beta}_{jp}^*)^\top$ is an optimal solution to (29), and γ_j is positive value and μ satisfies $0 < \mu \le \|\mathbf{X}^*\|_2^{-2}$, and then, the optimal solution $\widetilde{\beta}_{ji}^*$ can be given by*

$$\widetilde{\beta}_{ji}^* = \begin{cases} g_{\gamma_j \mu}\left(\left(B_\mu\left(\widetilde{\boldsymbol{\beta}}_j^*\right)\right)_i\right), & \text{if } \left|\left(B_\mu\left(\widetilde{\boldsymbol{\beta}}_j^*\right)\right)_i\right| > t, \\ 0, & \text{if } \left|\left(B_\mu\left(\widetilde{\boldsymbol{\beta}}_j^*\right)\right)_i\right| \le t, \end{cases} \quad (37)$$

where $i = 1, 2, \ldots, p$,

$$t = \begin{cases} \dfrac{\gamma_j \mu a}{2}, & \text{if } \gamma_j \le \dfrac{1}{a^2 \mu}, \\ \sqrt{\gamma_j \mu} - \dfrac{1}{2a}, & \text{if } \gamma_j > \dfrac{1}{a^2 \mu}. \end{cases} \quad (38)$$

Proof. According to $0 < \mu \le \|\mathbf{X}^*\|_2^{-2}$, we have

$$\begin{aligned} C_\mu\left(\boldsymbol{\beta}_j^*, \widetilde{\boldsymbol{\beta}}_j^*\right) &= \mu\left(\left\|\mathbf{y}_j^* - \mathbf{X}^* \boldsymbol{\beta}_j^*\right\|^2 + \gamma_j P_a\left(\boldsymbol{\beta}_j^*\right) - \left\|\mathbf{X}^* \boldsymbol{\beta}_j^* - \mathbf{X}^* \widetilde{\boldsymbol{\beta}}_j^*\right\|^2\right) + \left\|\boldsymbol{\beta}_j^* - \widetilde{\boldsymbol{\beta}}_j^*\right\|^2, \\ &= \mu\left(\left\|\mathbf{y}_j^* - \mathbf{X}^* \boldsymbol{\beta}_j^*\right\|^2 + \gamma_j P_a\left(\boldsymbol{\beta}_j^*\right)\right) - \mu\left\|\mathbf{X}^*\left(\boldsymbol{\beta}_j^* - \widetilde{\boldsymbol{\beta}}_j^*\right)\right\|^2 + \left\|\boldsymbol{\beta}_j^* - \widetilde{\boldsymbol{\beta}}_j^*\right\|^2 \\ &\ge \mu\left(\left\|\mathbf{y}_j^* - \mathbf{X}^* \boldsymbol{\beta}_j^*\right\|^2 + \gamma_j P_a\left(\boldsymbol{\beta}_j^*\right)\right) \\ &\ge \mu\left(\left\|\mathbf{y}_j^* - \mathbf{X}^* \widetilde{\boldsymbol{\beta}}_j^*\right\|^2 + \gamma_j P_a\left(\widetilde{\boldsymbol{\beta}}_j^*\right)\right), \\ &= C_\mu\left(\widetilde{\boldsymbol{\beta}}_j^*, \widetilde{\boldsymbol{\beta}}_j^*\right). \end{aligned} \quad (39)$$

That is to say, for $\beta_j^* \in R^p$, $\widetilde{\beta}_j^*$ is a local minimizer of $C_\mu(\beta_j^*, \widetilde{\beta}_j^*)$ as long as $\widetilde{\beta}_j^*$ is a solution to (29). According to Lemma 10 in [19] and Lemma 1, we complete the proof.

With the expression of $\widetilde{\beta}_j^*$ in Theorem 1, we can get an alternating thresholding algorithm for problem (29), that is,

$$\left(\left(\boldsymbol{\beta}_j^*\right)^{n+1}\right)_i = g_{\gamma_j \mu}\left(\left(B_\mu\left(\boldsymbol{\beta}_j^*\right)^n\right)_i\right), \quad (40)$$

for $i = 1, 2, \ldots, p$, where $B_\mu(\beta_j^*) = \beta_j^* + \mu(\mathbf{X}^*)^\top(\mathbf{y}_j^* - \mathbf{X}^* \beta_j^*)$ and $g_{\gamma_j \mu}(\cdot)$ is the thresholding operator in Lemma 1. We call this algorithm the alternating iterative FP-SPCA thresholding algorithm. □

Remark 2. The dot $\cdot$ in $g_{\gamma_j \mu}(\cdot)$ may be a real number as in Lemma 1 or a vector whose components are all real numbers. If the dot $\cdot$ represents a vector, $g_{\gamma_j \mu}(\cdot)$ means the result vector which components are the result that $g_{\gamma_j \mu}(\cdot)$ acts on every component of the vector $\cdot$ sequentially.

It is well known that the solution to problem (29) depends on the choice of the regularization parameter γ_j. Before we give out the proper choice for the regularization parameter γ_j, we need the following definition.

Definition 1. The nonincreasing rearrangement of the vector $\overline{\beta}_j \in \mathbb{R}^p$ is the vector $[\overline{\beta}_j] \in \mathbb{R}^p$ for which

$$\left[\overline{\boldsymbol{\beta}}_j\right]_1 \ge \left[\overline{\boldsymbol{\beta}}_j\right]_2 \ge \cdots \ge \left[\overline{\boldsymbol{\beta}}_j\right]_p, \quad (41)$$

and there is a permutation π: $\{1, 2, \ldots, p\} \longrightarrow \{1, 2, \ldots, p\}$ with $[\overline{\beta}_j]_i = |\overline{\beta}_{j\pi(i)}|$ for all $i \in (1, 2, \ldots, p)$, and $\overline{\beta}_{j\pi(i)}$ is the $\pi(i)$-th component of the vector $\overline{\beta}_j$.

Similar to the choice of the regularization parameter in FP algorithm (see Scheme 2 in [19]), here, we can approximate $\widetilde{\beta}_j^*$ by using the current step value $(\beta_j^*)^n$ and the γ_j can be chosen as

$$\gamma_j = \begin{cases} \dfrac{2}{a\mu}\left[B_\mu\left(\boldsymbol{\beta}_j^*\right)^n\right]_{k+1}, & \text{if } \dfrac{2}{a\mu}\left[B_\mu\left(\boldsymbol{\beta}_j^*\right)^n\right]_{k+1} \le \dfrac{1}{a^2 \mu}, \\ \dfrac{(1-\overline{\varepsilon})}{4a^2 \mu}\left(2a\left[B_\mu\left(\boldsymbol{\beta}_j^*\right)^n\right]_k + 1\right)^2, & \text{if } \dfrac{2}{a\mu}\left[B_\mu\left(\boldsymbol{\beta}_j^*\right)^n\right]_{k+1} > \dfrac{1}{a^2 \mu}, \end{cases} \quad (42)$$

where $\overline{\varepsilon}$ is a small positive number, such as 0.1, 0.01, or 0.001, and k is sparsity degree of $\widetilde{\beta}_j^*$. By this way, the iterative algorithm is adaptive and free from the regular parameter choice. Noticing that $0 < \mu < \|\mathbf{X}^*\|_2^{-2}$ in the (2) of Theorem 2, we take $\mu = (1 - \varepsilon/\|\mathbf{X}^*\|_2^2)$, where $\varepsilon \in (0, 1)$.

With abovementioned parameter-choosing method, according to the iterative method of (40), we have the following alternating iterative FP-SPCA thresholding algorithm. If $\mathbf{A}$ is given, for each $(j = 1, 2, \ldots, r)$, $\mathbf{y}_j^*$ is the same as in formula (24), by using the iterative formula (40), $(\beta_j^*)^{n+1}$ is obtained. If $\mathbf{B}$ is fixed, according to (21), it is only try to minimize $\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^\top \mathbf{x}_i\|^2 = \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|^2$ with respect to $\mathbf{A}$, where $\mathbf{A}^\top \mathbf{A} = I_r$. According to the reduced rank form of the Procrustes rotation in [26], if the singular

value decomposition of matrix $(\mathbf{X}^\top\mathbf{X})\mathbf{B}$ is $\mathbf{U}_1\mathbf{D}_1(\mathbf{V}_1)^\top$, the solution is $\hat{\mathbf{A}} = \mathbf{U}_1(\mathbf{V}_1)^\top$ [3]. The alternating iterative FP-SPCA thresholding algorithm is described in Algorithm 1.

At last, we discuss the convergence of the iterative FP-SPCA thresholding algorithm which converges to a stationary point of the iteration (40).

Theorem 2. *Let $\{(\beta_j^*)^n\}$ be the sequence generated by the iterative FP-SPCA thresholding algorithm (40) with $0 < \mu < \|\mathbf{X}^*\|_2^{-2}$. Then,*

(i) The sequence $C_{\gamma_j}((\beta_j^*)^n) = \|\mathbf{y}_j^* - \mathbf{X}^*(\beta_j^*)^n\|^2 + \gamma_j P_a((\beta_j^*)^n)$ is monotonically decreasing

(ii) $\{(\beta_j^*)^n\}$ is asymptotically regular, i.e., $\|(\beta_j^*)^{n+1} - (\beta_j^*)^n\| \longrightarrow 0$ as $n \longrightarrow \infty$

(iii) $\{(\beta_j^*)^n\}$ converges to a stationary point of the iteration (40)

Proof. The proof is similar to the proof of Theorem 4.1 in [29] and Lemma 2 in [30], so it is omitted. □

4. Numerical Experiments

In this section, we present numerical results of the FP-SPCA method and compare the result with the SPCA method [3]. The simulations are conducted on a personal computer (Intel(R) core(TM) i7-6700 cpu@3.40GHZ 3.40GHZ, RAM 16.0 GB) with MATLAB R2014a programming platform.

4.1. Experimental Preparation. In order to compare with the SPCA method, we use the synthetic data as in [3]. The synthetic data has three hidden factors,

$$\begin{aligned} V_1 &\sim N(0, 290), \\ V_2 &\sim N(0, 300), \\ V_3 &= -0.3V_1 + 0.925V_2 + \varepsilon, \\ \varepsilon &\sim N(0, 1), \end{aligned} \tag{43}$$

where V_1, V_2, and ε are independent.

The 10 observable variables are generated as follows:

$$\begin{aligned} X_i &= V_1 + \varepsilon_i^1, \varepsilon_i^1 \sim N(0,1), \quad i = 1,2,3,4, \\ X_i &= V_2 + \varepsilon_i^2, \varepsilon_i^2 \sim N(0,1), \quad i = 5,6,7,8, \\ X_i &= V_3 + \varepsilon_i^3, \varepsilon_i^3 \sim N(0,1), \quad i = 9,10, \end{aligned} \tag{44}$$

where $\{\varepsilon_i^j\}$ are independent, $j = 1,2,3$ and $i = 1,2,\ldots,10$. We use the matrix $\mathbf{X} = (X_1, X_2, \ldots, X_{10})$ to perform PCA, SPCA, and FP-SPCA.

The three hidden factors have three different variance 290, 300, and 283.7875. The number of variables with respect to the three factors are 4, 4, and 2, so V_1 and V_2 are nearly equally important, and they are more important than V_3. These facts suggest that we need to consider the relatively important variables with proper sparse representations only. The two derived variables should recover the factors V_1 and V_2 well using (X_1, X_2, X_3, X_4) and (X_5, X_6, X_7, X_8) accordingly.

The SPCA and FP-SPCA methods were carried out by the method proposed by Zou et al. [3] and the fractional penalty method proposed in [19]. We compared the performance of the two methods using the synthetic data matrix $\mathbf{X}$ and summarize the results in two cases below.

4.2. Parameter Selection and the Result of Experiments

Case 1. In this situation, the method proposed by Zou et al. [3] which is called SPCA and our proposed method which is called FP-SPCA have the same performance for the almost equally important hidden factors. That is, for most of the generated data matrices, the SPCA and FP-SPCA have the same loadings but different adjusted variance.

Table 1 reports a numerical result of PCA, SPCA, and FP-SPCA and summarizes the results in loadings and adjusted variance. The parameters are $(\lambda, \lambda_{1,j}, \varepsilon) = (0.01, 3.7, 0.01)$ in SPCA and $(\lambda, a, \varepsilon, \overline{\varepsilon}) = (0.01, 1.2, 0.01, 0.001)$ in FP-SPCA. We can see, from Table 1, that both SPCA and FP-SPCA have same orthogonal loadings of PC1 and PC2 after normalization, but the adjusted variance obtained from FP-SPCA is bigger than that from SPCA. That is to say, PC1 and PC2 obtained by FP-SPCA method have more information than PC1 and PC2 by SPCA. Numerical simulations using other generated data matrices have the same result also.

Case 2. For a part of the generated matrices, according to our experimental results, the FP-SPCA does work, but SPCA does not work well. That is, by adjusting parameters λ and $\lambda_{1,j}$ in the SPCA method, we cannot get the sparse loadings of PC1 or PC2. The parameters and the corresponding numerical results are reported in Table 2, where capital letter N represents no sparse loadings of PC1 and PC2 have been obtained when parameters are the value above the N and the value(s) on the left-hand side of the N. NaN means not a number when parameters acquire the value(s) in the same way as capital letter N in Table 2. Also, we can see, from Table 2, that the loadings of PC1 or PC2 show a meaningless result when $\lambda_{1,j}$ is between 3.5 and 4.0. Numerical simulations show the same meaningless result when $\lambda_{1,j}$ is increasing by a large degree. We give, in Table 3, a result of loadings of PC1 and PC2 when $(\lambda_{1,j}, \lambda, \varepsilon) = (1.8, 0.5, 0.01)$ using the SPCA method and the result of the FP-SPCA method when $(\lambda, a, \varepsilon, \overline{\varepsilon}) = (0.001, 1.3, 0.01, 0.001)$. In a word, the FP-SPCA method can find the sparse loadings of PC1 and PC2 while the SPCA method is at least difficult to obtain the sparse result in this case.

Initialize: Choose $(\beta_j^*)^0$, k, $\mu = (1 - \varepsilon/\|\mathbf{X}^*\|_2^2)$, $\overline{\epsilon}$, $\mathbf{W}$, a, ϵ, and λ.
while not converged **do**
  for $j = 1: r$
    $B_\mu((\beta_j^*)^n) = (\beta_j^*)^n + \mu(\mathbf{X}^*)^\top(\mathbf{y}_j^* - \mathbf{X}^*(\beta_j^*)^n)$
    $\gamma_{j1} = (2/a\mu)[B_\mu(\beta_j^*)]_{k+1}$; $\gamma_{j2}^n = ((1 - \overline{\epsilon})/4a^2\mu)(2a[B_\mu(\beta_j^*)^n]_k + 1))^2$
    if $\gamma_{j1} \le (1/a^2\mu)$ then
      $\gamma_j = \gamma_{j1}$; $t = (\gamma_j\mu a/2)$
    else
      $\gamma_j = \gamma_{j2}$; $t = \sqrt{\gamma_j\mu} - (1/2a)$
    end
    $\lambda_{1,j} = \gamma_j\sqrt{1 + \lambda}$;
    for $i = 1: p$
    if $|(B_\mu((\beta_j^*)^n))_i| > t$
      $((\beta_j^*)^{n+1})_i = g_{\gamma_j\mu}((B_\mu((\beta_j^*)^n))_i)$
    else
      $((\beta_j^*)^{n+1})_i = 0$
    end
  end
  $\mathbf{W}(:, j) = (\beta_j^*)^{n+1}$
  end;
  $\mathbf{C}: = (\mathbf{X}^*)^\top\mathbf{X}^*\mathbf{W}$
  $[\mathbf{U}_1, \mathbf{S}_1, \mathbf{V}_1] = \mathrm{svd}(\mathbf{C})$
  $\mathbf{V} = \mathbf{U}_1(\mathbf{V}_1)^\top$
  $n \longrightarrow n + 1$
end while
for $m = 1: r$
  $\widehat{\beta}_m = (\mathbf{W}(:, m)/\mathrm{norm}(\mathbf{W}(:, m)))$
end
output: $\widehat{\beta}_m, m = 1, 2, \ldots, r$.

Algorithm 1: Iterative FP-SPCA thresholding algorithm.

Table 1: Loadings and variance of numerical results of PCA, SPCA, and FP-SPCA methods in Case 1, where the SPCA and FP-SPCA methods have the same performance in obtaining the sparse loadings while FP-SPCA performs better than SPCA in adjusted variance. $AV^\dagger$ (%) denotes the adjusted variance.

	PCA			SPCA		FP-SPCA	
	PC1	PC2	PC3	PC1	PC2	PC1	PC2
X_1	0.3708	0.1993	−0.2698	0.5	0	0.5	0
X_2	0.3708	0.1993	−0.2698	0.5	0	0.5	0
X_3	0.3708	0.1993	−0.2698	0.5	0	0.5	0
X_4	0.3708	0.1993	−0.2698	0.5	0	0.5	0
X_5	0.2625	−0.4228	0.0483	0	−0.5	0	−0.5
X_6	0.2625	−0.4228	0.0483	0	−0.5	0	−0.5
X_7	0.2625	−0.4228	0.0483	0	−0.5	0	−0.5
X_8	0.2625	−0.4228	0.0483	0	−0.5	0	−0.5
X_9	0.2954	0.2510	0.5914	0	0	0	0
X_{10}	0.2954	0.2510	0.5914	0	0	0	0
$AV^\dagger$ (%)	58.99	33.10	7.91	32.49	29.86	40.00	36.76

From Case 1 and Case 2, it can be seen that the FP-SPCA method is more adaptable than the SPCA method. The reason why the FP-SPCA method is more preferable is that it has a parameter a in the penalty function which makes FP-SPCA more flexible and easier to adjust to get the desirable result.

Table 2: The parameters selection in experiments and the results, where N represents no sparse loadings of PC1 or PC2 have been obtained and NaN means not a number as the MATLAB experimental result shows. 3.5~4.0* denotes the numbers changing from the number 3.5 to 4.0 increasing by 0.1 each time.

$\lambda_{1,j}\lambda$	0.001	0.010	0.100	1.000
0.001	N	N	N	N
0.01	N	N	N	N
0.1 ~ 3.1	N	N	N	N
3.2 ~ 3.4	N	N	N	NaN
3.5~4.0*	NaN	NaN	NaN	NaN

Table 3: Loadings and variance of numerical results of PCA, SPCA, and FP-SPCA methods in Case 2, where the FP-SPCA method works better than SPCA in obtaining the sparse loadings. AV† (%) denotes the adjusted variance.

	PCA			SPCA		FP-SPCA	
	PC1	PC2	PC3	PC1	PC2	PC1	PC2
X_1	−0.3472	−0.3503	0.0819	−0.3565	−0.3466	−0.5	0
X_2	−0.3472	−0.3503	0.0819	−0.3565	−0.3466	−0.5	0
X_3	−0.3472	−0.3503	0.0819	−0.3565	−0.3466	−0.5	0
X_4	−0.3472	−0.3503	0.0819	−0.3565	−0.3466	−0.5	0
X_5	−0.3444	0.3566	0.0651	−0.3505	0.3604	0	0.5
X_6	−0.3444	0.3566	0.0651	−0.3505	0.3604	0	0.5
X_7	−0.3444	0.3566	0.0651	−0.3505	0.3604	0	0.5
X_8	−0.3444	0.3566	0.0651	−0.3505	0.3604	0	0.5
X_9	−0.1472	−0.0159	−0.6914	0	0	0	0
X_{10}	−0.1472	−0.0159	−0.6914	0	0	0	0
AV† (%)	49.23	32.10	18.67	44.61	29.88	40.00	38.44

5. Conclusions

In this paper, the FP-SPCA method is proposed based on the SPCA method [3, 4], where a fractional penalty function [19] is proposed to replace the l_1-norm in the SPCA method. Numerical simulations show that the proposed FP-SPCA method is more adaptable and flexible than the SPCA method. However, the FP-SPCA has its limitation, and during the numerical simulations, we find a few generated matrices that FP-SPCA fails to produce sparse loadings of principal components and the SPCA method does not work also. How to deal with this problem and how to compare the method we proposed with other existing penalty methods are remaining questions that will be worth investigating further.

Data Availability

The data used to support the findings of this study are available from the corresponding author on request.

Conflicts of Interest

The authors declare that there are no known conflicts of interest associated with this publication.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant nos. 11771347, 11871392, 91730306, and 41390454.

References

[1] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, USA, Second edition, 2002.

[2] X. Li, M. K. Ng, X. Xu, and Y. Ye, "Block principal component analysis for tensor objects with frequency or time information," *Neurocomputing*, vol. 302, pp. 12–22, 2018.

[3] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[4] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] Z. Hui, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[8] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[9] F. R. Bach, "Bolasso," in *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pp. 33–40, Helsinki, Finland, July 2008.

[10] A. Chatterjee and S. N. Lahiri, "Bootstrapping lasso estimators," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 608–625, 2011.

[11] D. Vidaurre, C. Bielza, and P. Larrañaga, "A survey ofL1-Regression," *International Statistical Review*, vol. 81, no. 3, pp. 361–387, 2013.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[13] T. Hesterberg, N. H. Choi, L. Meier, and C. Fraley, "Least angle and ℓ 1 penalized regression: a review," *Statistics Surveys*, vol. 2, pp. 61–93, 2008.

[14] P. Radchenko and G. M. James, "Variable inclusion and shrinkage algorithms," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1304–1315, 2008.

[15] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[16] E. Candès and T. Tao, "The dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[17] C. Leng and H. Wang, "On general adaptive sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 18, no. 1, pp. 201–215, 2009.

[18] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the lasso," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 531–547, 2003.

[19] H. Li, Q. Zhang, A. Cui, and J. Peng, "Minimization of fraction function penalty in compressed sensing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1626–1637, 2019.

[20] J. Li, X. Zhang, R. Cao, and M. Zhou, "Reduced-dimension music for angle and array gain-phase error estimation in bistatic mimo radar," *IEEE Communications Letters*, vol. 17, no. 3, pp. 443–446, 2013.

[21] J. Li, J. Ding, and D. Jiang, "Joint direction finding and array calibration method for mimo radar with unknown gain phase errors," *IET Microwaves, Antennas & Propagation*, vol. 10, no. 14, pp. 1563–1569, 2016.

[22] F. Wen, Z. Zhang, K. Wang, G. Sheng, and G. Zhang, "Angle estimation and mutual coupling self-calibration for ula-based bistatic mimo radar," *Signal Processing*, vol. 144, pp. 61–67, 2018.

[23] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[24] T. Liu, F. Wen, L. Zhang, and K. Wang, "Off-grid doa estimation for colocated mimo radar via reduced-complexity sparse bayesian learning," *IEEE Access*, vol. 7, pp. 99907–99916, 2019.

[25] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[26] K. V. Mardia, J. T. Kent, and J. M. Bibby, "Multivariate analysis," *Mathematical Gazette*, vol. 37, no. 1, pp. 123–131, 1979.

[27] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[28] P. Wang, Z. He, K. Xie, J. Gao, M. Antolovich, and B. Tan, "A hybrid algorithm for low-rank approximation of nonnegative matrix factorization," *Neurocomputing*, vol. 364, pp. 129–137, 2019.

[29] D. Peng, N. Xiu, and J. Yu, "$S_{1/2}$ regularization methods and fixed point algorithms for affine rank minimization problems," *Computational Optimization and Applications*, vol. 67, no. 3, pp. 543–569, 2017.

[30] J. Zeng, S. Lin, Y. Wang, and Z. Xu, "$L_{1/2}$regularization: convergence of iterative half thresholding algorithm," *IEEE Transactions on Signal Processing*, vol. 62, no. 9, pp. 2317–2329, 2014.