SPARSE PCA VIA HARD THRESHOLDING FOR BLIND SOURCE SEPARATION

Ming-Chun Wu and Kwang-Cheng Chen

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan
Email : { r02942050, ckc } @ntu.edu.tw

ABSTRACT

Principal Component Analysis (PCA) is adopted in diverse areas including signal processing and machine leaning. However, the derived principal components, the linear combinations of the original variables, are hard to be interpreted in many applications especially the blind source separation. Therefore, we propose regularized PCA via hard thresholding such that the derived loadings are sparse and easier to be interpreted. The proposed method has advantages due to the adoption of hard thresholding. First, the proposed method can be implemented by linear operators and thus computational efficient even in $p \gg n$ or large p scenarios. Second, the threshold can be objectively selected based on statistical decision theory without domain knowledge. Moreover, simulations show the superiority of our method compared to the L_1-penalized method. Therefore, our approach can be a strong competitor of the existing sparse PCA.

Index Terms— Principal component analysis (PCA), sparse PCA, regularization, hard thresholding, blind source separation.

1. INTRODUCTION

Principal Component Analysis (PCA) is adopted for diverse applications in many areas like signal processing and machine learning due to its computational efficiency and theoretic insights. PCA can be considered as a fundamental method of matrix decomposition and is briefly described below. Without loss of generality, assume that we have a $n \times p$ data matrix $\mathbf{X}$ such that all columns of $\mathbf{X}$ have zero mean. Apply the eigenvalue decomposition, then

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T. \tag{1}$$

Each column of $\mathbf{V}$ is a loading and each row of $\mathbf{U}\mathbf{\Lambda}$ represents a *Principal Component* (PC).

For dimension reduction, the first $k < \min(n, p)$ PCs are selected to represent the original data $\mathbf{X}$. However, the derived PCs, the linear combinations of the original p variables, are hard to be interpreted [1, 2]. One of the reasons is that PCs over-fit to noise and thus almost all the elements of $\mathbf{V}$ are nonzero. The other is that the loadings are orthogonal. The undesired properties of PCA degrade its performance in applications like the blind source separation.

Some works modify PCA by introducing regularizations such that the derived loading are sparse [3]. Hence, each resulting PC consists of a small subset of variables and thus are easier to be interpreted. An pioneer approach directly applies the L_1-penalty on the loadings [1]. The L_1-penalty shrinks the entries of loadings to zero until a sparse solution is derived as in the LASSO regression [4]. A more sophisticated work proved that PCA can be formulated as a regression type optimization, then we can obtain sparse loadings by solving a L_1 and L_2-penalized regression [2]. The resulting *Sparse PCA* (SPCA) is a promising method and is considered as the benchmark.

The SPCA requires additional algorithm to handle the *elastic net* optimization, L_1 and L_2-penalized regression, in each iteration [5]. Moreover, there are k tuning parameters, each one controls the sparsity of a single PC. The choice of the tuning parameters usually requires subjective domain knowledge of users or further searching steps. Another drawback of the SPCA is that the L_1-penalty introduces distortions by shrinking the entries of loading matrix to zero. Therefore, we propose a novel method, *Sparse PCA via Hard Thresholding* (SPCA-HT), in this paper.

Although the L_1-penalty has attractive ability of denoising [6], it introduces additional distortions by shrinking the elements of $\mathbf{V}$ to zero. Therefore, the proposed SPCA-HT uses hard thresholding to regularize PCA. Our approach only requires one tuning parameter, which is the hard threshold, and can be objectively determined by following statistical decision theory. Moreover, there are two benefits arising from the relief of L_1-penalty. 1) The SPCA-HT algorithm only requires linear operations and thus is computational efficient even in $p \gg n$ or large p scenarios. 2) Simulations show that the SPCA-HT better estimates principal directions and thus explains more variance of the data, since it does not shrink the coefficient of $\mathbf{V}$ to zero as the LASSO based methods do.

This paper considers blind source separation, which is frequently encountered in signal processing [7], as an example to illustrated the superiority of the proposed SPCA-HT compared with the SPCA. The simulations show that the SPCA-HT better estimates the principal directions and has higher explained variance than the LASSO based SPCA under the

same sparsity of $\mathbf{V}$. Therefore, the SPCA-HT can be a potential solution of blind source separation.

2. PRELIMERARIES

The first k columns of $\mathbf{V}$ in (1) can be obtained by solving the L_2-penalized optimization as described below.

Theorem 1 (PCA and Regularized Optimization [2])**.** *Let $\mathbf{A}$ and $\mathbf{B}$ both be $p \times k$ matrix such that $\mathbf{A} = [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_k]$ and $\mathbf{B} = [\boldsymbol{b}_1 \ \cdots \ \boldsymbol{b}_k]$. For any $\lambda > 0$, let $(\mathbf{A}^*, \mathbf{B}^*)$ be the solution of the optimization*

$$\begin{array}{ll} \underset{\mathbf{A},\mathbf{B}}{\textit{minimize}} & \sum_{j=1}^{k} \left\{ \|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\boldsymbol{b}_j\|^2 + \lambda \|\boldsymbol{b}_j\|^2 \right\} \\ \textit{subject to} & \mathbf{A}^T\mathbf{A} = \mathbf{I}. \end{array} \tag{2}$$

Then, $\boldsymbol{b}_j^*/\|\boldsymbol{b}_j^*\| = \boldsymbol{v}_j, \forall j = 1, 2, \ldots, k$.

The SPCA introduces the L_1-penalty into (2).

Definition 1 (Sparse PCA)**.** *Let $\mathbf{A}$ and $\mathbf{B}$ both be $p \times k$ matrix such that $\mathbf{A} = [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_k]$ and $\mathbf{B} = [\boldsymbol{b}_1 \ \cdots \ \boldsymbol{b}_k]$. Given $\lambda > 0$, nonnegative constants $\lambda_j, j = 1, 2, \ldots k$, and let $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ be the solution of the optimization*

$$\begin{array}{ll} \underset{\mathbf{A},\mathbf{B}}{\textit{minimize}} & \sum_{j=1}^{k} \left\{ \|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\boldsymbol{b}_j\|^2 + \lambda \|\boldsymbol{b}_j\|^2 + \lambda_j \|\boldsymbol{b}_j\|_1 \right\} \\ \textit{subject to} & \mathbf{A}^T\mathbf{A} = \mathbf{I}. \end{array} \tag{3}$$

Then, the sparse loadings are $\hat{\boldsymbol{v}}_j = \hat{\boldsymbol{b}}_j/\|\hat{\boldsymbol{b}}_j\|, \forall j$.

An iterative algorithm of SPCA is shown in Algorithm 1.

Algorithm 1 SPCA [2]

1: Initialize with $\mathbf{A} = [\boldsymbol{v}_1 \ \boldsymbol{v}_2 \cdots \boldsymbol{v}_k]$ where $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$.
2: **Update B given A:** For $j = 1, 2, \ldots, k$, update $\boldsymbol{b}_j$ by solving the elastic net optimization

$$\underset{\boldsymbol{b}}{\text{minimize}} \quad \|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\boldsymbol{b}\|^2 + \lambda_2 \|\boldsymbol{b}\|^2 + \lambda_{1,j} \|\boldsymbol{b}\|_1. \tag{4}$$

3: **Update A given B:** Derive the SVD, $\mathbf{X}^T\mathbf{X}\mathbf{B} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^T$, then $\mathbf{A} \leftarrow \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$.
4: Repeat step 2 and 3 until convergence.
5: $\boldsymbol{v}_j \leftarrow \boldsymbol{b}_j/\|\boldsymbol{b}_j\|, \forall j = 1, 2, \ldots, k$.

3. MAIN RESULTS

The L_1-penalty plays an important role in the SPCA. In the $L_q, q \in \mathbb{N}$ penalty family, [8] proves that only the LASSO (L_1-penalty) can produce a sparse solution and thus is capable of de-noising. That is the L_1-penalty prevents PCA from over-explaining the variance of noise. However, there are benefits after substituting hard thresholding for the L_1-penalty. 1) The elastic net (4) is reduced to a *ridge regression* [9] and can be solved by linear operations. 2) Hard thresholding does not introduce additional distortions of $\mathbf{V}$ by shrinking the elements to zero as the LASSO does. Therefore, we adopt hard-thresholding for the proposed method.

3.1. Sparse PCA via Hard-Thresholding

We use a regularization matrix to control the sparsity of $\mathbf{V}$.

Definition 2 (Regularization Matrix $\mathbf{G}$)**.** *Any matrix $\mathbf{G} \in \{0, 1\}^{p \times k}$ can be a valid regularization matrix and $[\mathbf{V}]_{ij} = 0$ if $[\mathbf{G}]_{ij} = 0$.*

With a well-designed regularization matrix $\mathbf{G}$, we can replace the L_1-penalty in (3) by the zero constraints introduced by $\mathbf{G}$ and still be able to derive sparse loadings.

Definition 3 (Spare PCA via Hard Thresholding)**.** *Given a sparse regularization matrix $\mathbf{G}$. For any $\lambda > 0$, let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ be the solution of the optimization*

$$\begin{array}{ll} \underset{\mathbf{A},\mathbf{B}}{\textit{minimize}} & \sum_{j=1}^{k} \left\{ \|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\boldsymbol{b}_j\|^2 + \lambda \|\boldsymbol{b}_j\|^2 \right\} \\ \textit{subject to} & \mathbf{A}^T\mathbf{A} = \mathbf{I}, \ [\mathbf{B}]_{ij} = 0 \textit{ if } [\mathbf{G}]_{ij} = 0. \end{array} \tag{5}$$

Then, the sparse loadings are $\tilde{\boldsymbol{v}}_j = \tilde{\boldsymbol{b}}_j/\|\tilde{\boldsymbol{b}}_j\|, \forall j$.

Moreover, we show that the SPCA-HT can be obtained by an algorithm using linear operations, which is critical for computational efficiency.

Theorem 2 (Equivalent Form of the SPCA-HT)**.** *Let $\mathbf{D}_j$ be the diagonal matrix with $[\mathbf{D}_j]_{ii} = [\mathbf{G}]_{ij}$, that is $\mathbf{D}_j = \mathrm{diag}(\boldsymbol{g}_j)$. Then $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ in Definition 3 can be obtained by solving*

$$\begin{array}{ll} \underset{\mathbf{A},\mathbf{B}}{\textit{minimize}} & \sum_{j=1}^{k} \left\{ \|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\mathbf{D}_j\boldsymbol{b}_j\|^2 + \lambda \|\boldsymbol{b}_j\|^2 \right\} \\ \textit{subject to} & \mathbf{A}^T\mathbf{A} = \mathbf{I}. \end{array} \tag{6}$$

Therefore, the elastic net (4) of Algorithm 1 is reduced to a ridge regression with solution

$$\boldsymbol{b}_j = (\mathbf{D}_j\mathbf{X}^T\mathbf{X}\mathbf{D}_j + \lambda\mathbf{I})^{-1}\mathbf{D}_j\mathbf{X}^T\mathbf{X}\boldsymbol{a}_j. \tag{7}$$

Moreover, when $p \gg n$, (7) is further simplified to

$$\boldsymbol{b}_j = \mathbf{D}_j\mathbf{X}^T\mathbf{X}\boldsymbol{a}_j. \tag{8}$$

Proof of Theorem 2 . Let $\boldsymbol{b}_j = [b_{1j} \ b_{2j} \cdots b_{pj}]^T$ and $\mathbf{D}_j\boldsymbol{b}_j = [\tilde{b}_{1j} \ \tilde{b}_{2j} \cdots \tilde{b}_{pj}]^T$, then

$$\tilde{b}_{ij} = \begin{cases} b_{ij} & \text{, if } [\mathbf{D}_j]_{ii} = 1. \\ 0 & \text{, otherwise.} \end{cases} \tag{9}$$

$$= \begin{cases} b_{ij} & \text{, if } [\mathbf{G}]_{ij} = 1. \\ 0 & \text{, otherwise.} \end{cases} \tag{10}$$

Now consider the optimization, for any $\lambda > 0$

$$\underset{\boldsymbol{a}_j, \boldsymbol{b}_j}{\text{minimize}} \quad \underbrace{\|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\mathbf{D}_j\boldsymbol{b}_j\|^2}_{(a)} + \underbrace{\lambda\|\boldsymbol{b}_j\|^2}_{(b)}. \tag{11}$$

From (10) we know that b_{ij} has no influence on term (a) of (11). Hence, b_{ij} should be zero if $[\mathbf{G}]_{ij} = 0$ due to the penalty term (b). Note that $b_{ij} = [\mathbf{B}]_{ij}$, then (11) is equivalent to

$$\begin{array}{ll} \underset{\boldsymbol{a}_j, \boldsymbol{b}_j}{\text{minimize}} & \|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\boldsymbol{b}_j\|^2 + \lambda\|\boldsymbol{b}_j\|^2 \\ \text{subject to} & [\mathbf{B}]_{ij} = 0 \text{ if } [\mathbf{G}]_{ij} = 0. \end{array} \tag{12}$$

Combine all j and add the orthonormal constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, then we show that (5) is equivalent to (6). Therefore, the elastic net (4) is reduced to the ridge regression

$$\boldsymbol{b}_j = \arg\min_{\boldsymbol{b}} \|\mathbf{X}\boldsymbol{a}_j - \mathbf{X}\mathbf{D}_j\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|^2. \tag{13}$$

After doing some linear algebra, one can obtain

$$\boldsymbol{b}_j = (\mathbf{D}_j\mathbf{X}^T\mathbf{X}\mathbf{D}_j + \lambda\mathbf{I})^{-1}\mathbf{D}_j\mathbf{X}^T\mathbf{X}\boldsymbol{a}_j. \tag{14}$$

For $p \gg n$, note that λ is arbitrary nonnegative constant. Let $\lambda \to \infty$, we have $\boldsymbol{b}_j \propto \mathbf{D}_j\mathbf{X}^T\mathbf{X}\boldsymbol{a}_j$. Since $\tilde{\boldsymbol{v}}_j = \tilde{\boldsymbol{b}}_j/\|\tilde{\boldsymbol{b}}_j\|$, we can let $\boldsymbol{b}_j = \mathbf{D}_j\mathbf{X}^T\mathbf{X}\boldsymbol{a}_j$. □

Hence, the SPCA-HT can be obtained by Algorithm 2.

Algorithm 2 Proposed SPCA-HT

1: Given $\mathbf{G} = [\boldsymbol{g}_1\ \boldsymbol{g}_2 \cdots \boldsymbol{g}_k]$ and initialize with $\mathbf{A} = [\boldsymbol{v}_1\ \boldsymbol{v}_2 \cdots \boldsymbol{v}_k]$ where $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$.
2: **Update B given A:** $\forall j = 1, 2, \ldots, k$, $\mathbf{D}_j \leftarrow \text{diag}(\boldsymbol{g}_j)$.

$$\boldsymbol{b}_j \leftarrow \begin{cases} \mathbf{D}_j\mathbf{X}^T\mathbf{X}\boldsymbol{a}_j & \text{, if } p \gg n. \\ (\mathbf{D}_j\mathbf{X}^T\mathbf{X}\mathbf{D}_j + \lambda\mathbf{I})^{-1}\mathbf{D}_j\mathbf{X}^T\mathbf{X}\boldsymbol{a}_j & \text{, otherwise.} \end{cases}$$

3: **Update A given B:** Derive the SVD, $\mathbf{X}^T\mathbf{X}\mathbf{B} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^T$, then $\mathbf{A} \leftarrow \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$.
4: Repeat step 2 and 3 until convergence.
5: $\boldsymbol{v}_j \leftarrow \boldsymbol{b}_j/\|\boldsymbol{b}_j\|, \forall j = 1, 2, \ldots, k$.

3.2. The Regularization Matrix G

For data reduction, we want to choose a small number of PCs that explain most of the variance of the data. On the other hand, for sparse PCA, each PC should only consist of a small number of highly related variables. Therefore, we summarize: 1) Suppose variable i has large variance and thus significant. There should be at least one principal component (PC) consist of variable i. 2) Highly correlated variables should be explained by the same PC.

To identify highly correlated variables, we have the hypotheses, for all i, j

$$\begin{array}{ll} H_{0,ij} & : \text{ variable } i \text{ and } j \text{ are not correalted.} \\ H_{1,ij} & : \text{ not } H_{0,ij}. \end{array} \tag{15}$$

Without loss of generality, we consider two-tailed tests in this paper, $H_{0,ij} : \rho_{ij} = 0$ and $H_{1,ij} : \rho_{ij} \neq 0$, where ρ_{ij} is the correlation coefficient of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The standard testing statistic is the sample correlation coefficient of variable i and j defined as $\hat{\rho}_{ij} = \frac{\sum_{l=1}^n (x_{li}-\bar{x}_i)(x_{lj}-\bar{x}_j)}{\|\boldsymbol{x}_i - \bar{x}_i\mathbf{1}\|\|\boldsymbol{x}_j - \bar{x}_j\mathbf{1}\|}$ where $\bar{x}_i = \frac{1}{n}\sum_{l=1}^n x_{li}$. The resulting decision rule is rejecting $H_{0,ij}$ if and only if $|\hat{\rho}_{ij}| \geq \rho$ where ρ is the decision threshold. The proposed method uses ρ to perform hard thresholding and the choice of ρ can be objectively judged by statistical decision theories [10, 11]. With appropriate choice of ρ, we have the decision matrix $\mathbf{H}$.

Definition 4 (Decision Matrix $\mathbf{H}$)**.**
The matrix $\mathbf{H} \in \{0, 1\}^{p\times p}$ is a decision matrix when $[\mathbf{H}]_{ij} = 1$ if and only if $H_{0,ij}$ is rejected.

Suppose variable j has large variance, at least one PC should consist of variable j along with its highly correlated variables which are indicated by the j-th column of $\mathbf{H}$, $\boldsymbol{h}_j$. Therefore, $\boldsymbol{h}_j$ indicates the nonzero elements of the corresponding loadings and $\boldsymbol{h}_j$ should be a column of $\mathbf{G}$. With the same argument, we can find another significant variable without being selected, then forms a new column of $\mathbf{G}$. Repeat the procedures until k columns of $\mathbf{H}$ are chosen, we then have the regularization matrix $\mathbf{G}$. Hence, we propose Algorithm 3 to implement the rational above.

Algorithm 3 Regularization Matrix $\mathbf{G}$

1: Let α be a permutation of $\{1, 2, \ldots, p\}$ such that $\text{var}(\boldsymbol{x}_{\alpha_1}) \geq \text{var}(\boldsymbol{x}_{\alpha_2}) \geq \ldots \text{var}(\boldsymbol{x}_{\alpha_p})$ and $\mathbf{G}$ is empty.
2: **while** number of columns of $\mathbf{G} < k$ **do**
3: **if** $j = 1, 2, \ldots p$ and $\boldsymbol{h}_{\alpha_j}$ has not been chosen **then**
4: $\mathbf{G}$ has new column $\boldsymbol{h}_{\alpha_j}$.
5: **end if**
6: **end while**

4. SIMULATION STUDY

The general model of blind source separation in signal processing is $\mathbf{X}^T = \mathbf{F}\mathbf{S} + \mathbf{W}$ [12]. Suppose there are k sources, then the $k \times n$ matrix $\mathbf{S}$ is the signal matrix with i-th row being the time series of signals of i-th source. The $p \times k$ matrix $\mathbf{F}$ is the signature/mixing matrix with j-th column being the signature vector of j-th source. The $p \times n$ noise matrix $\mathbf{W}$ is uncorrelated to $\mathbf{S}$. The challenge is to estimate $\mathbf{F}$ and $\mathbf{S}$ simultaneously. Suppose $\mathbf{F}$ is normalized such that all columns have unit Euclidean norm, then we can use sparse loadings to estimate $\mathbf{F}$.

To compare the SPCA and the SPCA-HT, one important performance measure is the explained variance of the derived PCs. Since both methods may produce linear correlated PCs, we use *Adjusted Explain Variance* (AEV) instead of explained variance. AEV adjusts the overrated explained variance by

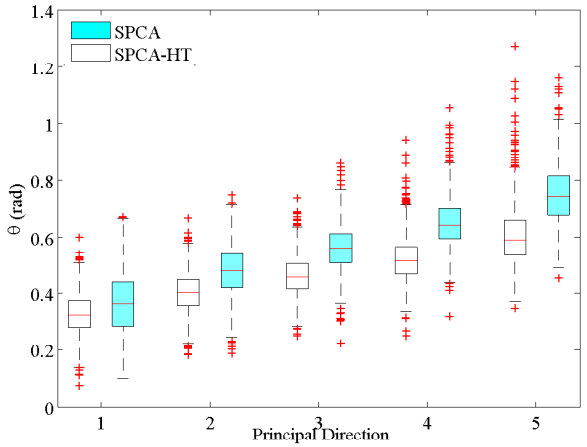

Fig. 1. Box plot of estimation accuracy of sparse principal directions. Note that θ is the angle between the true signature $\boldsymbol{f}$ and the estimated one $\boldsymbol{v}$. The LASSO based SPCA introduces distortions of the coefficients of principal directions due to its shrink-to-zero property. In contrast, the SPCA-HT does not have such drawbacks and thus has superior performance.

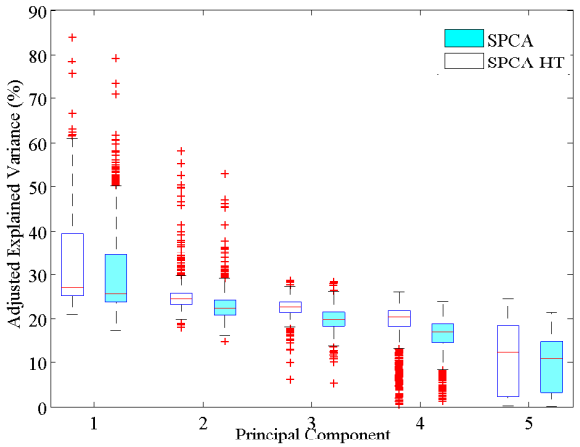

Fig. 2. Box plot of percentage adjusted explained variance. The percentage is respect to the total variance of the noiseless signal $\mathbf{S}$. Since Figure 1 has already shown the superior estimation performance of the SPCA-HT, the higher AEV of the SPCA-HT in Figure 2 is due to the better preservation of the signal variance.

considering the linear dependency between PCs. Apply the QR decomposition to the estimated PC, $\mathbf{X}^T\widehat{\mathbf{V}} = \mathbf{QR}$ where $\mathbf{R}$ is the triangular matrix. The i-th PC has $\text{AEV}_i = [\mathbf{R}]_{ii}^2$. The other performance measure is the estimation accuracy. Suppose we use $\boldsymbol{v}_j$ to estimate $\boldsymbol{f}_j$, the j-th column of $\mathbf{F}$, we use the angle between $\boldsymbol{v}_j$ and $\boldsymbol{f}_j$, $\theta_j = \arccos(\boldsymbol{v}_j^T\boldsymbol{f}_j)$, to evaluate the estimation performance. Since both $\boldsymbol{v}_j$ and $\boldsymbol{f}_j$ are normalized to have unit L_2 norm, using θ_j is equivalent to using the mean squared error.

In simulations, we use sinusoidal signals such that $[\mathbf{S}]_{ij} = a\sin(2\pi\frac{i}{n}j)$, randomly generated $\mathbf{F}$ with normalized columns and AWGN, $[\mathbf{W}]_{ij} \overset{i.i.d.}{\sim} N(0,\sigma^2)$. We consider the $p \gg n$ scenario to illustrate the effectiveness and efficiency of the proposed method. With $p = 300, n = 30, k = 5$, $\text{SNR} = 17\text{dB}$ and $\rho = 0.7$, the comparisons are shown in Figure 1 and 2.

In Figure 1, the SPCA-HT is a better estimator of $\mathbf{F}$, that is, has small θ. The reason is that although the LASSO well estimates the strong coefficient of $\mathbf{F}$, it introduces distortions to the medium level coefficients due to the shrink-to-zero property. In contrast, the SPCA-HT has no such drawbacks and thus has superior performance of estimating $\mathbf{F}$.

For both methods, Figure 2 shows that the first PC explains above 25% of the variance of the noiseless signal $\mathbf{S}$ in most of the time. It seems unsatisfactory that the first PC only explains 25% variance in PCA. However, due to the inability of de-noising, PCA usually overfits to noise by explaining too much variance of noise especially in large p scenarios. Hence, a 25% AEV is indeed good for a sparse PC. Since Figure 1 has already shown the superior estimation performance of the SPCA-HT, the higher AEV of the SPCA-HT in Figure 2 is due to the better preservation of the signal variance. Unfortunately, due to the limitation of papers, more data examples cloud not be presented here.

5. CONCLUSIONS AND FINAL REMARKS

We propose a regularized PCA based on hard thresholding to produce sparse principal directions. Compared to LASSO based methods, the proposed SPCA-HT enjoys some advantages due to the relief of the L_1-penalty. First, it is easy to implement since only linear operations are required in the algorithm. Second, the threshold ρ, the tuning parameter, can be objectively chosen based on statistical decision theory. Third, the hard thresholding does not introduces additional distortions of the principal directions by shrinking them to zero as the LASSO does. In application to blind source separation, the proposed SPCA-HT outperforms the LASSO based SPCA in simulations and thus has great potential to be a promising approach. This work can be considered as an early development of a general method of regularized PCA via hard thresholding. However, further efforts are required to develop both theories and algorithms to construct a compact methodology of sparse PCA.

6. REFERENCES

[1] 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.

[2] 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.

[3] M. O. Ulfarsson and V. Solo, "Vector l_0 sparse variable pca," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 1949–1958, May 2011.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[5] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[6] D. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[8] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, 2001.

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[10] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[11] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[12] A. Cichocki, R. Zdunek, and S.-I. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proceedings. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, May 2006.