Statistica Sinica **22** (2012), 933-956
doi:http://dx.doi.org/10.5705/ss.2010.149

PRINCIPAL COMPONENT ANALYSIS IN VERY HIGH-DIMENSIONAL SPACES

Young Kyung Lee[1], Eun Ryung Lee[2] and Byeong U. Park[2]

[1]*Kangwon National University and* [2]*Seoul National University*

Abstract: Principal component analysis (PCA) is widely used as a means of dimension reduction for high-dimensional data analysis. A main disadvantage of the standard PCA is that the principal components are typically linear combinations of all variables, which makes the results difficult to interpret. Applying the standard PCA also fails to yield consistent estimators of the loading vectors in very high-dimensional settings where the dimension of the data is comparable to, or even larger than, the sample size. In this paper we propose a modification of the standard PCA that works for such high-dimensional data when the loadings of principal components are sparse. Our method starts with an initial subset selection, and then performs a penalized PCA based on the selected subset. We show that our procedure identifies correctly the sparsity of the loading vectors and enjoys the oracle property, meaning that the resulting estimators of the loading vectors have the same first-order asymptotic properties as the oracle estimators that use knowledge of the indices of the nonzero loadings. Our theory covers a variety of penalty schemes. We also provide some numerical evidence of the proposed method, and illustrate it through gene expression data.

Key words and phrases: Adaptive lasso, eigenvalues, eigenvectors, high-dimensional data, MC penalization, penalized principal component analysis, SCAD, sparsity.

1. Introduction

Let $\mathbf{X}^1, \ldots, \mathbf{X}^n$ be n observations on a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$. In this paper, we are interested in estimating the principal components of $\mathbf{X}$ based on the observations $\mathbf{X}^i$. Principal components are $\boldsymbol{\beta}_j^\top \mathbf{X}$ for those orthonormal loading vectors $\boldsymbol{\beta}_j$ that give the largest variances. The orthonormal loading vectors are the eigenvectors of the covariance matrix $\boldsymbol{\Sigma} \equiv \mathrm{var}(\mathbf{X})$ that correspond to the largest eigenvalues. Principal component analysis (PCA) has wide applications ranging from biology to finance and offers a way to compress the data with minimal information loss by capturing directions of maximal variance in the data. A particular disadvantage of PCA is that the principal components are typically linear combinations of all variables X_j, which makes the results difficult to interpret, especially when d is very large. Recent years have seen several proposals that give 'sparse' solutions, that is, solutions that

involve only a few nonzero loadings; see Jolliffe, Trendafilov, and Uddin (2003), Zou, Hastie, and Tibshirani (2006), d'Aspremont et al. (2007), d'Aspremont, Bach, and Ghaoui (2008), Shen and Huang (2008), Leng and Wang (2009), and Witten, Tibshirani, and Hastie (2009).

We are concerned with the case where d, the dimension of $\mathbf{X}$, is comparable to, or even larger than, the sample size n. The standard PCA is known to yield inconsistent results in such a high-dimensional case, see Johnstone and Lu (2009). We propose a method that gives consistent estimators of the principal component loading vectors. The underlying assumption for our work is that the principal loading vectors $\boldsymbol{\beta}_j$ are sparse, which means that $\beta_{j\ell} = 0$ for all but a finite number of $1 \leq \ell \leq d$. The resulting estimators correctly identify the nonzero loadings, and have the same first-order asymptotic properties as the oracle estimators which use knowledge of the nonzero loadings.

Our method consists of two steps: initial dimension reduction; performing a penalized PCA. In the first step, the method tries to choose the variables X_ℓ such that $\beta_{j\ell} \neq 0$ for some j. In the second, it solves a penalized PCA optimization problem, using the observations on those variables chosen in the first step, to extract a given number of leading eigenvectors $\boldsymbol{\beta}_j$. The theory for the method is developed for a general penalty scheme, and covers various choices of penalty such as the adaptive lasso of Zou (2006), the smoothly clipped absolute deviation (SCAD) of Fan and Li (2001), and the minimax concave (MC) penalty of Zhang (2010). A numerical algorithm for the second step is also provided.

The idea of initial dimension reduction is inspired by the work of Johnstone and Lu (2009) that showed that the standard PCA gives consistent estimators of the loading vectors if and only if $d/n \to 0$. It also asserts that with an initial dimension reduction the standard PCA gives consistent solutions under a single component model that corresponds to the special case $M = 1$ of our model (2.1). The main difference between our model and theirs is in the 'sparsity' assumption. In Johnstone and Lu (2009), both $\mathbf{X}$ and $\boldsymbol{\beta}_1$ are represented in a fixed orthonormal basis, and the coefficients of the basis expansion for $\boldsymbol{\beta}_1$ decay rapidly, a 'sparsity' that differs from ours.

Performing a penalized PCA without dimension reduction does not guarantee a consistent result: this involves searching among many different vector components for those that give a significant 'spike' on the covariance of $\mathbf{X}$, allowing stochastic fluctuations to be mistaken for information when many variables are assessed. For example, the adaptive lasso of Zou (2006) needs initial consistent estimators of $\boldsymbol{\beta}_j$ to determine proper penalty weights, but this is possible only when $d/n \to 0$ as Johnstone and Lu (2009) and Paul (2007) demonstrate with the standard PCA. There are some penalized versions of PCA that give sparse solutions in our sense, but they do not appear to produce consistent results when

$d/n \to \infty$, since they put some type of penalization onto the standard PCA without dimension reduction. An interesting result on the standard PCA was obtained by Jung and Marron (2009) under the setting of fixed n and diverging d.

Sparsity of the loading vectors is closely related to sparsity of the covariance matrix $\mathbf{\Sigma}$. To estimate a high-dimensional sparse covariance matrix, Bickel and Levina (2008a,b) considered regularizing the sample covariance matrix by hard thresholding, banding, or tapering. Other related works include Johnstone (2001) and Paul (2007), in which the eigenstructure of the sample covariance matrix was studied. In regression settings, a number of efforts have been made to deal with the case where the number of predictors diverges. Some important developments include Fan and Peng (2004), Bair, Hastie, and Tibshirani (2006), Paul et al. (2008), Meinshausen and Yu (2009), Xie and Hunag (2009), and Zou and Zhang (2009).

The rest of this paper is organized as follows. In the next section, we describe the underlying sparse model, introduce the penalized PCA method with initial subset selection, and provide a numerical algorithm to calculate the sparse loading vectors. In Section 3, we show the consistency and the oracle property of the proposed method for a general penalty scheme. We report the results of a simulation study, and illustrate the proposed method through gene expression data. All the technical details are given in the Appendix.

2. Methodology

2.1. Underlying sparse models

Without loss of generality, assume $E(\mathbf{X}) = 0$. We consider the M-factor model

$$\mathbf{X}^i = \sum_{j=1}^{M} \lambda_j^{1/2} Z_j^i \boldsymbol{\beta}_j + \sigma \boldsymbol{\varepsilon}^i, \tag{2.1}$$

where Z_j^i are i.i.d. with mean 0 and variance 1, $\boldsymbol{\varepsilon}^i$ are i.i.d. having mean vector 0 and variance $\mathbf{I}_d$, independent of Z_j^i, and

$$\lambda_1 > \cdots > \lambda_M > 0.$$

Here and below, $\mathbf{I}_d$ represents the identity matrix of dimension d. The vectors $\boldsymbol{\beta}_j$ are orthonormal, $\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j = 1$ and $\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_k = 0$ for $j \neq k$. Thus, the strength of the signal is governed by the size of λ_j. From (2.1), the spectral decomposition of $\mathrm{var}(\mathbf{X})$ is

$$\mathrm{var}(\mathbf{X}) = \sum_{j=1}^{M} \lambda_j \boldsymbol{\beta}_j \boldsymbol{\beta}_j^\top + \sigma^2 \mathbf{I}_d.$$

Our theory is based on this spectral decomposition, rather than on (2.1).

In our treatment, the dimension d tends to infinity as n goes to infinity. Specifically, we assume $d = O(n^c)$ for some $c > 0$. In (2.1), we also allow M to grow to infinity as $n \to \infty$, and σ^2 may vary with n. One should note that λ_j and $\boldsymbol{\beta}_j$ depend on d, and thus on the sample size n. Under this model, the eigenvalues are $\lambda_j^0 = \lambda_j + \sigma^2$ for $1 \le j \le M$ and $\lambda_j^0 = \sigma^2$ for $M+1 \le j \le d$. The corresponding eigenvectors are $\boldsymbol{\beta}_j$. Note that separation of the eigenvalues λ_j^0 for $1 \le j \le M$ makes the corresponding eigenvectors $\boldsymbol{\beta}_j$, $1 \le j \le M$, identifiable. We also assume that the M loading vectors $\boldsymbol{\beta}_j$ in (2.1) are 'sparse' in the sense that

$$\beta_{j\ell} = 0 \quad \text{for all} \quad \ell \notin \mathcal{J}(j), \tag{2.2}$$

where $\text{card}[\mathcal{J}(j)] \le q$ and q is bounded. Let $\mathcal{I} = \bigcup_{j=1}^{M} \mathcal{J}(j), d_0 = \text{card}(\mathcal{I})$. Our treatment includes the case where $d_0 \to \infty$ as $n \to \infty$.

Suppose we are interested in estimating the loading vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ corresponding to the first $K(\le M)$ principal components, K fixed. Note that if $\ell \notin \mathcal{I}$, then $\beta_{j\ell} = 0$ for all $1 \le j \le M$. If we delete the variables that correspond to $\ell \notin \mathcal{I}$ and perform the principal component analysis with the remaining variables, then we get estimators of $\beta_{j\ell}$ for $1 \le j \le K$ and $\ell \in \mathcal{I}$. Thus, in the model (2.1) with (2.2), if we were to know $\mathcal{I}$, deleting X_ℓ^i for $\ell \notin \mathcal{I}$ is no harm for estimating $\boldsymbol{\beta}_j$, $1 \le j \le K$, since performing principal component analysis with the remaining X_ℓ^i with $\ell \in \mathcal{I}$ and setting those loading coefficients $\beta_{j\ell}$, $1 \le j \le K$, $\ell \notin \mathcal{I}$ to be zero would give better estimators of $\beta_{j\ell}$ than performing principal component analysis with the full data set.

Under (2.1),

$$\sigma_\ell^2 \equiv \text{var}(X_\ell) = \sum_{j=1}^{M} \lambda_j \beta_{j\ell}^2 + \sigma^2,$$

and under (2.2), for all $\ell \notin \mathcal{I}$,

$$\sigma_\ell^2 = \sigma^2 < \min\Big\{\sigma_k^2 : k \in \mathcal{I}\Big\}.$$

Thus, if we select d_0 indices according to the magnitudes of σ_ℓ^2, then the set of the selected indices is $\mathcal{I}$. Since the sample variances $\hat{\sigma}_\ell^2$ are good estimators of σ_ℓ^2, we expect the subset selection rule based on ranking of the sample variances to work well.

2.2. Penalized PCA

Here we describe our methods of estimating the sparse loading vectors of the first K principal components. To identify the sparsity, we adopt the idea of

variable selection in the linear regression setting where one adds a weighted L_1-penalty to the traditional least squares criterion. The latter works when d, the dimension of $\mathbf{X}$, is of moderate size. To handle the case where d is much larger than the sample size n, we first perform an initial dimension reduction. The subset of $\{X_\ell : 1 \le \ell \le d\}$ selected in this stage aims at $\{X_\ell : \ell \in \mathcal{I}\}$. Then, we do a penalized version of PCA based on the observations on the selected variables X_ℓ. This stage identifies and estimates the nonzero $\beta_{j\ell}$ for each $1 \le j \le K$.

Initial Dimension Reduction. Compute the sample variances $\hat{\sigma}_j^2 \equiv n^{-1}\sum_{i=1}^n X_j^{i2}$ for $1 \le j \le d$, and find

$$\hat{\mathcal{I}} \equiv \left\{\ell : \hat{\sigma}_\ell^2 \ge \left[1 + n^{-1/2}(\log n)^C\right]\hat{\sigma}^2\right\},$$

where C is any constant such that $C > 1/2$ and $\hat{\sigma}^2$ is a consistent estimate of σ^2.

Penalization. Let $\mu_{j\ell}$ for $1 \le j \le K$ and $\ell \in \hat{\mathcal{I}}$ be the penalty weights, allowed to be random or deterministic, given priori or chosen in some way. Let $\tilde{\mathbf{X}}^i = (X_j^i : j \in \hat{\mathcal{I}})^\top$. Maximize, successively for $j = 1, \ldots, K$,

$$\boldsymbol{\beta}_j^\top \left(n^{-1}\sum_{i=1}^n \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top}\right)\boldsymbol{\beta}_j - \sum_{\ell \in \hat{\mathcal{I}}} \mu_{j\ell} \cdot |\beta_{j\ell}| \tag{2.3}$$

subject to $\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j = 1$ and $\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_k = 0$ for $1 \le k < j$, to obtain $\hat{\beta}_{j\ell}$, $\ell \in \hat{\mathcal{I}}$.

Filling Up Loading Vectors. Set $\hat{\beta}_{j\ell} = 0$ for $1 \le j \le K$ and $\ell \notin \hat{\mathcal{I}}$.

In the initial dimension reduction step, one may use $\hat{\sigma}^2 = \text{median}(\hat{\sigma}_\ell^2)$. This is a consistent estimator of σ^2 if $d_0/d \to 0$, see the discussion after Theorem 1 in the next section. The proposed procedure does not depend on the knowledge of M, q, or d_0. The successive maximization of (2.3) for $1 \le j \le K$ is equivalent to maximizing

$$\sum_{j=1}^K \boldsymbol{\beta}_j^\top \left(n^{-1}\sum_{i=1}^n \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top}\right)\boldsymbol{\beta}_j - \sum_{j=1}^K \sum_{\ell \in \hat{\mathcal{I}}} \mu_{j\ell} \cdot |\beta_{j\ell}| \tag{2.4}$$

subject to $\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j = 1$ for $1 \le j \le K$ and $\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_k = 0$ for $1 \le j \ne k \le K$. Also, it can be shown that

$$n^{-1}\sum_{j=1}^K \boldsymbol{\beta}_j^\top \left(\sum_{i=1}^n \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top}\right)\boldsymbol{\beta}_j \;=\; n^{-1}\sum_{i=1}^n \tilde{\mathbf{X}}^{i\top}\tilde{\mathbf{X}}^i - n^{-1}\sum_{i=1}^n \|\tilde{\mathbf{X}}^i - \mathbf{B}\mathbf{B}^\top \tilde{\mathbf{X}}^i\|^2, \tag{2.5}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ is a $\hat{d}_0 \times K$ matrix. Here and below, $\hat{d}_0 = \text{card}(\hat{\mathcal{I}})$. Note that maximization of (2.5) is the usual eigen-analysis problem.

The theory we develop covers various choices of the penalty scheme $\mu_{j\ell}$. For example, it includes the adaptive lasso (Zou (2006)) where $\mu_{j\ell} = \tau_j |\tilde{\beta}_{j\ell}|^{-\gamma}$ for some $\gamma > 0$ and a regularization parameters $\tau_j > 0$. Here, $\tilde{\beta}_{j\ell}$ are some initial estimators of $\beta_{j\ell}$. Furthermore, it includes treatment of a general penalty function in which $\sum_{\ell \in \hat{\mathcal{I}}} \mu_{j\ell} \cdot |\beta_{j\ell}|$ is replaced by $\sum_{\ell \in \hat{\mathcal{I}}} p_{\tau_j}(|\beta_{j\ell}|)$ for some nonnegative, monotone increasing and differentiable functions p_{τ_j} with a regularization parameter τ_j. By linear approximation

$$p_{\tau_j}(|\beta_{j\ell}|) \simeq p_{\tau_j}(|\tilde{\beta}_{j\ell}|) + \left[\frac{\partial}{\partial \beta} p_{\tau_j}(\beta)\right]_{\beta=|\tilde{\beta}_{j\ell}|} (|\beta_{j\ell}| - |\tilde{\beta}_{j\ell}|),$$

and taking $\mu_{j\ell} = [\partial p_{\tau_j}(\beta)/\partial \beta]_{\beta=|\tilde{\beta}_{j\ell}|}$, reduces the general penalty scheme to (2.3). The general penalty scheme p_{τ_j} includes various penalization methods as special cases. For example, $p_{\tau_j}(x) = \tau_j\, x$ corresponds to the lasso, $p_{\tau_j}(x) = \tau_j^2\, p(x/\tau_j)$ with $p'(x) = I(x \le 1) + \frac{(\gamma - x)_+}{\gamma - 1} I(x > 1)$ for some $\gamma > 2$ to the SCAD penalty (Fan and Li (2001)), and $p_{\tau_j}(x) = \tau_j^2 \int_0^{x/\tau_j} (1 - u/\gamma)_+\, du$ for some $\gamma > 0$ to the MC penalty (Zhang (2010)). The one-step approximation of a non-convex penalty function p_{τ_j} is known to have various theoretical and computational advantages, see Zou and Li (2008) and Noh and Park (2010) for more details.

2.3. Numerical algorithm

Due to (2.5), maximization of (2.3) is equivalent to minimization of

$$n^{-1} \sum_{i=1}^{n} \|\tilde{\mathbf{X}}^i - \mathbf{B}\mathbf{B}^\top \tilde{\mathbf{X}}^i\|^2 + \sum_{j=1}^{K} \sum_{\ell \in \hat{\mathcal{I}}} \mu_{j\ell} \cdot |\beta_{j\ell}|.$$

A difficulty is that the optimization problem has the constraint $\mathbf{B}^\top \mathbf{B} = \mathbf{I}_K$. We suggest using an iterative algorithm that is a slight modification of the general SPCA algorithm proposed by Zou, Hastie, and Tibshirani (2006).

An iterative algorithm.

[S1] Take the $\hat{d}_0 \times K$ matrix $\mathbf{A}$ whose jth column vector is the normalized eigenvector corresponding to the jth largest eigenvalue of $n^{-1} \sum_{i=1}^{n} \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top}$.

[S2] Minimize

$$n^{-1} \sum_{i=1}^{n} \|\tilde{\mathbf{X}}^i - \mathbf{A}\mathbf{B}^\top \tilde{\mathbf{X}}^i\|^2 + \sum_{j=1}^{K} \sum_{\ell \in \hat{\mathcal{I}}} \mu_{j\ell} \cdot |\beta_{j\ell}|$$

with respect to $\mathbf{B}$.

[S3] Maximize

$$\operatorname{tr}[\mathbf{A}^\top(\tilde{\mathbb{X}}^\top\tilde{\mathbb{X}})\mathbf{B}] = \sum_{j=1}^{K} \boldsymbol{\alpha}_j^\top \left(\sum_{i=1}^{n} \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top}\right) \boldsymbol{\beta}_j$$

with respect to $\mathbf{A}$ subject to $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_K$.

[S4] Repeat [S2] and [S3] until $\mathbf{B}$ converges, and normalize $\mathbf{B}$ by $\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j/\|\boldsymbol{\beta}_j\|$.

3. Theoretical Properties

3.1. Oracle properties

In this section we give some theoretical properties of our sparse PCA method. Without loss of generality, assume $\mathcal{I} = \{1, \ldots, d_0\}$. The first theorem demonstrates that, with probability tending to one, the initial dimension reduction correctly identifies the index set $\mathcal{I}$ of the variables X_ℓ that make nonzero contribution to some of the first M principal components.

Theorem 1. *Under* (2.1) *and* (2.2), *suppose* $d = O(n^c)$ *for some* $c > 0$, $\sigma_\ell^2 \geq (1+\alpha_n)\sigma^2$ *for all* $\ell \in \mathcal{I}$, *where* $\alpha_n = n^{-1/2}(\log n)^\kappa$ *for some* $\kappa > C$ *and* C *is the constant in the definition of* $\hat{\mathcal{I}}$. *Assume that the* X_ℓ^i *have* m *moments for some* $m > 4(c+1)$, *and that* $P\big[\,|\hat{\sigma}^2 - \sigma^2| > n^{-1/2}(\log n)^{C_1}\sigma^2\big] \to 0$ *for some* C_1 *with* $1/2 < C_1 < C$. *Then,* $P(\hat{\mathcal{I}} = \mathcal{I}) \to 1$.

A proof of the theorem is given in the Appendix. There we prove

$$\sum_{\ell=d_0+1}^{d} P\Big(\hat{\sigma}_\ell^2 \geq \big[1 + c_1 n^{-1/2}(\log n)^C\big]\sigma^2\Big) \to 0, \tag{3.1}$$

$$\sum_{\ell=1}^{d_0} P\Big(\hat{\sigma}_\ell^2 < \big[1 + c_2 n^{-1/2}(\log n)^C\big]\sigma^2\Big) \to 0 \tag{3.2}$$

for all $c_1, c_2 > 0$. These two properties imply

$$P\left(\sup_{d_0+1\leq\ell\leq d} \hat{\sigma}_\ell^2 < \inf_{1\leq\ell\leq d_0} \hat{\sigma}_\ell^2\right) \to 1. \tag{3.3}$$

Also, one can prove similarly as in the proofs of (3.1) and (3.2) that

$$\sum_{\ell=d_0+1}^{d} P\Big(|\hat{\sigma}_\ell^2 - \sigma^2| \geq n^{-1/2}(\log n)^{C_1}\sigma^2\Big) \;\to\; 0, \tag{3.4}$$

Let $\hat{\sigma}^2 = \operatorname{median}\big(\hat{\sigma}_\ell^2\big)$. Then, from (3.3) and (3.4) the estimator $\hat{\sigma}^2$ satisfies the condition of Theorem 1: $P\big[|\hat{\sigma}^2 - \sigma^2| > n^{-1/2}(\log n)^{C_1}\sigma^2\big] \to 0$, provided $d_0/d \to 0$.

We note that the standard PCA fails under the condition of Theorem 1. To see this, let $\sigma^2 = 1$ and $\boldsymbol{\beta}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ for $1 \le j \le M$, with 1 appearing at the jth position for simplicity. Then, the covariance structure of $\mathbf{X}^i$ in our model exactly matches the one considered in Paul (2007), and $\sigma_j^2 = \lambda_j + 1$. In this case $\mathcal{I} = \{1, \ldots, M\}$. In Paul (2007), the case where $\sigma_j^2 \le 1 + \sqrt{\gamma}$ with $\gamma > 0$ being the limit of d/n corresponds to the "more difficult phase" where the standard PCA fails. In our theorems we allow the case $\sigma_j^2 = 1 + n^{-1/2}(\log n)^\kappa$ for $j \in \mathcal{I}$. Thus, all $\boldsymbol{\beta}_j$ for $1 \le j \le M$ belong to this more difficult phase. Johnstone and Lu (2009) also showed that the standard PCA fails when $d/n \to \gamma > 0$.

Next, we present the oracle properties of the penalization method. These properties are that, for each $1 \le j \le K$ and with probability tending to one, our method selects the nonzero $\beta_{j\ell}$ correctly, and that our estimators $\hat{\boldsymbol{\beta}}_j$ have the same first-order properties as those obtained from an oracle PCA that uses knowledge of the true index sets $\mathcal{J}(j)$ for $1 \le j \le K$.

Note that $d_0 \le Mq$ under (2.2). Let $\hat{\mathcal{J}}(j) = \{\ell \in \hat{\mathcal{I}} : \hat{\beta}_{j\ell} \ne 0\}$ and $\tilde{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^n \mathbf{X}^i \mathbf{X}^{i\top}$. We assume

$$d_0 \sim n^{C'} \tag{3.5}$$

for some $0 \le C' < 1/4$. We let $C' = 0$ when d_0 is a bounded number. For the penalty constants we focus on the case where they are random, and assume that for each $1 \le j \le K$,

$$n^{1/2} \max_{\ell \in \mathcal{J}(j)} \mu_{j\ell} \to 0, \quad n^{1/2-2C'} \inf_{\ell \notin \mathcal{J}(j)} \mu_{j\ell} \to \infty \tag{3.6}$$

in probability for any $0 < c < \infty$. The theory we present is also valid for the deterministic $\mu_{j\ell}$ if we assume (3.6).

Theorem 2. *Under the conditions in Theorem* 1, (3.5), *and* (3.6), *suppose that the X_ℓ^i have m moments for some $m > 4(2C'+1)$, that K is fixed and* $\operatorname{card}(\mathcal{J}(j))$ *for $1 \le j \le K$ are bounded numbers, that there exists a constant $\delta > 0$ such that $\inf_{\ell \in \mathcal{J}(j)} |\beta_{j\ell}| \ge \delta$ for all $1 \le j \le K$, and that $\limsup_{n\to\infty} \sup_{1 \le \ell \le d_0} \sigma_\ell^2 < \infty$. Then,* (i) *with probability tending to one as $n \to \infty$, the method correctly selects those components $\beta_{j\ell} \ne 0$ in the sense that $P[\hat{\mathcal{J}}(j) = \mathcal{J}(j)] \to 1$ for all $1 \le j \le K$, and* (ii) *the estimator $\big(\hat{\beta}_{j\ell} : \ell \in \mathcal{J}(j),\ 1 \le j \le K\big)^\top$ has the same first-order asymptotic properties that the PCA with the constraints $\beta_{j\ell} = 0$ for $\ell \notin \mathcal{J}(j)$, $1 \le j \le K$, would enjoy.*

One may extend the results of the above theorem to the case where K is not fixed but tends to infinity as the sample size n grows. This requires stronger conditions on the penalty weights $\mu_{j\ell}$ than those given at (3.6) and conditions on K.

The constrained PCA in the above theorem is an oracle procedure that is based on the knowledge of $\beta_{j\ell} = 0$ for $\ell \notin \mathcal{J}(j)$, $1 \le j \le K$. Let $\hat{\boldsymbol{\beta}}_j^{\mathrm{ora}}$ denote the resulting oracle estimators of $\boldsymbol{\beta}_j$. The constrained PCA is performed as follows. First, set $\hat{\beta}_{1\ell}^{\mathrm{ora}} = 0$ for $\ell \notin \mathcal{J}(1)$, and maximize $\sum_{\ell\in\mathcal{J}(1)}\sum_{\ell'\in\mathcal{J}(1)} \tilde{\Sigma}_{\ell\ell'}\beta_{1\ell}\beta_{1\ell'}$ subject to $\sum_{\ell\in\mathcal{J}(1)}\beta_{1\ell}^2 = 1$ to get $\hat{\beta}_{1\ell}^{\mathrm{ora}}$ for $\ell \in \mathcal{J}(1)$, where $\tilde{\Sigma}_{\ell\ell'}$ is the (ℓ,ℓ')th entry of $\tilde{\boldsymbol{\Sigma}}$. Next, set $\hat{\beta}_{2\ell}^{\mathrm{ora}} = 0$ for $\ell \notin \mathcal{J}(2)$, and maximize $\sum_{\ell\in\mathcal{J}(2)}\sum_{\ell'\in\mathcal{J}(2)} \tilde{\Sigma}_{\ell\ell'}\beta_{2\ell}\beta_{2\ell'}$ subject to $\sum_{\ell\in\mathcal{J}(2)}\beta_{2\ell}^2 = 1$ and $\sum_{\ell\in\mathcal{J}(1)\cap\mathcal{J}(2)}\hat{\beta}_{1\ell}^{\mathrm{ora}}\beta_{2\ell} = 0$ to get $\hat{\beta}_{2\ell}^{\mathrm{ora}}$ for $\ell \in \mathcal{J}(2)$. Continue the procedure to obtain $\hat{\beta}_j^{\mathrm{ora}}$ for $j = 3$, and so on until $j = K$.

The success of the procedure hinges on the success of the initial dimension reduction. In fact, what we need asymptotically in the dimension reduction is

$$P\Big(\bigcup_{j=1}^{K}\mathcal{J}(j) \subset \hat{\mathcal{I}}\Big) \to 1. \tag{3.7}$$

This means that we do not need to recover the set $\mathcal{I} = \bigcup_{j=1}^{M}\mathcal{J}(j)$ but only $\mathcal{J}(j)$ for $j \le K$, where K is the number of loading vectors that we want to estimate. Note that the conclusion of Theorem 1 implies (3.7). If $\hat{\mathcal{I}}$ misses some indices in $\bigcup_{j=1}^{K}\mathcal{J}(j)$, the estimators $\hat{\boldsymbol{\beta}}_j$ $(1 \le j \le K)$ target vectors different from the true $\boldsymbol{\beta}_j$. For example, suppose that $M = K = 2$, $\lambda_1 = 3$, $\lambda_2 = 2$, $\sigma^2 = 1$, $\boldsymbol{\beta}_1 = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, \ldots, 0)$, and $\boldsymbol{\beta}_2 = (1/\sqrt{2}, -1/\sqrt{2}, 0, \ldots, 0)$, and that one chooses $\hat{\mathcal{I}} = \{2, 3\}$ missing X_1. The resulting estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ aim at $\boldsymbol{\beta}_1^* = (0, v_{12}, v_{13}, 0, \ldots, 0)$ and $\boldsymbol{\beta}_2^* = (0, v_{22}, v_{23}, 0, \ldots, 0)$, where (v_{12}, v_{13}) and (v_{22}, v_{23}) are the normalized eigenvectors of the covariance matrix of (X_2, X_3).

3.2. Penalty weights

In this section we consider the one-step approximation of the general penalty function p_{τ_j} discussed in Section 2.2. We discuss how the conditions on the penalty weights $\mu_{j\ell}$ at (3.6) are satisfied for $\mu_{j\ell} = [\partial p_{\tau_j}(\beta)/\partial\beta]_{\beta=|\tilde{\beta}_{j\ell}|}$, where $\tilde{\beta}_{j\ell}$ are some initial estimators of $\beta_{j\ell}$. We consider two forms of penalty function p_{τ_j}: $p_{\tau_j} = \tau_j^2 p(\cdot/\tau_j)$ and $p_{\tau_j} = \tau_j p$, for a nonnegative, monotone increasing and differentiable function p. The former was studied by Noh (2009) and Zhang (2010), and the latter by Lv and Fan (2009) in the linear regression problem. The SCAD and MC penalization methods take the first form, with $p'(x) = I(x \le 1) + \frac{(\gamma - x)_+}{\gamma - 1} I(x > 1)$ for the SCAD and $p(x) = \int_0^x (1 - u/\gamma)_+ \, du$ for the MC. The adaptive lasso corresponds to the second form with $p'(u) = u^{-\gamma}$ for some $\gamma > 0$. We derive a set of sufficient conditions on the function p and the regularization parameters τ_j, in each form of p_{τ_j}, for the the penalty weights $\mu_{j\ell}$ to satisfy (3.6).

The initial estimators $\tilde{\boldsymbol{\beta}}_j$, $1 \le j \le K$, we take are those $\hat{d}_0$-dimensional orthonormal eigenvectors obtained from performing PCA on $\tilde{\mathbf{X}}^i = (X_j^i : j \in \hat{\mathcal{I}})^\top$.

The initial estimators need to have a certain rate of convergence. For this we assume

$$\lambda_j - \lambda_{j+1} \ \geq \ (\log n)^{-C''}, \ 1 \leq j \leq K \tag{3.8}$$

for some $C'' > 0$. This condition requires that the spacings between the leading eigenvalues are not too small. This is automatically satisfied when the λ_j are fixed and do not depend on n.

Theorem 3. *Under* (2.1), (2.2), (3.5), *and* (3.8), *suppose that the* X_ℓ^i *have* m *moments for some* $m > 4(2C'+1)$, *and that* $\limsup_{n\to\infty} \sup_{1\leq \ell \leq d_0} \sigma_\ell^2 < \infty$, *where* C' *is the constant at (3.5). Let* $\bar{\boldsymbol{\beta}}_j$, $1 \leq j \leq K$ *be the loading vectors for the first* K *principal component of* $\tilde{\boldsymbol{\Sigma}}_0$, *the* $d_0 \times d_0$ *top-left block of* $\tilde{\boldsymbol{\Sigma}}$. *Then it follows that*

$$\sup_{1\leq \ell \leq d_0} |\bar{\beta}_{j\ell} - \beta_{j\ell}| = O_p\left(n^{-1/2+C'}(\log n)^{1/2+C''}\right), \ 1 \leq j \leq K,$$

where C'' *is the constant at* (3.8).

3.2.1. The case $p_\tau = \tau^2 p(\cdot/\tau)$.

In this case, $\mu_{j\ell} = \tau_j p'(|\tilde{\beta}_{j\ell}|/\tau_j)$. Suppose that p has a nonnegative and nonincreasing derivative p' on $(0,\infty)$, and that

$$\lim_{u\to 0+} p'(u) > 0, \quad p'(u) = O(u^{-a}) \ \text{ as } \ u \to \infty \ \text{ for some } \ a > 0. \tag{3.9}$$

We verify that the conditions at (3.6) are satisfied if, for each $1 \leq j \leq K$,

$$n^{1/2}\tau_j^{1+a} \to 0, \ \ n^{1/2-2C'}\tau_j \to \infty, \tag{3.10}$$

where C' is the constant at (3.5). The first part of (3.6) follows easily from the second condition of (3.9) and the first condition of (3.10). To see that the second part of (3.6) holds, we note that from the monotonicity of p',

$$\inf_{\ell\notin\mathcal{J}(j)} p'(|\tilde{\beta}_{j\ell}|/\tau_j) \geq p'\left(\frac{n^{1/2-C'}(\log n)^{-(1/2+C'')} \sup_{1\leq\ell\leq d_0}|\tilde{\beta}_{j\ell} - \beta_{j\ell}|}{\tau_j\, n^{1/2-C'}(\log n)^{-(1/2+C'')}}\right). \tag{3.11}$$

By Theorem 3, the first condition of (3.9), and the second condition of (3.10), the right hand side of (3.11) converges in probability to a strictly positive constant. This implies that

$$n^{1/2-2C'} \inf_{\ell\notin\mathcal{J}(j)} \mu_{j\ell} \ = \ n^{1/2-2C'}\tau_j \inf_{\ell\notin\mathcal{J}(j)} p'(|\tilde{\beta}_{j\ell}|/\tau_j) \ \xrightarrow{p} \ \infty.$$

Both the one-step SCAD and MC penalty functions satisfy (3.9) for *all* constants $a > 0$ since $p'(u)$ in those cases vanishes for all u greater than a fixed

positive constant. Thus, for these methods, the first condition of (3.10) only needs to hold for an *arbitrarily large* constant $a > 0$. If $\tau_j = O(n^{-b})$ for some $b > 0$, then the first condition of (3.10) always hold by taking $a > 0$ sufficiently large. Thus, for the one-step SCAD and MC penalty functions, one only needs the second condition of (3.10).

3.2.2. The case $p_\tau = \tau p$.

Here, $\mu_{j\ell} = \tau_j p'(|\tilde{\beta}_{j\ell}|)$. Suppose that p has a nonnegative and nonincreasing derivative p' on $(0, \infty)$, and that

$$p'(u)^{-1} = O(u^\gamma) \text{ as } u \to 0 \text{ for some } \gamma > 0. \tag{3.12}$$

The condition (3.12) implies that $p'(u)$ tends to infinity as u decreases to zero, and this makes sense with the weight scheme $\mu_{j\ell} = \tau_j p'(|\tilde{\beta}_{j\ell}|)$ since one needs to put a large penalty for β_j close to zero. The conditions at (3.6) are satisfied if

$$n^{1/2}\tau_j \to 0, \quad n^{\gamma(1/2-C')+(1/2-2C')}\tau_j \, (\log n)^{-\gamma(1/2+C'')} \xrightarrow{p} \infty, \tag{3.13}$$

where C' and C'' are the constants at (3.5) and (3.8), respectively. The first condition of (3.6) is immediate from the first condition of (3.13). To check the second condition of (3.6), note that

$$\inf_{\ell \notin \mathcal{J}(j)} p'(|\tilde{\beta}_{j\ell}|) \geq p'\left(\frac{n^{1/2-C'}(\log n)^{-(1/2+C'')} \sup_{1 \leq \ell \leq d_0} |\tilde{\beta}_{j\ell} - \beta_{j\ell}|}{n^{1/2-C'}(\log n)^{-(1/2+C'')}} \right).$$

By Theorem 3 and (3.12), the inverse of the right hand side of this inequality is bounded, in probability, by $n^{-\gamma(1/2-C')}(\log n)^{\gamma(1/2+C'')}$ multiplied by some strictly positive constant. The second condition of (3.6) now follows from the second condition of (3.13).

Recall that the adaptive lasso corresponds to the one-step penalized method with $p_\tau = \tau p$ and $p'(u) = u^{-\gamma}$ for some $\gamma > 0$. This means that the penalty weights $\mu_{j\ell} = \tau_j |\tilde{\beta}_{j\ell}|^{-\gamma}$ of the adaptive lasso satisfy (3.6) if (3.13) holds.

4. Numerical Properties

We investigated the finite sample performance of the proposed methods. The penalization methods we took were the adaptive lasso method, SCAD, and MC. Our aims were to see how effectively the proposed methods identify the sparse loadings, and to compare the mean squared errors of $\hat{\boldsymbol{\beta}}_j$ for the three penalization methods. For the constant γ of the three penalization methods, we chose $\gamma = 3.7$ for the SCAD penalty, as suggested by Fan and Li (2001), and for the MC penalty we used $\gamma = 2/(1 - \max_{j \neq k} |\mathbf{x}_j^\top \mathbf{x}_k|/n)$, where $\mathbf{x}_j$ denotes the

jth column of the design matrix with $\mathbf{x}_j^\top \mathbf{x}_j = n$, which is the minimal value that affords the theoretical results in Zhang (2010). For the adaptive lasso penalty we took $\gamma = 0.5$, 1, 2.

In the comparison we also considered the SPCA algorithm of Zou, Hastie, and Tibshirani (2006). The latter does not have an initial dimension reduction stage, and in the case where $d \gg n$, its algorithm runs as does ours but with $\tilde{\mathbf{X}}^i$ and $\hat{d}_0$ replaced by $\mathbf{X}^i$ and d, respectively, and in [S2] it minimizes

$$\sum_{j=1}^{K} \left[-2\,\boldsymbol{\beta}_j^\top \left(n^{-1} \sum_{i=1}^{n} \mathbf{X}^i \mathbf{X}^{i\top} \right) \boldsymbol{\alpha}_j + \boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j + \tau_j \sum_{\ell=1}^{d} |\beta_{j\ell}| \right]. \tag{4.1}$$

The minimization problem has the explicit solution

$$\hat{\boldsymbol{\beta}}_j = \left(\left| n^{-1} \sum_{i=1}^{n} \mathbf{X}^i \mathbf{X}^{i\top} \boldsymbol{\alpha}_j \right| - \frac{\tau_j}{2} \right)_+ \operatorname{sign} \left(\sum_{i=1}^{n} \mathbf{X}^i \mathbf{X}^{i\top} \boldsymbol{\alpha}_j \right).$$

We also added a hard-thresholding rule to the comparison. The method performs the same initial dimension reduction step as the proposed methods. After the initial dimension reduction, it carries out PCA on $\tilde{\mathbf{X}}^i$ to get $\tilde{\boldsymbol{\beta}}_j$, and then takes $\hat{\beta}_{j\ell} = \tilde{\beta}_{j\ell} I(|\tilde{\beta}_{j\ell}| > M_j n^{-2/5})$ for $1 \le j \le K$ and $\ell \in \hat{\mathcal{I}}$, where $M_j > 0$ is a tuning parameter to be chosen. A similar idea was proposed by Johnstone and Lu (2009) in the case of the single factor model with $M = 1$.

The regularization parameters τ_j for our methods were selected by a BIC-type criterion. The constant C in the initial dimension reduction stage was selected by the same criterion. For a given τ_j and a cut-off constant C, let $\hat{\boldsymbol{\alpha}}_{j,C,\tau_j}$ and $\hat{\boldsymbol{\beta}}_{j,C,\tau_j}$ be the limit of the iterative algorithm described in Section 2.3. We minimized

$$\begin{aligned} \mathrm{BIC}(C; \tau_1, \ldots, \tau_K) = {} & \sum_{j=1}^{K} (\hat{\boldsymbol{\beta}}_{j,C,\tau_j} - \hat{\boldsymbol{\alpha}}_{j,C,\tau_j})^\top \left(n^{-1} \sum_{i=1}^{n} \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top} \right) (\hat{\boldsymbol{\beta}}_{j,C,\tau_j} - \hat{\boldsymbol{\alpha}}_{j,C,\tau_j}) \\ & + \sum_{j=1}^{K} \mathrm{df}_{C,\tau_j} \frac{\log n}{n} \end{aligned}$$

to select C and τ_j, where df_{C,τ_j} is the number of nonzero loading coefficients identified in $\hat{\boldsymbol{\beta}}_{j,C,\tau_j}$. The BIC-type criterion was also used in Leng and Wang (2009). In the case of the Zou et al.'s SPCA, we used a different BIC-type criterion to select τ_j. Observing that the objective function at (4.1) is equivalent to $\sum_{j=1}^{K} \|\boldsymbol{\beta}_j - \tilde{\boldsymbol{\Sigma}} \boldsymbol{\alpha}_j\|^2 + \tau_j \sum_{\ell=1}^{d} |\beta_{j\ell}|$, we minimized

$$\mathrm{BIC}(\tau_j) = \left\| \hat{\boldsymbol{\beta}}_{j,\tau_j} - n^{-1} \sum_{i=1}^{n} \mathbf{X}^i \mathbf{X}^{i\top} \hat{\boldsymbol{\alpha}}_{j,\tau_j} \right\|^2 + \mathrm{df}_{\tau_j} \frac{\log n}{n}.$$

As for the hard-thresholding method, we minimized a BIC-type criterion to select the tuning parameters M_j and the constant C in the initial dimension reduction:

$$\mathrm{BIC}(C; M_1, \ldots, M_K) = \sum_{j=1}^{K} \left\| \hat{\boldsymbol{\beta}}_{j,C,M_j} - \tilde{\boldsymbol{\beta}}_{j,C} \right\|^2 + \sum_{j=1}^{K} \mathrm{df}_{C,M_j} \frac{\log n}{n},$$

where the definitions of $\hat{\boldsymbol{\beta}}_{j,C,M_j}$, $\tilde{\boldsymbol{\beta}}_{j,C}$ and df_{C,M_j} are obvious.

We took $d = 1,000$ for the dimension of the variables $\mathbf{X}^i$, and $n = 100$, 400 for the sample size. We generated $\mathbf{X}^i$ according to (2.1). We chose $M = 2$, $\lambda_1 = 2$, $\lambda_2 = 1$, and

(1) $\boldsymbol{\beta}_1 = \frac{1}{\sqrt{5}}(1_5, 0_{995})^\top$, $\boldsymbol{\beta}_2 = \frac{1}{\sqrt{5}}(0_5, 1_5, 0_{990})^\top$,

(2) $\boldsymbol{\beta}_1 = \frac{1}{\sqrt{5}}(1_5, 0_{995})^\top$, $\boldsymbol{\beta}_2 = \frac{1}{2}(1, -1, 1, -1, 0_{996})^\top$,

where a_k denotes k-dimensional row vector with all elements being a. For each model, we considered the noise levels $\sigma = 0.5$ and $\sigma = 1$. We generated ε_i from the d-dimensional standard normal distribution $N(\mathbf{0}, \mathbf{I}_d)$.

The results, based on 100 Monte Carlo samples, are reported in Tables 1–4 as the average numbers of correctly and incorrectly identified zero loadings. The correct zeros are when $\hat{\beta}_{j\ell} = 0 = \beta_{j\ell}$, and the incorrect zeros are those with $\hat{\beta}_{j\ell} = 0$ but $\beta_{j\ell} \neq 0$. Thus, for model (1) it is better to have the number of correct zeros closer to 995 for both $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, and for model (2) to 995 for $\hat{\boldsymbol{\beta}}_1$ and to 996 for $\hat{\boldsymbol{\beta}}_2$. The tables also report the Monte Carlo mean of the squared errors $\sum_{\ell=1}^{d} (\hat{\beta}_{j\ell} - \beta_{j\ell})^2$.

Compared with the Zou et al.'s SPCA (SPCA-ZHT), the tables suggest that our methods with the adaptive lasso, the SCAD, and the MC penalty schemes have better performance in general. In terms of identifying zero loadings, the SPCA-ZHT largely fails except in the case where $n = 400$ and $\sigma = 0.5$. In particular, the method identifies only 6% of the true zero loadings in the case of smaller n and larger σ. Our methods performed well in all cases. Our methods also defeated the SPCA-ZHT in terms of the squared error performance. The superiority of the proposed methods are evident in the case of the smaller sample size. When n is large, our methods were far better than the SPCA-ZHT for the higher noise level, and were comparable to the latter for the lower noise. In terms of incorrect zero performance, the SPCA-ZHT was better than our method in the case of smaller n and larger σ. This is mainly due to the fact that the SPCA-ZHT produces non-sparse solutions in general, and especially in the case of smaller n and larger σ. In comparison of the three penalization methods with the hard-thresholding rule (HARD-THRES), we found that the HARD-THRES was slightly better for the lower noise level, but worse for the higher noise level.

Table 1. Performance of the Methods for Model 1 ($n = 100$).

				Avg. # Zero Loadings	
		Method	Squared Error	Correct	Incorrect
$\sigma = 0.5$	$\boldsymbol{\beta}_1$	SPCA-ZHT	0.2518	690.76	0.00
		HARD-THRES	0.0138	994.88	0.00
		A.LASSO ($\gamma = 0.5$)	0.0249	995.00	0.02
		A.LASSO ($\gamma = 1$)	0.0251	995.00	0.04
		A.LASSO ($\gamma = 2$)	0.0306	995.00	0.07
		SCAD	0.0248	995.00	0.04
		MC	0.0278	995.00	0.01
	$\boldsymbol{\beta}_2$	SPCA-ZHT	0.9578	682.46	0.32
		HARD-THRES	0.0581	994.83	0.15
		A.LASSO ($\gamma = 0.5$)	0.0946	995.00	0.22
		A.LASSO ($\gamma = 1$)	0.1091	995.00	0.31
		A.LASSO ($\gamma = 2$)	0.1525	995.00	0.52
		SCAD	0.1264	995.00	0.36
		MC	0.1162	995.00	0.25
$\sigma = 1$	$\boldsymbol{\beta}_1$	SPCA-ZHT	1.8061	62.08	0.11
		HARD-THRES	0.7322	992.40	2.02
		A.LASSO ($\gamma = 0.5$)	0.7007	991.24	2.00
		A.LASSO ($\gamma = 1$)	0.6674	993.21	2.06
		A.LASSO ($\gamma = 2$)	0.6354	994.12	2.16
		SCAD	0.6679	992.90	2.14
		MC	0.6430	993.26	2.13
	$\boldsymbol{\beta}_2$	SPCA-ZHT	1.9144	66.56	0.25
		HARD-THRES	1.7956	988.75	4.39
		A.LASSO ($\gamma = 0.5$)	1.2986	991.22	4.32
		A.LASSO ($\gamma = 1$)	1.1760	993.23	4.52
		A.LASSO ($\gamma = 2$)	1.0665	994.30	4.69
		SCAD	1.1649	993.20	4.53
		MC	1.1459	993.47	4.57

The three penalization schemes showed similar performance in general, although different choices of γ within the adaptive lasso scheme gave somewhat different squared error performance for model (2) with the lower noise level.

5. Real Data Analysis

The Golub data has $d = 7,129$ genes and $n = 72$ tumor samples. The data have been analyzed by Golub et al. (1999) to classify two cancer classes: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). We applied the proposed sparse PCA method to find the two leading sparse principal component (PC) scores ($K = 2$). We used the adaptive lasso penalty with $\gamma = 1$. The constant C in the initial dimension reduction and the regularization

Table 2. Performance of the Methods for Model 1 ($n = 400$).

				Avg. # Zero Loadings	
		Method	Squared Error	Correct	Incorrect
$\sigma = 0.5$	$\boldsymbol{\beta}_1$	SPCA-ZHT	0.0028	987.01	0.00
		HARD-THRES	0.0015	995.00	0.00
		A.LASSO ($\gamma = 0.5$)	0.0028	995.00	0.00
		A.LASSO ($\gamma = 1$)	0.0023	995.00	0.00
		A.LASSO ($\gamma = 2$)	0.0031	995.00	0.00
		SCAD	0.0036	995.00	0.00
		MC	0.0034	995.00	0.00
	$\boldsymbol{\beta}_2$	SPCA-ZHT	0.0052	993.27	0.00
		HARD-THRES	0.0037	994.96	0.00
		A.LASSO ($\gamma = 0.5$)	0.0080	994.99	0.00
		A.LASSO ($\gamma = 1$)	0.0074	995.00	0.00
		A.LASSO ($\gamma = 2$)	0.0079	995.00	0.00
		SCAD	0.0124	995.00	0.00
		MC	0.0099	995.00	0.00
$\sigma = 1$	$\boldsymbol{\beta}_1$	SPCA-ZHT	1.1635	167.09	0.06
		HARD-THRES	0.0196	994.79	0.03
		A.LASSO ($\gamma = 0.5$)	0.0209	994.92	0.03
		A.LASSO ($\gamma = 1$)	0.0187	995.00	0.03
		A.LASSO ($\gamma = 2$)	0.0171	995.00	0.03
		SCAD	0.0212	994.92	0.03
		MC	0.0220	994.93	0.03
	$\boldsymbol{\beta}_2$	SPCA-ZHT	1.8342	156.18	0.52
		HARD-THRES	0.6764	989.85	1.87
		A.LASSO ($\gamma = 0.5$)	0.5308	993.01	1.81
		A.LASSO ($\gamma = 1$)	0.5050	994.04	1.85
		A.LASSO ($\gamma = 2$)	0.4923	994.44	1.89
		SCAD	0.5050	993.63	1.85
		MC	0.5147	993.46	1.84

parameters τ_j were selected by the BIC-type criterion introduced in Section 4. As few as 2.3% of the 7,129 genes went into the first principal component, and 3.0% the second. Figure 1 depicts the two PC scores of 72 tumor samples marked according to their cancer classes by black (AML) and white (ALL) circles. The figure shows that the two leading sparse PC scores effectively classify the cancer classes.

Acknowledgements

We thank an associate editor and two referees for their helpful comments on the earlier version of the paper. The work of Y. K. Lee was supported by Basic Science Research Program through the National Research Foundation of Korea

Table 3. Performance of the Methods for Model 2 ($n = 100$).

				Avg. # Zero Loadings	
		Method	Squared Error	Correct	Incorrect
$\sigma = 0.5$	$\boldsymbol{\beta}_1$	SPCA-ZHT	0.2732	691.53	0.00
		HARD-THRES	0.0415	995.00	0.07
		A.LASSO ($\gamma = 0.5$)	0.0662	995.00	0.15
		A.LASSO ($\gamma = 1$)	0.1850	995.00	0.67
		A.LASSO ($\gamma = 2$)	0.4667	995.00	1.80
		SCAD	0.2082	995.00	0.78
		MC	0.1138	995.00	0.37
	$\boldsymbol{\beta}_2$	SPCA-ZHT	0.8735	687.99	0.32
		HARD-THRES	0.0401	995.93	0.03
		A.LASSO ($\gamma = 0.5$)	0.0782	996.00	0.12
		A.LASSO ($\gamma = 1$)	0.1724	996.00	0.49
		A.LASSO ($\gamma = 2$)	0.4186	996.00	1.30
		SCAD	0.1885	996.00	0.54
		MC	0.1171	996.00	0.26
$\sigma = 1$	$\boldsymbol{\beta}_1$	SPCA-ZHT	1.8036	64.32	0.13
		HARD-THRES	0.4429	993.81	1.27
		A.LASSO ($\gamma = 0.5$)	0.4646	992.68	1.17
		A.LASSO ($\gamma = 1$)	0.5077	994.09	1.64
		A.LASSO ($\gamma = 2$)	0.5601	994.70	2.00
		SCAD	0.5197	993.93	1.69
		MC	0.5509	994.13	1.86
	$\boldsymbol{\beta}_2$	SPCA-ZHT	1.8995	63.61	0.07
		HARD-THRES	1.1917	991.70	2.17
		A.LASSO ($\gamma = 0.5$)	1.0431	990.91	1.74
		A.LASSO ($\gamma = 1$)	0.9804	994.02	2.42
		A.LASSO ($\gamma = 2$)	0.9320	995.42	2.90
		SCAD	0.9773	993.92	2.41
		MC	0.9991	994.37	2.60

(NRF) funded by the Ministry of Education, Science and Technology (2010-0021396). The work of B. U. Park and E. R. Lee was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2008-314-C00046).

Appendix

A.1. Proof of Theorem 1

Without loss of generality, assume $\sigma_1^2 \geq \cdots \geq \sigma_d^2$. Note that with this re-enumeration, $\sigma_{d_0+1}^2 = \cdots = \sigma_d^2 = \sigma^2$. We prove (3.1) and (3.2) for all $c_1, c_2 > 0$. Because of the condition on the estimator $\hat{\sigma}^2$, these two properties imply $\sum_{\ell=d_0+1}^{d} P(\ell \in \hat{\mathcal{I}}) \to 0$ and $\sum_{\ell=1}^{d_0} P(\ell \notin \hat{\mathcal{I}}) \to 0$, respectively, concluding the

Table 4. Performance of the Methods for Model 2 ($n = 400$).

				Avg. # Zero Loadings	
		Method	Squared Error	Correct	Incorrect
$\sigma = 0.5$	$\boldsymbol{\beta}_1$	SPCA-ZHT	0.0125	987.68	0.00
		HARD-THRES	0.0083	995.00	0.00
		A.LASSO ($\gamma = 0.5$)	0.0087	995.00	0.00
		A.LASSO ($\gamma = 1$)	0.0090	995.00	0.00
		A.LASSO ($\gamma = 2$)	0.0220	995.00	0.05
		SCAD	0.0088	995.00	0.00
		MC	0.0087	995.00	0.00
	$\boldsymbol{\beta}_2$	SPCA-ZHT	0.0154	994.18	0.00
		HARD-THRES	0.0075	995.99	0.00
		A.LASSO ($\gamma = 0.5$)	0.0109	996.00	0.00
		A.LASSO ($\gamma = 1$)	0.0107	996.00	0.00
		A.LASSO ($\gamma = 2$)	0.0220	996.00	0.04
		SCAD	0.0116	996.00	0.00
		MC	0.0105	996.00	0.00
$\sigma = 1$	$\boldsymbol{\beta}_1$	SPCA-ZHT	1.1594	171.65	0.04
		HARD-THRES	0.0269	994.84	0.01
		A.LASSO ($\gamma = 0.5$)	0.0336	994.94	0.01
		A.LASSO ($\gamma = 1$)	0.0343	994.99	0.03
		A.LASSO ($\gamma = 2$)	0.0446	995.00	0.09
		SCAD	0.0357	994.98	0.03
		MC	0.0388	994.99	0.03
	$\boldsymbol{\beta}_2$	SPCA-ZHT	1.8506	160.27	0.22
		HARD-THRES	0.1027	993.62	0.00
		A.LASSO ($\gamma = 0.5$)	0.0633	995.13	0.01
		A.LASSO ($\gamma = 1$)	0.0629	995.54	0.04
		A.LASSO ($\gamma = 2$)	0.0714	995.76	0.10
		SCAD	0.0652	995.42	0.03
		MC	0.0670	995.28	0.03

proof.

We use Bernstein's inequality for a sum of independent random variables W_i:

$$P\left(\left|\sum_{i=1}^{n} W^i\right| > x\right) \leq 2\,\exp\left(-\frac{1}{2}\frac{x^2}{V + Lx/3}\right), \tag{A.1}$$

where V is the upper bound for the variance of $\sum_{i=1}^{n} W^i$ and L is the bound for the absolute values of W^i, i.e., $|W^i| \leq L$. To apply the inequality we use the

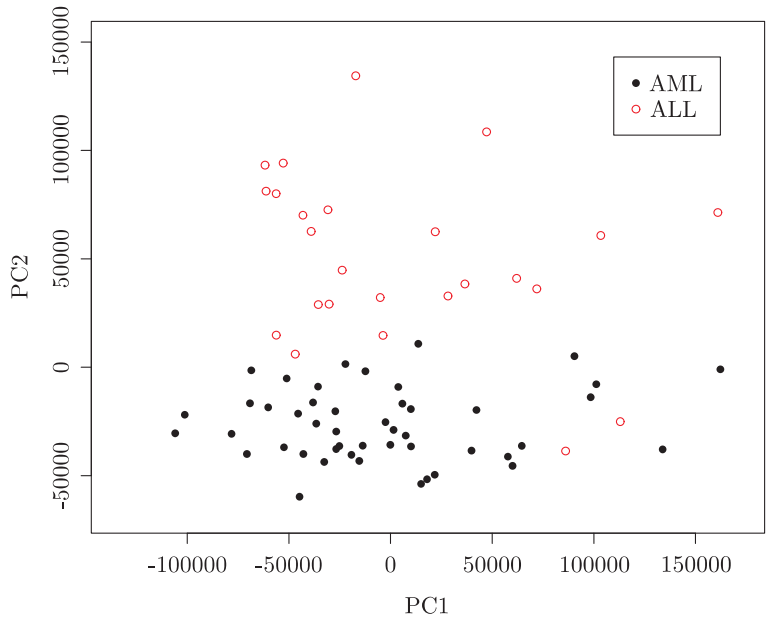

Figure 1. Two leading sparse PC scores for the Golub gene expression data.

truncation technique. Let $Y_\ell^i = X_\ell^{i2}/\sigma_\ell^2 - 1$ and take ξ such that $2(c+1)/m < \xi < 1/2$. Then,

$$P\big(|Y_\ell^i| > n^\xi \text{ for some } 1 \le i \le n\big) \le C_3 n^{1-m\xi/2} \le n^{-c-\eta},$$
$$E|Y_\ell^i| I(|Y_\ell^i| > n^\xi) \le C_4 n^{-(m/2-1)\xi} \le n^{-(1+c)/2},$$

the former holding for some $\eta > 0$. Define $Y_\ell^{i*} = Y_\ell^i I(|Y_\ell^i| \le n^\xi) - EY_\ell^i I(|Y_\ell^i| \le n^\xi)$. Applying (A.1) we get, for $\ell \ge d_0 + 1$,

$$\begin{aligned}
&P\Big(\hat\sigma_\ell^2 \ge \big[1 + c_1 n^{-1/2} (\log n)^C\big]\sigma^2\Big) \\
&\quad = P\Big(n^{-1/2}\sum_{i=1}^n Y_\ell^i > c_1 (\log n)^C\Big) \\
&\quad \le P\Big(n^{-1/2}\sum_{i=1}^n Y_\ell^{i*} > c_1 (\log n)^C - n^{-c/2}\Big) + n^{-c-\eta} \\
&\quad \le \exp\Big[-C_5 \frac{\big(c_1(\log n)^C - n^{-c/2}\big)^2}{1 + n^{\xi-1/2}\big(c_1(\log n)^C - n^{-c/2}\big)}\Big] + n^{-c-\eta} \\
&\quad \le n^{-C_6 (\log n)^{2C-1}} + n^{-c-\eta}.
\end{aligned}$$

This shows the left hand side of (3.1) is bounded by

$$O(n^c)\Big\{n^{-C_6(\log n)^{2C-1}} + n^{-c-\eta}\Big\} \to 0.$$

We can get a similar bound for the left hand side of (3.2). For this term, we have, if $1 \le \ell \le d_0$, then

$$P\Big(\hat{\sigma}_\ell^2 < \big[1 + c_2 n^{-1/2} (\log n)^C\big]\sigma^2\Big) \le P\left(n^{-1/2}\sum_{i=1}^n Y_\ell^i < -\frac{(\log n)^\kappa}{4}\right) \le n^{-C_7(\log n)^{2\kappa-1}} + n^{-c-\eta}.$$

A.2. Proof of Theorem 2

We first prove the second part. By Theorem 1, we may assume $\hat{\mathcal{I}} = \{1, \dots, d_0\}$ so that we can take $\tilde{\mathbf{X}}^i = (X_1^i, \dots, X_{d_0}^i)^\top$ and $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \dots, \hat{\beta}_{jd_0})^\top$ in this proof. Let $\boldsymbol{\beta}_j$ denote $(\beta_{j1}, \dots, \beta_{jd_0})^\top$, and $\tilde{\boldsymbol{\Sigma}}_0$ the $d_0 \times d_0$ top-left block of $\tilde{\boldsymbol{\Sigma}}$. Let $\hat{\mathbf{u}}_j = n^{1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ for $1 \le j \le K$. Then $\{\hat{\mathbf{u}}_j : 1 \le j \le K\}$ is the solution to the following problem: maximize

$$\sum_{j=1}^K \mathbf{u}_j^\top \tilde{\boldsymbol{\Sigma}}_0 \mathbf{u}_j + 2n^{1/2}\sum_{j=1}^K \boldsymbol{\beta}_j^\top \tilde{\boldsymbol{\Sigma}}_0 \mathbf{u}_j - n\sum_{j=1}^K \sum_{\ell=1}^{d_0} \mu_{j\ell}\left(|\beta_{j\ell} + n^{-1/2}u_{j\ell}| - |\beta_{j\ell}|\right)$$

subject to $\mathbf{u}_j^\top(2\boldsymbol{\beta}_j + n^{-1/2}\mathbf{u}_j) = 0$ for $1 \le j \le K$ and $\mathbf{u}_j^\top \boldsymbol{\beta}_{j'} + \mathbf{u}_{j'}^\top \boldsymbol{\beta}_j + n^{-1/2}\mathbf{u}_j^\top \mathbf{u}_{j'} = 0$ for $1 \le j \ne j' \le K$. Let $\mathbf{W}_j = n^{1/2}(\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\boldsymbol{\beta}_j$, where $\boldsymbol{\Sigma}_0$ is the $d_0 \times d_0$ top-left block of $\boldsymbol{\Sigma}$. Using the arguments in the proof of Theorem 1, one can verify that

$$\mathbf{u}_j^\top(\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\mathbf{u}_j = O_p\left(n^{-1/2+2C'}(\log n)^{1/2}\right),$$

uniformly for $\mathbf{u}_j \in \mathcal{U} \equiv \{\mathbf{v} : \sup_{1 \le \ell \le d_0} |v_\ell| \le A\}$, where $A > 0$ is arbitrary large and C' is the constant in (3.5). Thus, from the constraint $\mathbf{u}_j^\top(2\boldsymbol{\beta}_j + n^{-1/2}\mathbf{u}_j) = 0$, we get

$$\mathbf{u}_j^\top \tilde{\boldsymbol{\Sigma}}_0 \mathbf{u}_j + 2n^{1/2}\boldsymbol{\beta}_j^\top \tilde{\boldsymbol{\Sigma}}_0 \mathbf{u}_j \;=\; \mathbf{u}_j^\top(\boldsymbol{\Sigma}_0 - \lambda_j^0 \mathbf{I}_{d_0})\mathbf{u}_j + 2\mathbf{W}_j^\top \mathbf{u}_j + o_p(1)$$

uniformly for $\mathbf{u}_j$ in $\mathcal{U}$.

Next, since $\inf_{\ell \in \mathcal{J}(j)} |\beta_{j\ell}| \ge \delta > 0$ for $1 \le j \le K$, we have

$$\sup_{\ell \in \mathcal{J}(j)} \left| n^{1/2}\left(|\beta_{j\ell} + n^{-1/2}u_{j\ell}| - |\beta_{j\ell}|\right) - u_{j\ell}\operatorname{sgn}(\beta_{j\ell})\right| \;\to\; 0$$

uniformly for $\mathbf{u}_j$ in $\mathcal{U}$. This implies, with the first part of the condition (3.6), that

$$n\sum_{j=1}^K \sum_{\ell \in \mathcal{J}(j)} \mu_{j\ell}\left(|\beta_{j\ell} + n^{-1/2}u_{j\ell}| - |\beta_{j\ell}|\right) \;\xrightarrow{p}\; 0$$

uniformly for $\mathbf{u}_j \in \mathcal{U}$, $1 \le j \le K$. From this we deduce that

$$\begin{aligned} & n \sum_{j=1}^{K} \sum_{\ell=1}^{d_0} \mu_{j\ell} \left(|\beta_{j\ell} + n^{-1/2} \hat{u}_{j\ell}| - |\beta_{j\ell}| \right) \\ &= n^{1/2} \sum_{j=1}^{K} \sum_{\ell \notin \mathcal{J}(j)} \mu_{j\ell} |\hat{u}_{j\ell}| + o_p(1) \\ &\ge n^{1/2} \sum_{j=1}^{K} \Big(\inf_{\ell \notin \mathcal{J}(j)} \mu_{j\ell} \Big) \Big(\sup_{\ell \notin \mathcal{J}(j)} |\hat{u}_{j\ell}| \Big) + o_p(1). \end{aligned} \tag{A.2}$$

Let $N_j(\epsilon) = \big\{ \sup_{\ell \notin \mathcal{J}(j)} |\hat{u}_{j\ell}| \ge \epsilon \big\}$. On the event $\bigcup_{j=1}^{K} N_j(\epsilon)$,

$$n^{1/2} \sum_{j=1}^{K} \Big(\inf_{\ell \notin \mathcal{J}(j)} \mu_{j\ell} \Big) \Big(\sup_{\ell \notin \mathcal{J}(j)} |\hat{u}_{j\ell}| \Big) \ge \epsilon\, n^{1/2} \inf_{\ell \notin \mathcal{J}(j)} \mu_{j\ell}$$

which, by the second part of the condition (3.6), goes to infinity faster than $n^{2C'}$. On the other hand, using the arguments in the proof of Theorem 1, one can show

$$\sum_{j=1}^{K} \sup_{\mathbf{u}_j \in \mathcal{U}} |\mathbf{W}_j^\top \mathbf{u}_j| = O_p \left(n^{C'} (\log n)^{1/2} \right).$$

This implies that $P\big[\bigcup_{j=1}^{K} N_j(\epsilon) \big] \to 0$ for any $\epsilon > 0$.

Let $\{ \tilde{u}_{j\ell} : \ell \in \mathcal{J}(j),\, 1 \le j \le K \}$ be the maximizer of

$$\sum_{j=1}^{K} \sum_{\ell \in \mathcal{J}(j)} \sum_{\ell' \in \mathcal{J}(j)} (\Sigma_{\ell\ell'} - \lambda_j^0 d_{\ell\ell'}) u_{j\ell} u_{j\ell'} + 2 \sum_{j=1}^{K} \sum_{\ell \in \mathcal{J}(j)} W_{j\ell} u_{j\ell}$$

subject to $\sum_{\ell \in \mathcal{J}(j)} u_{j\ell} \beta_{j\ell} = 0$ for $1 \le j \le K$ and $\sum_{\ell \in \mathcal{J}(j) \cap \mathcal{J}(j')} (u_{j\ell} \beta_{j'\ell} + u_{j'\ell} \beta_{j\ell}) = 0$, where $d_{\ell\ell'} = 1$ if $\ell = \ell'$ and zero otherwise. Then, the foregoing arguments establish that, for all $1 \le j \le K$,

$$\sup_{\ell \in \mathcal{J}(j)} |\hat{u}_{j\ell} - \tilde{u}_{j\ell}| = o_p(1), \quad \sup_{\ell \notin \mathcal{J}(j)} |\hat{u}_{j\ell}| = o_p(1). \tag{A.3}$$

Note also that since $\sup_{\ell \in \mathcal{J}(j)} |W_{j\ell}| = O_p(1)$,

$$\sup_{\ell \in \mathcal{J}(j)} |\tilde{u}_{j\ell}| = O_p(1) = \sup_{\ell \in \mathcal{J}(j)} |\hat{u}_{j\ell}| = n^{1/2} \sup_{\ell \in \mathcal{J}(j)} |\hat{\beta}_{j\ell} - \beta_{j\ell}| \tag{A.4}$$

for $1 \le j \le K$. The first result of (A.3) concludes part (ii) of the theorem.

To prove the first part of the theorem, we verify

$$P[\mathcal{J}(j) \subset \hat{\mathcal{J}}(j)] \to 1 \text{ for all } 1 \le j \le K, \tag{A.5}$$

$$P[\mathcal{J}(j) \supset \hat{\mathcal{J}}(j)] \to 1 \text{ for all } 1 \le j \le K. \tag{A.6}$$

The property (A.5) is immediate from (A.4), since for $\delta_j \equiv \inf_{\ell \in \mathcal{J}(j)} |\beta_{j\ell}| > 0$ the latter entails $\inf_{\ell \in \mathcal{J}(j)} |\hat{\beta}_{j\ell}| > \delta_j/2$ with probability tending to one. To show (A.6), we consider the Lagrangian problems that maximize, successively for $j = 1, \ldots, K$,

$$\boldsymbol{\beta}_j^\top \tilde{\boldsymbol{\Sigma}}_0 \boldsymbol{\beta}_j - \sum_{\ell=1}^{d_0} \mu_{j\ell} |\beta_{j\ell}| - \zeta_j (\boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j - 1) - \sum_{j'=1}^{j-1} \zeta_{jj'} \boldsymbol{\beta}_j^\top \boldsymbol{\beta}_{j'},$$

where ζ_j and $\zeta_{jj'}$ are the Lagrange multipliers. If there exists $\ell_0 \notin \mathcal{J}(j_0)$ but $\ell_0 \in \hat{\mathcal{J}}(j_0)$ for some $1 \le j_0 \le K$, then for such j_0 and ℓ_0, by the Kuhn-Tucker Theorem,

$$(\tilde{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\beta}}_{j_0})_{\ell_0} - \zeta_{j_0} \hat{\beta}_{j_0 \ell_0} - \sum_{j'=1}^{j_0-1} \zeta_{j_0 j'} \hat{\beta}_{j'\ell_0} \;=\; \frac{1}{2} \mu_{j_0 \ell_0} \operatorname{sgn}(\hat{\beta}_{j_0 \ell_0}). \tag{A.7}$$

Since $\sup_{1 \le \ell \le d_0} |\hat{\beta}_{j\ell} - \beta_{j\ell}| = O_p(n^{-1/2})$ for all $1 \le j \le K$ by (A.4), $\beta_{j_0 \ell_0} = 0$, $(\boldsymbol{\Sigma}_0 \boldsymbol{\beta}_{j_0})_{\ell_0} = \lambda_{j_0}^0 \beta_{j_0 \ell_0} = 0$, and

$$\tilde{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\beta}}_j = (\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + n^{-1/2} \mathbf{W}_j + \boldsymbol{\Sigma}_0 (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) + \boldsymbol{\Sigma}_0 \boldsymbol{\beta}_j,$$

we can deduce that

$$\left| (\tilde{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\beta}}_{j_0})_{\ell_0} - \zeta_{j_0} \hat{\beta}_{j_0 \ell_0} - \sum_{j'=1}^{j_0-1} \zeta_{j_0 j'} \hat{\beta}_{j'\ell_0} \right| = O_p\left(n^{-1/2+C'} (\log n)^{1/2} \right). \tag{A.8}$$

On the other hand, by the second part of the condition (3.6),

$$|\mu_{j_0 \ell_0}|^{-1} = o_p\left(n^{1/2-C'} (\log n)^{-1/2} \right).$$

This with (A.7) and (A.8) establishes (A.6).

A.3. Proof of Theorem 3

Let $\boldsymbol{\beta}_j = (\beta_{j\ell} : 1 \le \ell \le d_0)^\top$. Since eigenvectors are determined uniquely up to sign, we take $\bar{\boldsymbol{\beta}}_j^\top \boldsymbol{\beta}_j \ge 0$. By using the arguments in the proof of Theorem 1, it can be proved that

$$\sup_{1 \le j, \ell \le d_0} \left| n^{-1} \sum_{i=1}^n X_j^i X_\ell^i - E X_j X_\ell \right| = O_p\left(n^{-1/2} (\log n)^{1/2} \right).$$

For a matrix A, let $\|A\|_{op} = \sup_{x:\|x\|=1} \|Ax\|$ denote the operator norm of A. Then

$$\|\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0\|_{op} = O_p\left(n^{-1/2+C'}(\log n)^{1/2}\right), \tag{A.9}$$

where $\boldsymbol{\Sigma}_0$ is the $d_0 \times d_0$ top-left block of $\boldsymbol{\Sigma}$.

Now, by Corollary 8.1.6 of Golub and Van Loan (1996) we have

$$|\tilde{\lambda}_j^0 - \lambda_j^0| \le \|\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0\|_{op}, \tag{A.10}$$

where $\tilde{\lambda}_1^0 \ge \cdots \ge \tilde{\lambda}_{d_0}^0 \ge 0$ is an enumeration of the eigenvalues of $\tilde{\boldsymbol{\Sigma}}_0$. Due to (3.8), this means that

$$P\left(|\tilde{\lambda}_j^0 - \lambda_\ell^0| > 0 \text{ for all } \ell \ne j \text{ and } 1 \le j \le K\right) \to 1.$$

Note that $\boldsymbol{\beta}_\ell, 1 \le \ell \le d_0$, form an orthonormal basis for $\mathbb{R}^{d_0}$. Thus, on a set with probability tending to one, we obtain

$$\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j = \sum_{\ell:\ell\ne j} (\tilde{\lambda}_j^0 - \lambda_\ell^0)^{-1} \bar{\boldsymbol{\beta}}_j^\top (\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\boldsymbol{\beta}_\ell \boldsymbol{\beta}_\ell + (\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top \boldsymbol{\beta}_j \boldsymbol{\beta}_j,\ 1 \le j \le K$$

because $\bar{\boldsymbol{\beta}}_j^\top (\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\boldsymbol{\beta}_\ell = (\tilde{\lambda}_j^0 - \lambda_\ell^0)(\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top \boldsymbol{\beta}_\ell$ for all $j \ne \ell$. This basis expansion of $\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$ with respect to $\{\boldsymbol{\beta}_\ell\}_{\ell=1}^{d_0}$ gives $\|\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|^2 = P_j^2 + Q_j^2$, where $P_j^2 = \sum_{\ell:\ell\ne j} (\tilde{\lambda}_j^0 - \lambda_\ell^0)^{-2} \left[\bar{\boldsymbol{\beta}}_j^\top (\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\boldsymbol{\beta}_\ell\right]^2$ and $Q_j^2 = \left[(\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top \boldsymbol{\beta}_j\right]^2$. Note that P_j is the norm of the projection of $\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$ onto the linear subspace of $\mathbb{R}^{d_0}$ generated by $\{\boldsymbol{\beta}_\ell : \ell \ne j\}$, so that it also is the norm of the projection of $\bar{\boldsymbol{\beta}}_j$ onto that subspace. Thus, $P_j^2 + (\bar{\boldsymbol{\beta}}_j^\top \boldsymbol{\beta}_j)^2 = 1$. This gives

$$Q_j^2 = \left[1 - (1 - P_j^2)^{1/2}\right]^2 = 2\left[1 - (1 - P_j^2)^{1/2}\right] - P_j^2,$$

so that $\|\bar{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|^2 \le 2P_j^2$.

It suffices to prove

$$P_j^2 = O_p\left(n^{-1+2C'}(\log n)^{1+2C''}\right).$$

By (A.9), (A.10) and (3.8),

$$P\left[(\tilde{\lambda}_j^0 - \lambda_\ell^0)^{-2} \le C_0(\log n)^{2C''} \text{ for all } \ell \ne j \text{ and } 1 \le j \le K\right] \to 1$$

for some $C_0 > 0$. Since $\sum_{\ell=1}^{d_0} \left[\bar{\boldsymbol{\beta}}_j^\top (\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\boldsymbol{\beta}_\ell\right]^2 = \|(\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\bar{\boldsymbol{\beta}}_j\|^2$, it follows that $P_j^2 \le C_0(\log n)^{2C''}\|\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0\|_{op}^2$ on a set with probability tending to one.

References

Bair, E., Hastie, T. and Tibshirani, T. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101**, 119-137.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.

d'Aspremont, A., Bach, F. and Ghaoui, L. E. (2008). Optimal solution for sparse principal component analysis. *J. Machine Learning Research* **9**, 1269-1294.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I. and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming, *SIAM Rev.* **49**, 434-448.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations.* Johns Hopkins University Press, Baltimore.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295-327.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104**, 682-693.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12**, 531-547.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37**, 4104-4130.

Leng, C. and Wang, H. (2009). On general adaptive sparse principal component Analysis. *J. Comput. Graph. Statist.* **18**, 201-215.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498-3528.

Meinshausen, N. and Yu, B. (2009). LASSO-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 246-270.

Noh, H. S. (2009). Variable selection in sparse nonparametric models. Ph.D. dissertation, Seoul National University.

Noh, H. S. and Park, B. U. (2010). Sparse varying coefficient models for longitudinal data. *Statist. Sinica* **20**, 1183-1202.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17**, 1617-1742.

Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.* **36**, 1595-1618.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99**, 1015-1034.

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

Xie, H. and Hunag, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37**, 673-696.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concavity penalty. *Ann. Statist.* **38**, 894-942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15**, 265-286.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509-1566.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic net with diverging number of parameter. *Ann. Statist.* **37**, 1733–1751.

Department of Statistics, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do, 200-701 Republic of Korea.

E-mail: youngklee@kangwon.ac.kr

Department of Statistics, Seoul National University, Seoul 151–747, Korea.

E-mail: silverryuee@gmail.com

Department of Statistics, Seoul National University, Seoul 151–747, Korea.

E-mail: bupark@stats.snu.ac.kr

(Received July 2010; accepted August 2011)