Sparse Unsupervised Dimensionality Reduction Algorithms

Wenjun Dou, Guang Dai, Congfu Xu, and Zhihua Zhang

College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang 310027, China
{xucongfu,zhzhang}@cs.zju.edu.cn

Abstract. Principal component analysis (PCA) and its dual—principal coordinate analysis (PCO)—are widely applied to unsupervised dimensionality reduction. In this paper, we show that PCA and PCO can be carried out under regression frameworks. Thus, it is convenient to incorporate sparse techniques into the regression frameworks. In particular, we propose a sparse PCA model and a sparse PCO model. The former is to find sparse principal components, while the latter directly calculates sparse principal coordinates in a low-dimensional space. Our models can be solved by simple and efficient iterative procedures. Finally, we discuss the relationship of our models with other existing sparse PCA methods and illustrate empirical comparisons for these sparse unsupervised dimensionality reduction methods. The experimental results are encouraging.

1 Introduction

Unsupervised dimensionality reduction methods are widely used in many applications such as image processing, microarray data analysis, information retrieval, etc. PCA [13] and PCO (or the classical multidimension scaling) [9,16] are two classical unsupervised techniques for dimensionality reduction. PCA aims to find the principal components (PCs) with the largest variance, while PCO directly calculates the coordinate configurations in the dimension-reduced space.

However, it is sometimes difficult to interpret the results with PCA, because each principal component is a linear combination of all the original variables and its loadings are typically nonzero. Many approaches have been developed to deal with this drawback of PCA. Recently, a sparse approach has been introduced. Roughly speaking, the approach is to impose some sparsity constraints such as lasso [19] and elastic net [22] to loadings, then some of loadings are naturally zero. There are mainly two families of sparse PCA methods in the literature. The first one uses the maximum-variance property of principal components, such as SCoTLASS [14], DSPCA [2], sPCA-rSVD [17], SOCA [20], sPCA-DC [18], etc. The other family is based on regression-type problems such as sparse PCA [23].

In this paper we develop a new sparse PCA model to achieve sparseness. Our model is built on the notion of optimal scoring, which was originally used to carry out the Fisher discriminant analysis [11]. Recently, Clemmensen [1] proposed a sparse discriminant analysis method by optimal scoring. Zhang and Dai [21] showed that some

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part I, LNAI 6321, pp. 361–376, 2010.

unsupervised learning methods, such as spectral clustering and PCA, can be cast into an optimal scoring framework. This work immediately motivates our sparse PCA model. An advantage of our model over the other sparse PCA methods is that our model can achieve more sparseness when the total explained variance of principal components is approximately same. Moreover, since our sparse PCA is derived from the optimal scoring, it is more appropriately applied to discriminant analysis problems.

When applying these sparse PCA methods, we are able to obtain sparse principal loadings. However, the coordinate configurations in the dimension-reduced space are not necessarily sparse. In practical applications, it would be sometimes interesting to find sparse principal coordinates. As we know, however, there is no work about this theme. In this paper we present a sparse PCO model. In particular, we exploit the Eckart-Young theorem [3], because the theorem shows a dual relationship between the conventional PCA model and the conventional PCO model. Moreover, we note that it also provides a regression framework to perform PCO. Introducing the elastic net penalty for principal coordinates into this framework, we devise our sparse PCO.

The rest of this paper is organized as follows. Sections 2 and 3 present our sparse PCA and PCO models respectively. Section 4 discusses the relationship of our sparse models with other existing sparse PCA methods. In Section 5, we conduct our experimental evaluations. Finally, we give our conclusions in Section 6.

2 Sparse PCA

Optimal scoring was first introduced by Hastie [11] to formulate the Fisher linear discriminant analysis as a multiple linear regression problem. In recent work, Zhang and Dai [21] extended the concept of optimal scoring to unsupervised learning problems and developed a framework for unsupervised clustering. Based on the optimal scoring framework, we now develop our sparse principal component analysis model, namely *sparse PCA via optimal scoring* (sPCA-OS).

First of all, we list some notations. Throughout this paper, $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, $\mathbf{1}_m$ the $m \times 1$ of ones, and $\mathbf{0}$ the zero matrix or vector whose dimensionality is dependent upon the context.

For an $n \times m$ matrix $\mathbf{A} = [a_{ij}]$, let $\mathsf{vec}(\mathbf{A}) = (a_{11}, \ldots, a_{n1}, a_{12}, \ldots, a_{nm})^T$ be the $nm \times 1$ vector, $\|\mathbf{A}\|_F = \|\mathsf{vec}(\mathbf{A})\|_2 = \sqrt{\sum_{i,j} a_{ij}^2}$ be the Frobenius norm of $\mathbf{A}$ or the 2-norm of $\mathsf{vec}(\mathbf{A})$, and $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$ be the 1-norm of $\mathsf{vec}(\mathbf{A})$. In addition, $\mathbf{A} \otimes \mathbf{B} = [a_{ij}\mathbf{B}]$ represents the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, and

$$\mathcal{O}^{n \times m} = \{\mathbf{A} : \mathbf{A} \in \mathbb{R}^{n \times m} \text{ and } \mathbf{A}^T\mathbf{A} = \mathbf{I}_m\}.$$

2.1 Optimal Scoring for PCA

We are given an $n \times p$ data matrix $\mathbf{X} = [\mathbf{x}_{ij}]$, where n is the number of observations and p is the number of variables. Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_q]$ be an $n \times q$ sample scoring matrix such that $\mathbf{1}_n^T\mathbf{Y} = \mathbf{0}$ and $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_q$ and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_q]$ be a $p \times q$ weight matrix. For dimensionality reduction problems, q represents the number of PCs (or loadings) and must be less than p.

Without loss of generality, suppose that $\mathbf{X}$ is centered; that is, $\mathbf{1}_n^T\mathbf{X} = \mathbf{0}$. The framework of optimal scoring for unsupervised learning is then defined by

$$\min_{\mathbf{Y},\mathbf{W}} \left\{ f(\mathbf{Y},\mathbf{W}) \equiv \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2 + \frac{\delta^2}{2}\mathsf{tr}(\mathbf{W}^T\mathbf{W}) \right\}$$

subject to $\mathbf{1}_n^T\mathbf{Y} = \mathbf{0}$ and $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_q$. Zhang and Dai [21] proved that the minimum of f is obtained when $\mathbf{Y}$ is the $n \times q$ matrix of the top orthonormal eigenvectors of $\mathbf{X}(\mathbf{X}^T\mathbf{X} + \delta^2\mathbf{I}_p)^{-1}\mathbf{X}^T$ and $\mathbf{W} = (\mathbf{X}^T\mathbf{X} + \delta^2\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}$. Obviously, $\mathbf{W}$ can be treated as a non-orthogonal matrix of loadings and then $\mathbf{X}\mathbf{W}$ is the low-dimensional principal coordinate matrix of $\mathbf{X}$.

Let $r = \min\{n, p\}$, we make the full singular value decomposition (SVD) [7] of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ $(n \times n)$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ $(p \times p)$ are orthogonal matrices, and $\mathbf{D} = \mathsf{diag}(d_1, \ldots, d_r)$ $(n \times p)$ is a diagonal matrix with $d_1 \geq d_2 \geq \cdots \geq d_r \geq 0$. We then obtain the minimizers of $f(\mathbf{Y},\mathbf{W})$ as $\mathbf{Y} = [\mathbf{u}_1, \ldots, \mathbf{u}_q]$ and $\mathbf{W} = \mathbf{V}_1(\mathbf{D}_1^2 + \delta^2\mathbf{I}_q)^{-1}\mathbf{D}_1$, where $\mathbf{V}_1 = [\mathbf{v}_1, \ldots, \mathbf{v}_q]$ and $\mathbf{D}_1 = \mathsf{diag}(d_1, \ldots, d_q)$. This immediately leads us to the following theorem.

Theorem 1. *Assume that* $\mathbf{Y}$ *is an* $n \times q$ *optimal scoring matrix and* $\mathbf{W}$ *is a* $p \times q$ *loading matrix. Consider*

$$(\hat{\mathbf{Y}}, \hat{\mathbf{W}}) = \operatorname*{argmin}_{\mathbf{Y},\mathbf{W}} \; f(\mathbf{Y},\mathbf{W}) \tag{1}$$

under the constraints $\mathbf{1}_n^T\mathbf{Y} = 0$ *and* $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_q$. *Then* $\hat{\mathbf{w}}_j \propto \mathbf{v}_j$ *for* $j = 1, \ldots, q$.

Zou [23] proposed a regression approach for solving PCA. Theorem 1 shows that we can develop an alternative regression formulation of PCA.

2.2 Sparse PCA via Optimal Scoring

It is natural to impose a sparse penalty to the loading matrix $\mathbf{W}$. Particularly, we exploit the elastic net in our sPCA-OS method. That is, we consider the following optimization problem:

$$\begin{aligned} &\operatorname*{argmin}_{\mathbf{Y},\mathbf{W}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2 + \frac{\delta^2}{2}\|\mathbf{W}\|_F^2 + \sum_{j=1}^{q}\lambda_j\|\mathbf{w}_j\|_1 \\ &\text{s.t. } \mathbf{1}_n^T\mathbf{Y} = \mathbf{0} \text{ and } \mathbf{Y}^T\mathbf{Y} = \mathbf{I}_q, \end{aligned} \tag{2}$$

where δ is applied to all the q components and the λ_j is used to let each loading $\mathbf{w}_j$ has different degree of sparseness.

Problem (2) can be solved by an iternative procedure. First, with fixed $\mathbf{Y}$, the optimization problem (2) is converted into the conventional elastic net problem [22]; namely,

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{j=1}^{q}\|\mathbf{y}_j-\mathbf{X}\mathbf{w}_j\|_2^2 + \frac{\delta^2}{2}\sum_{j=1}^{q}\|\mathbf{w}_j\|_2^2 + \sum_{j=1}^{q}\lambda_j\|\mathbf{w}_j\|_1.$$

This problem can be decomposed into q separable optimization problems; that is, for $j = 1, \ldots, q$, we have

$$\min_{\mathbf{w}_j} \frac{1}{2}\|\mathbf{y}_j - \mathbf{X}\mathbf{w}_j\|_2^2 + \frac{\delta^2}{2}\|\mathbf{w}_j\|_2^2 + \lambda_j\|\mathbf{w}_j\|_1. \tag{3}$$

Second, with fixed $\mathbf{W}$, we can ignore the penalty terms in the optimization problem (2) and it becomes a Procrustes problem [5] as follows:

$$\begin{aligned} &\operatorname*{argmin}_{\mathbf{Y}} \ \frac{1}{2}\sum_{j=1}^{q}\|\mathbf{y}_j - \mathbf{X}\mathbf{w}_j\|_2^2 \\ &\text{s.t. } \mathbf{1}_n^T\mathbf{Y} = \mathbf{0} \text{ and } \mathbf{Y}^T\mathbf{Y} = \mathbf{I}_q. \end{aligned} \tag{4}$$

This problem is easily solved in a closed form (see, e.g., Gower and Dijksterhuis [10]). Let the thin SVD [7] of $\mathbf{XW}$ be $\boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T$ where $\boldsymbol{\Psi}$ ($n \times q$) and $\boldsymbol{\Phi}$ ($q \times q$) satisfy $\boldsymbol{\Psi}^T\boldsymbol{\Psi} = \mathbf{I}_q$ and $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \mathbf{I}_q$ and $\boldsymbol{\Delta}$ ($q \times q$) is the diagonal matrix with nonnegative entries. Then $\mathbf{XW} = \boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T$ is a solution of (4). The procedure for solving our sPCA-OS is summarized in Algorithm 1.

Algorithm 1. SparsePCA via Optimal Scoring(sPCA-OS)

1: Initialize a $\mathbf{Y}$ subject to $\mathbf{1}_n^T\mathbf{Y} = \mathbf{0}$ and $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_q$.
2: With a fixed $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_q]$, solve the elastic net problems for $j = 1, \ldots, q$:

$$\operatorname*{argmin}_{\mathbf{w}_j} \frac{1}{2}\|\mathbf{y}_j - \mathbf{X}\mathbf{w}_j\|_2^2 + \frac{\delta^2}{2}\|\mathbf{w}_j\|_2^2 + \lambda_j\|\mathbf{w}_j\|_1.$$

3: With a fixed $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_q]$, perform the thin SVD of $\mathbf{XW}$ as $\mathbf{XW} = \boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T$ and update $\mathbf{Y}$ by $\mathbf{Y} = \boldsymbol{\Psi}\boldsymbol{\Phi}^T$.
4: Repeat Steps 2 and 3 until convergence.

3 Sparse PCO

Although $\mathbf{W}$ obtained via Algorithm 1 is sparse, the coordinate matrix $\mathbf{Z} = \mathbf{XW}$ is not necessarily sparse. However, it would be also interesting in the situation that $\mathbf{Z}$ is sparse. We thus attempt to develop a sparse PCO algorithm in which the coordinate matrix is sparse. First of all, we present the following theorem.

Theorem 2. *Let the full SVD of* $\mathbf{X}$ *be* $\mathbf{X} = \mathbf{UDV}^T$ *where* $\mathbf{U} \in \mathcal{O}^{n\times n}$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p] \in \mathcal{O}^{p\times p}$ *and* $\mathbf{D} = \mathsf{diag}(d_1, \ldots, d_r)$ *with* $r = \min\{n, p\}$ *and* $d_1 \geq d_2 \geq \ldots \geq d_r \geq 0$. *Assume that* g *is defined by*

$$g(\mathbf{A}, \mathbf{Z}) = \|\mathbf{X} - \mathbf{Z}\mathbf{A}^T\|_F^2 + \gamma\|\mathbf{Z}\|_F^2 \tag{5}$$

where $\mathbf{A} \in \mathcal{O}^{p\times q}$, $\mathbf{Z} \in \mathbb{R}^{n\times q}$ *and* $\gamma \geq 0$. *Then the minimum of* g *is obtained when* $\mathbf{A} = [\mathbf{v}_1, \ldots, \mathbf{v}_q]$ *and* $\mathbf{Z} = \frac{1}{1+\gamma}\mathbf{XA}$.

The proof of this theorem is given in Appendix A. When $\gamma = 0$, Theorem 2 degenerates to the Eckart-Young theorem [3,15].

Theorem 2 shows that $\mathbf{Z}$ is just the coordinate matrix up to the constant $\frac{1}{1+\gamma}$. In order to make $\mathbf{Z}$ sparse, we impose the elastic net penalty on it. Accordingly, we have our sparse PCO model which is defined by the following optimization problem:

$$\min_{\mathbf{A}\in\mathcal{O}^{p\times q},\mathbf{Z}\in\mathbb{R}^{n\times q}} \frac{1}{2}\|\mathbf{X}-\mathbf{Z}\mathbf{A}^T\|_F^2 + \frac{\gamma_1}{2}\|\mathbf{Z}\|_F^2 + \gamma_2\|\mathbf{Z}\|_1, \tag{6}$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ are regularization parameters. We also resort to an alternatively iterative procedure to solve the problem (6).

With a fixed $\mathbf{Z}$, the optimization problem (6) becomes

$$\min_{\mathbf{A}\in\mathcal{O}^{p\times q}} \frac{1}{2}\|\mathbf{X}-\mathbf{Z}\mathbf{A}^T\|_F^2,$$

which is a reduced rank Procrustes problem illustrated by Zou [23]. Suppose that the thin SVD of $\mathbf{X}^T\mathbf{Z}$ is $\mathbf{X}^T\mathbf{Z} = \boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T$. Then $\mathbf{A} = \boldsymbol{\Psi}\boldsymbol{\Phi}^T$ is the solution of this Procrustes problem.

With a fixed $\mathbf{A}$, the optimization problem (6) degenerates to

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\alpha}-(\mathbf{A}\otimes\mathbf{I}_n)\boldsymbol{\beta}\|_2^2 + \frac{\gamma_1}{2}\|\boldsymbol{\beta}\|_2^2 + \gamma_2\|\boldsymbol{\beta}\|_1, \tag{7}$$

where $\boldsymbol{\alpha} = \mathsf{vec}(\mathbf{X})$ and $\boldsymbol{\beta} = \mathsf{vec}(\mathbf{Z})$. Since $(\mathbf{A}\otimes\mathbf{I}_n)^T(\mathbf{A}\otimes\mathbf{I}_n) = (\mathbf{A}^T\mathbf{A})\otimes\mathbf{I}_n = \mathbf{I}_{qn}$, this problem can be directly solved via the soft thresholding algorithm [19,22].

In summary, we have our SPCO algorithm which is given in Algorithm 2.

Algorithm 2. Sparse PCO (SPCO)

1: Give $\mathbf{X}$ and initialize $\mathbf{A}$.
2: Fix $\mathbf{A}$ and solve the elastic net problem w.r.t. $\boldsymbol{\beta}$.

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\alpha}-(\mathbf{A}\otimes\mathbf{I}_n)\boldsymbol{\beta}\|_2^2 + \frac{\gamma_1}{2}\|\boldsymbol{\beta}\|_2^2 + \gamma_2\|\boldsymbol{\beta}\|_1.$$

3: Fix $\boldsymbol{\beta}$ and perform the thin SVD of $\mathbf{X}^T\mathbf{Z}$ as $\mathbf{X}^T\mathbf{Z} = \boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T$ and update $\mathbf{A}$ by $\mathbf{A} = \boldsymbol{\Psi}\boldsymbol{\Phi}^T$.
4: Repeat Step 2 and 3 until convergence.

4 Related Work

In the literature [20], the authors illustrated the relationship among SCoTLASS [14], sPCA-rSVD [17], the SPCA method of Zou *et al.* [23] (called sPCA-ZHT), and the SPCA method of Witten *et al.* [20] (called SPCA-WTH). We further discuss the relationship of our sPCA-OS and SPCO with these existing sparse PCA methods.

First, the SPCA method of Zou *et al.* [23] (SPCA-ZHT) is defined as

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times q}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{A}^T\|_F^2+\frac{\gamma_1}{2}\|\mathbf{W}\|_F^2+\sum_{j=1}^{q}\gamma_{2,j}\|\mathbf{w}_j\|_1,$$
$$\text{s.t. } \mathbf{A}^T\mathbf{A}=\mathbf{I}_q.$$

Comparing this model with our SPCO model in (6), a connection between these two models is immediately obtained via letting $\mathbf{Z}=\mathbf{X}\mathbf{W}$. In our SPCO, we devise a different penalty term

$$\frac{\gamma_1}{2}\mathsf{tr}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})+\gamma_2\|\mathbf{X}\mathbf{W}\|_1,$$

which can be regarded as a weighted norm of $\mathbf{W}$ with respect to $\mathbf{X}$.

Since $\mathbf{A}$ ($p\times q$) is orthogonal, there exists a $p\times(p-q)$ matrix $\mathbf{A}_0$ such that $\mathbf{A}_0^T\mathbf{A}_0=\mathbf{I}_{p-q}$ and $\mathbf{A}^T\mathbf{A}_0=\mathbf{0}$. Thus,

$$\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{A}^T\|_F^2=\|\mathbf{X}\mathbf{A}_0\|_F^2+\|\mathbf{X}\mathbf{A}-\mathbf{X}\mathbf{W}\|_F^2.$$

This implies that SPCA-ZHT is equivalent to

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times q}} \frac{1}{2}\|\mathbf{X}\mathbf{A}-\mathbf{X}\mathbf{W}\|_F^2+\frac{\gamma_1}{2}\|\mathbf{W}\|_F^2+\sum_{j=1}^{q}\gamma_{2,j}\|\mathbf{w}_j\|_1,$$
$$\text{s.t. } \mathbf{A}^T\mathbf{A}=\mathbf{I}_q,$$

which with (2) together shows an interesting connection between SPCA-ZHT and our sPCA-OS by setting $\mathbf{Y}=\mathbf{X}\mathbf{A}$. However, in our model we employ the constraint $\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}=\mathbf{I}_q$.

Second, when $q=1$, the sPCA-rSVD model of Shen & Huang [17] is defined as

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times q}} \frac{1}{2}\|\mathbf{X}-\mathbf{Z}\mathbf{A}^T\|_F^2+\frac{\gamma_1}{2}\|\mathbf{A}\|_F^2+\gamma_2\|\mathbf{A}\|_1,$$
$$\text{s.t. } \mathbf{Z}^T\mathbf{Z}=\mathbf{I}_q.$$

We see that there is a duality between sPCA-rSVD and SPCO, in which the roles of $\mathbf{Z}$ and $\mathbf{A}$ are exchanged.

We now study the relationship of sPCA-OS with SPCA-WTH [20] and SCoTLASS [14]. Assume $q=1$, we write (2) for sPCA-OS as

$$\min_{\mathbf{y},\mathbf{w}} f(\mathbf{y},\mathbf{w})=\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2,$$
$$\text{s.t. } \|\mathbf{w}\|_2^2\le c_1,\ \|\mathbf{w}\|_1\le c_2,\ \|\mathbf{y}\|_2^2=1,$$

which can be used to associate sPCA-OS with SPCA-WTH and SCoTLASS, because SPCA-WTH is based on the following problem

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}\mathbf{v},$$
$$\text{s.t. } \|\mathbf{v}\|_2^2\le 1,\ \|\mathbf{v}\|_1\le c_2, \|\mathbf{u}\|_2^2\le 1,$$

while SCoTLASS is based on the problem:

$$\max_{\mathbf{w}} \ \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w},$$
$$\text{s.t. } \|\mathbf{w}\|_2^2 \leq c_1, \ \|\mathbf{w}\|_1 \leq c_2.$$

5 Experiments

In this section we compare our sPCA-OS and SPCO algorithms with two closely related sparse PCA methods, i.e., SPCA-ZHT [23] and SPCA-WTH [20]. The experiments are conducted on the `pitprops` dataset, two synthetic datasets, six UCI datasets and one gene microarray dataset. Following the setting in [17,23], we also employ *cumulative percentage of explained variance* (CPEV) as an evaluation criterion.

5.1 Evaluations on the `Pitprops` Dataset

The `pitprops` dataset was first put forward by Jeffers [12] for difficulty of interpreting PCs. The dataset consists of 13 variables and 180 observations, and has become a standard example illustrating the potential difficulty of interpreting principal components. Jeffers [12] suggested explaining the first six components. Thus, we also select the first six PCs to analyze SPCA-ZHT, SPCA-WTH and sPCA-OS. Similar to [23], the corresponding regularized parameters are identified on the basis of that each sparse approximation explains almost the same amount of variance as the ordinary PC does. Tables 1-3 show the experimental results with these sparse PCA methods.

It is seen from Tables 1-3 that with regard to CPEV, the sPCA-OS method should be competitive with the other two sparse PCA methods, because the values of CPEV for SPCA-ZHT, SPCA-WTH and sPCA-OS are 75.76%, 75.22% and 75.47%, respectively. Moreover, sPCA-OS and SPCA-ZHT have higher sparseness than SPCA-WTH on the whole, specially as the number of principal components increases. In addition, Figure 1 shows the corresponding variation of variance with respect to the different principal components. It is also worth mentioning that unlike SPCA-WTH, the variances of sPCA-OS and SPCA-ZHT strictly monotonously decrease.

5.2 Evaluations on Two Synthetic Datasets

We also conduct our comparison on the synthetic dataset employed in [23]. In particular, three hidden factors are first created as follows:

$$v_1 \sim \mathcal{N}(0, 290), \ v_2 \sim \mathcal{N}(0, 300), \ v_3 = -0.3v_1 + 0.925v_2 + \epsilon, \ \epsilon \sim \mathcal{N}(0, 1),$$

where v_1, v_2 and ϵ are mutually independent. Then, 10 observed variables $x_1, \ldots, x_{10}$ are generated by:

$$x_i = v_j + \epsilon_i^j, \ \ \epsilon_i^j \sim \mathcal{N}(0, 1),$$

where $j = 1$ corresponds to $i = 1, 2, 3, 4$, $j = 2$ corresponds to $i = 5, 6, 7, 8$, and $j = 3$ corresponds to $i = 9, 10$, and the ϵ_i^j are independent of each other. From the

Table 1. The first six PCs obtained by SPCA-ZHT on the `pitprops` dataset

Variable	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$
topdiam	-0.4796	0	0	0	0	0
length	-0.4689	0	0	0	0	0
moist	0	0.7769	0	0	0	0
testsg	0	0.6289	0	0	0	0
ovensg	0.1903	0	0.6551	0	0	0
ringtop	0	0	0.6048	0	0	0
ringbut	-0.2802	0	0.4528	0	0	0
bowmax	-0.3401	-0.0312	0	0	0	0
bowdist	-0.4148	0	0	0	0	0
whorls	-0.3844	0	0	0	0	0
clear	0	0	0	-1	0	0
knots	0	0	0	0	-1	0
diaknots	0	0	0	0	0	1
CPEV(%)	28.06	42.06	55.16	62.61	69.45	75.76

Table 2. The first six PCs obtained by SPCA-WTH on the `pitprops` dataset

Variable	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$
topdiam	0	-0.6974	0	0	0	0
length	0	-0.6988	0	0	0	0
moist	0	0	0.6825	0	0	0
testsg	0	0	0.6967	0	0	0
ovensg	0	0.1494	0	0	0.2810	0.4391
ringtop	0	0	0	0	0.4995	0
ringbut	-0.8099	0	0	0	0	0
bowmax	0	0	0	-0.2208	0	0
bowdist	0	-0.0544	0	-0.2752	0	-0.8740
whorls	-0.5214	0	0	0	0	-0.0021
clear	0	0	0	0	-0.8195	0.1054
knots	0	0	0	0.8152	0	-0.1795
diaknots	0.2687	0	0	0	0	0
CPEV(%)	14.49	30.89	46.37	57.99	68.12	75.22

formulation above, it is interesting to note that the three hidden factors have about the same variances, and the variables (x_1, x_2, x_3, x_4) are independent of the variables (x_5, x_6, x_7, x_8). Moreover, as the mixed roles, the variables (x_9, x_{10}) have closed relationship with the variables (x_5, x_6, x_7, x_8).

We implement classical PCA, SPCA-ZHT, SPCA-WTH and sPCA-OS on the 500 synthetic points here. Table 4 summarizes the comparison results. Obviously, unlike PCA, the three sparse PCA algorithms have the correct sparse representations

Table 3. The first six PCs obtained by sPCA-OS on the `pitprops` dataset

Variable	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$
topdiam	0.6304	0	0	0	0	0
length	0.6092	0	0	0	0	0
moist	0	0	0	0.6894	0	0
testsg	0	0	0	0.7244	0	0
ovensg	-0.4739	0	0	0	0	0
ringtop	0	0	0	0	0	0.8517
ringbut	0	0	0	0	0	0.5240
bowmax	0	0	0	0	0.5033	0
bowdist	0	0.7976	0	0	0	0
whorls	0	0.0599	-0.4074	0	0	0
clear	0	0	0.9135	0	0	0
knots	0	0	0	0	0	0
diaknots	0.0832	-0.6002	0	0	0	0
CPEV(%)	23.42	40.60	53.69	62.76	70.25	75.47

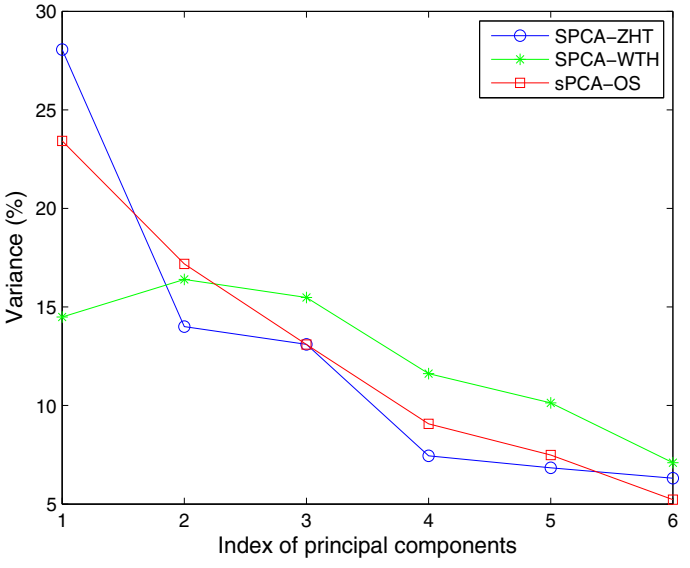

Fig. 1. Variation of variance (%) corresponding to the principal components on the `pitprops` dataset

recovering the same hidden factors. Moreover, the variance of nonzero loadings of sPCA-OS is less than those of the other two sparse PCA. This shows that our sPCA-OS should be more robust than SPCA-ZHT and SPCA-WTH. This agrees with that our sPCA-OS based on the optimal scoring can capture the underlying discriminative property

Table 4. Results of the simulation example: loadings and variance

	PCA		SPCA-ZHT		SPCA-WTH		sPCA-OS	
	$PC1$	$PC2$	$PC1$	$PC2$	$PC1$	$PC2$	$PC1$	$PC2$
X_1	-0.0901	-0.4767	0	0.3945	-0.0024	0	0	-0.4982
X_2	-0.0875	-0.5107	0	0.6348	-0.7270	0	0	-0.5028
X_3	-0.0886	-0.4719	0	0.4789	-0.4704	0	0	-0.5007
X_4	-0.0856	-0.4801	0	0.4604	-0.5001	0	0	-0.4983
X_5	0.4134	-0.0840	0.3616	0	0	-0.1826	0.5292	0
X_6	0.3948	-0.1266	0.3831	0	0	-0.8578	0.5328	0
X_7	0.3991	-0.1442	0.4218	0	0	-0.4115	0.4538	0
X_8	0.4047	-0.1173	0.7380	0	0	-0.2481	0.4798	0
X_9	0.3996	0.0270	0	0	0	0	0	0
X_{10}	0.3996	0.0221	0	0	0	0	0	0
CPEV(%)	61.03	94.01	60.86	75.27	36.50	72.71	61.59	79.66

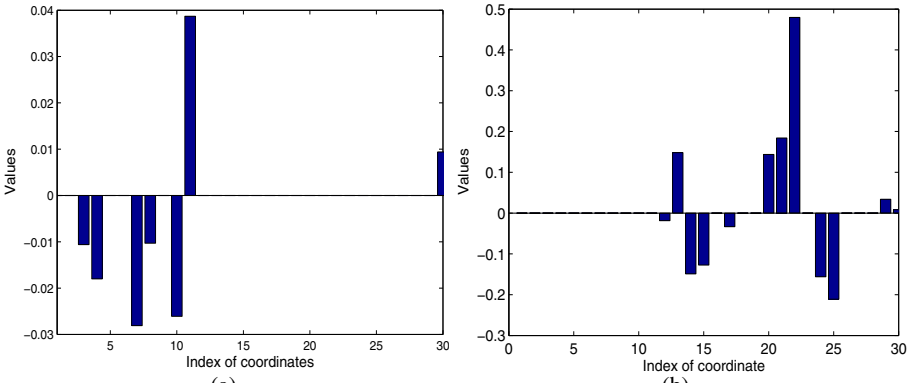

Fig. 2. Sparse results of SPCO on the synthetic dataset: (a) principal coordinate on the first principal component; (b) principal coordinate on the second principal component

in datasets. In addition, our CPEV is higher than the others, when all the sparse PCA algorithms keep the same number of nonzero loadings.

In order to reveal the effectiveness of SPCO, we also generate another synthetic dataset via the following Gaussian distributions:

$$G_1 \sim \mathcal{N}(10 * \mathbf{1}_{50},\ \mathbf{I}_{50}),\ G_2 \sim \mathcal{N}(-10 * \mathbf{1}_{50},\ 9 * \mathbf{I}_{50}),\ G_3 \sim \mathcal{N}(\mathbf{0},\ 17 * \mathbf{I}_{50}).$$

Obviously, G_2 is more close to G_3 in comparison with G_1. This synthetic dataset consists of 30 points with 50 variables, and each 10 points corresponds to one distribution. Without loss of generality, the points $(\mathbf{x}_{(i-1)*10+1}, \ldots, \mathbf{x}_{i*10})$ are sampled from $G_i, i = 1, 2, 3$. The results of implementing SPCO are shown in Figure 2, where the i-th coordinate correspond to the i-th point. Figure 2 (a) depicts that the nonzero coordinates correspond to the first 10 points, while the coordinates for the other points are zeros.

Figure 2 (b) depicts that the coordinates for the first 10 points are zeros, but the nonzero coordinates are associated with the other points. In essence, it also tells us an appealing property that SPCO can effectively capture the underlying relationship among data by strengthening zero or nonzero coordinates. We also carry out the similar experiments with the three sparse PCA algorithms. Unfortunately, the resulting principal coordinates are not sparse so that the property mentioned above does not remain.

5.3 Evaluations on Classification

In this experiment, we conduct the comparison of the four sparse dimensionality reduction algorithms in classification problems. We first implement sPCA-OS, SPCO, SPCA-ZHT and SPCA-WTH for dimensionality reduction. Then, we perform classification respectively on the dimensionality-reduced data matrices by simply applying nearest neighbor classifier.

Our experiments are implemented on six UCI datasets, the details of which are summarized in Table 5.

In order to make comparison fair, we keep CPEV as equal as possible when we carry out the four dimensionality reduction methods. This can be done by adjusting the regularization parameters. Note that for SPCO, there is no obvious effect on the explained variance when varying the regularization parameter λ. For each dataset, we randomly sample 90% of the instances for training and the remaining 10% for test. This procedure is repeated 20 times for each dataset, and the evaluation criteria are reported in the classification accuracy rate(%) and corresponding standard deviation.

The classification results are listed in Table 6. We find that when the explained variance obtained from the four sparse dimensionality reduction methods are nearly equal, our sPCA-OS and SPCO outperform SPCA-ZHT and SPCA-WTH on the whole. It also successfully confirms that as shown in [21], our sPCA-OS based on optimal scoring can effectively detect the underlying discriminative information among data, and that SPCO can also effectively detect the underlying distribution among data.

It should be also worth mentioning here that the sparsity of the loadings obtained by sPCA-OS, SPCA-ZHT, SPCA-WTH are different when keeping approximately same explained variance. In particular, Figure 3 depicts the sparsity of loadings obtained by sPCA-OS, SPCA-ZHT and SPCA-WTH on the six UCI datasets. Obviously, the similar conclusion for the `pitprops` dataset can be followed here.

Table 5. The summaries of datasets. (c—the number of classes, p—the number of the variables and n—the number of instances.).

Datasets	c	p	n
`dermatology`	6	34	358
`segmentation`	7	18	2310
`glass`	6	9	214
`letter`	10	16	1978
`pageblocks`	5	10	5473
`pendigits`	10	16	7494

Table 6. Classification results using the four sparse methods on the different datasets. ("CPEV" for "the cumulative percentage of explained variance", "acc" for "the accuracy of classification" and "std" for "the standard deviation".).

Dataset	CPEV	acc(± std)	Algorithm
dermatology	34.67	80.54 (±3.39)	SPCA-ZHT
	34.70	80.02 (±3.46)	SPCA-WTH
	34.64	82.43(±5.48)	sPCA-OS
	35.97	83.56(±5.48)	SPCO
segmentation	70.70	66.52(±2.25)	SPCA-ZHT
	73.00	70.80(±1.53)	SPCA-WTH
	73.35	77.80(±1.53)	sPCA-OS
	71.04	71.13(±0.94)	SPCO
glass	82.14	45.78(±5.95)	SPCA-ZHT
	82.10	45.81(±4.95)	SPCA-WTH
	82.01	46.87(±6.42)	sPCA-OS
	82.69	47.56(±6.00)	SPCO
letter	81.34	60.96(±2.74)	SPCA-ZHT
	81.30	63.04(±2.74)	SPCA-WTH
	81.67	64.26(±1.78)	sPCA-OS
	80.80	63.12(±2.93)	SPCO
pageblocks	67.83	93.83(±0.48)	SPCA-ZHT
	67.90	94.03(±0.31)	SPCA-WTH
	67.90	93.80(±0.45)	sPCA-OS
	67.62	94.35(±0.43)	SPCO
pendigits	82.86	93.02(±0.76)	SPCA-ZHT
	82.50	93.48(±0.57)	SPCA-WTH
	82.80	93.18(±0.58)	sPCA-OS
	82.63	93.67(±0.61)	SPCO

5.4 Application in Gene Microarray

The SRBCT microarray dataset [8] has 2308 genes (variables) and 63 samples (observations). We implement SPCO on this dataset to explore sparse representation of the original data in a dimensionality-reduced space. Here q is the number of principal coordinates and it is also the dimensionality of the low-dimensional space.

Similarly, a new evaluation criterion, i.e., *cumulative percentage of zero coordinates* (CPZC) is defined to measure SPCO, and it is the ratio of the number of zero coordinates to the total number of coordinates in the dimensionality-reduced data matrix.

A sequence of principal coordinates are taken from $q = 3$ to $q = 63$. As shown in Figure 4, the CPEV of the principal component $\mathbf{A}$ increases sharply while the CPZC of data matrix $\mathbf{Z}$ in the low-dimensional space dramatically declines, when q is set from 3 to 10. When $q > 20$, however, the obtained CPZC is nearly same while the CPEV still increases steadily. Thus, in practice, we can choose a suitable q through the trade-off between CPEV and CPZC.

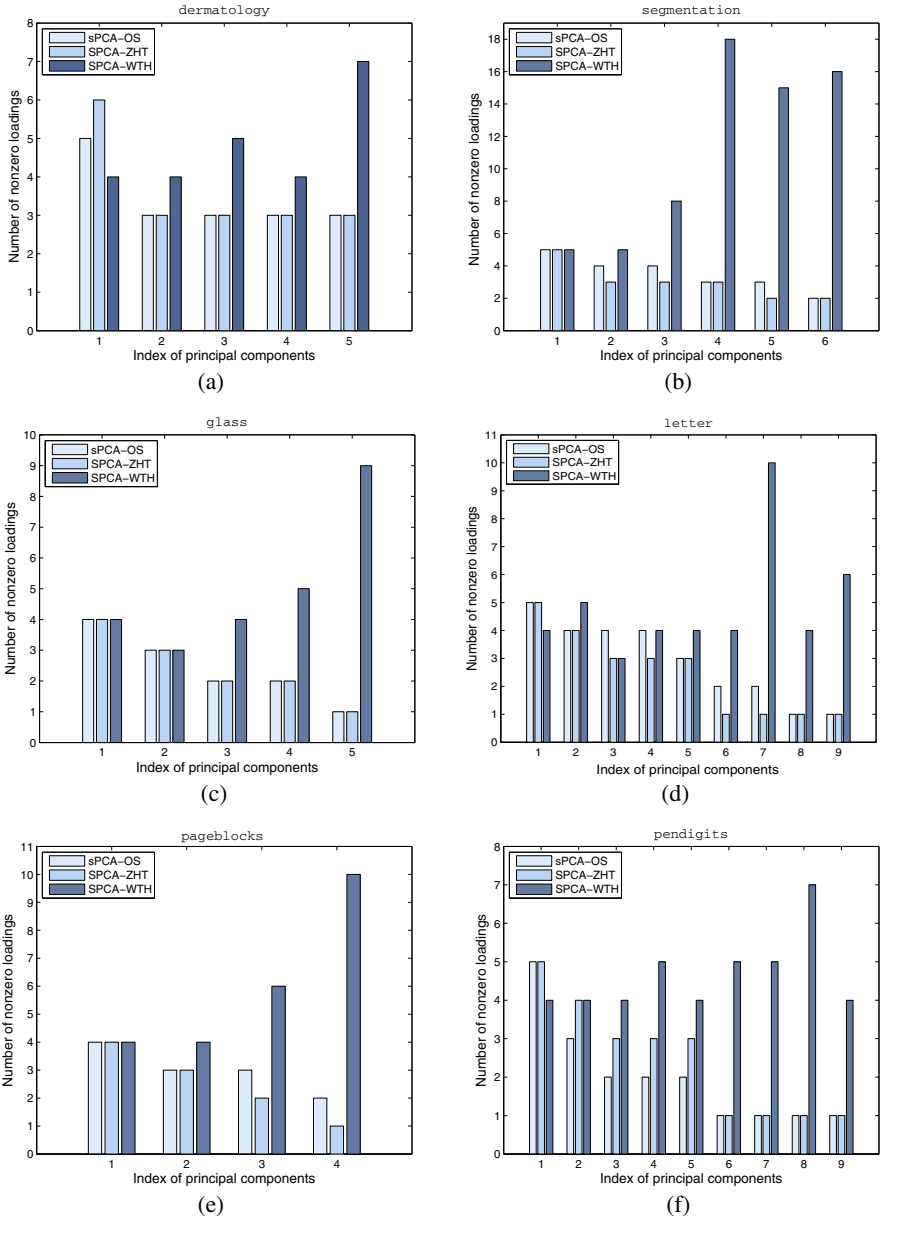

Fig. 3. The sparsity of loadings obtained by sPCA-OS, SPCA-ZHT and SPCA-WTH on the six UCI datasets

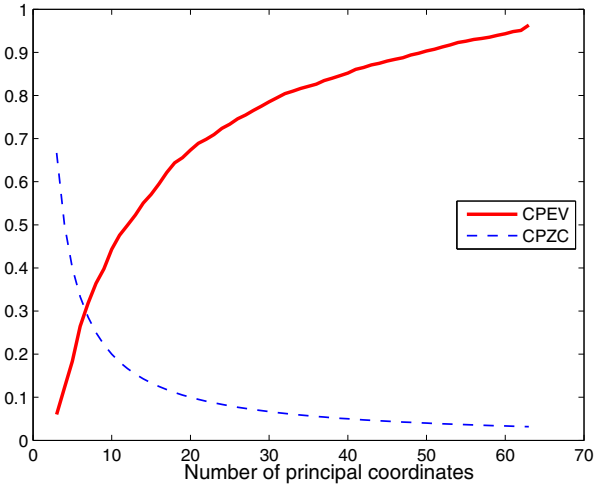

Fig. 4. Variation of CPEV and CPZC with respect to the number of principal coordinates

6 Conclusion

In this paper we have developed a new SPCA method to compute sparse principal components and an SPCO method to compute sparse principal coordinates. Our SPCA is built on the optimal scoring theorem, while SPCO is based on the Eckart-Young theorem. Since the optimal scoring theorem and the Eckart-Young theorem respectively provide the theoretical foundation to carry out the conventional PCA and PCO by regression-type problems, our SPCA and SPCO can be efficiently solved by existing algorithms for sparse regression models such that in [4,6,19,22]. There are a lot of treatments for sparse PCA in the literature. To our knowledge, however, we have first provided an attempt for sparse PCO. In our future work, we will dig out the potential application of our sparse PCO method in multivariate data analysis.

Acknowledgments. This work is supported by the Natural Science Foundations of China (No. 60970081) and the 973 Program of China (No. 2010CB327903). Zhihua Zhang also acknowledges support from Doctoral Program of Specialized Research Fund of Chinese Universities (No. J20091608) and from the Fundamental Research Funds for the Central Universities.

References

1. Clemmensen, L., Hastie, T., Erbøll, B.: Sparse discriminant analysis. Technical report (June 2008)
2. d'Aspremont, A., El Ghaoui, L., Jordan, M.I., Lanckriet, G.R.G.: A direct formulation for sparse PCA using semidefinite programming. SIAM Review 49(3), 434–448 (2007)

3. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218 (1936)
4. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression (with discussions). The Annals of Statistics 32(2), 407–499 (2004)
5. Elden, L., Park, H.: A procrustes problem on the stiefel manifold. Numerische Mathematik (1999)
6. Friedman, J.H., Hastie, T., Hoefling, H., Tibshirani, R.: Pathwise coordinate optimization. The Annals of Applied Statistics 2(1), 302–332 (2007)
7. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore (1996)
8. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–536 (1999)
9. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate data analysis. Biometrika 53, 315–328 (1966)
10. Gower, J.C., Dijksterhuis, G.B.: Procrustes Problems. Oxford University Press, Oxford (2004)
11. Hastie, T., Tibshirani, R., Buja, A.: Flexible discriminant analysis by optimal scoring. Journal of the American Statistical Association 89(428), 1255–1270 (1994)
12. Jeffers, J.: Two case studies in the application of principal compnent. Appl. Statist. 16, 225–236 (1967)
13. Jolliffe, I.T.: Principal component analysis, 2nd edn. Springer, New York (2002)
14. Jolliffe, I.T., Trendafilov, N.T., Uddin, M.: A modified principal component technique based on the lasso. Journal of Computational and Graphical Statistics 12, 531 (2003)
15. Magnus, J.R., Neudecker, H.: Matrix Calculus with Applications in Statistics and Econometric. John Wiley & Sons, New York (1999) (revised edn.)
16. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press, New York (1979)
17. Shen, H., Huang, J.: Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis 99, 1015–1034 (2008)
18. Sriperumbudur, B.K., Torres, D., Lanckriet, G.R.G.: Sparse eigen methods by d.c. programming. In: ICML (2007)
19. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1996)
20. Witten, M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatisitics 10(3), 515–534 (2009)
21. Zhang, Z., Dai, G.: Optimal scoring for unsupervised learning. In: Advances in Neural Information Processing Systems, vol. 23 (2009)
22. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301–320 (2005)
23. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics 15, 265–286 (2006)

A The Proof of Theorem 2

Consider the Lagrange function

$$L(\mathbf{Z}, \mathbf{A}, \mathbf{C}) = \frac{1}{2}\Big[\|\mathbf{X}-\mathbf{Z}\mathbf{A}^T\|_F^2 + \gamma\|\mathbf{Z}\|_F^2 - \mathsf{tr}(\mathbf{C}(\mathbf{A}^T\mathbf{A}-\mathbf{I}_q))\Big],$$

where $\mathbf{C}$ is a $q\times q$ symmetric matrix of Lagrange multipliers. The first-order conditions are

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{Z}} &= (\mathbf{Z}\mathbf{A}^T - \mathbf{X})\mathbf{A} + \gamma\mathbf{Z} = \mathbf{0},\\
\frac{\partial L}{\partial \mathbf{A}} &= (\mathbf{A}\mathbf{Z}^T - \mathbf{X}^T)\mathbf{Z} - \mathbf{A}\mathbf{C} = \mathbf{0},\\
\frac{\partial L}{\partial \mathbf{C}} &= \mathbf{A}^T\mathbf{A} - \mathbf{I}_q = \mathbf{0},
\end{aligned} \tag{8}$$

which yield

$$\mathbf{Z} = \frac{1}{1+\gamma}\mathbf{X}\mathbf{A}.$$

Premultiplying both sides of (8) by $\mathbf{A}^T$ then gives

$$\mathbf{C} = \mathbf{Z}^T\mathbf{Z} - \mathbf{Z}^T\mathbf{X}\mathbf{A} = -\gamma\mathbf{Z}^T\mathbf{Z}.$$

Hence, (8) can be rewritten as

$$(\mathbf{X}^T\mathbf{X})\mathbf{A} = \mathbf{A}(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}).$$

Suppose that the full SVD of $\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}$ is $\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A} = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T$ where $\mathbf{G}$ is an orthogonal $q\times q$ matrix and $\boldsymbol{\Lambda}$ is a diagonal $q\times q$ matrix with the eigenvalues (singular values) of $\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}$ on its diagonal. Then we have

$$(\mathbf{X}^T\mathbf{X})\mathbf{A}\mathbf{G} = \mathbf{A}\mathbf{G}\boldsymbol{\Lambda},$$

which implies that $\mathbf{A}\mathbf{G}$ is the orthogonal eigenvector matrix of $\mathbf{X}^T\mathbf{X}$ and the diagonal entries of $\boldsymbol{\Lambda}$ are its corresponding eigenvalues.

Now given $\mathbf{Z} = \frac{1}{1+\gamma}\mathbf{X}\mathbf{A}$, we have

$$\begin{aligned}
g(\mathbf{A}, \mathbf{Z}) &= \mathsf{tr}(\mathbf{X}^T\mathbf{X}) - \frac{1}{1+\gamma}\mathsf{tr}(\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A})\\
&= \mathsf{tr}(\mathbf{X}^T\mathbf{X}) - \frac{1}{1+\gamma}\mathsf{tr}(\boldsymbol{\Lambda}).
\end{aligned}$$

In order to minimize g, we should maximize $\mathsf{tr}(\boldsymbol{\Lambda})$. We thus take $\boldsymbol{\Lambda} = \mathsf{diag}(d_1, \ldots, d_q)$. Moreover, we can also let $\mathbf{A} = [\mathbf{v}_1, \ldots, \mathbf{v}_q]$.