2013

Sparse Principal Component Analysis and Iterative Thresholding

Zongming Ma
University of Pennsylvania

Recommended Citation

Ma, Z. (2013). Sparse Principal Component Analysis and Iterative Thresholding. *Annals of Statistics, 41* (2), 772-801. http://dx.doi.org/10.1214/13-AOS1097

Sparse Principal Component Analysis and Iterative Thresholding

Abstract

.

Principal component analysis (PCA) is a classical dimension reduction method which projects data onto the principal subspace spanned by the leading eigenvectors of the covariance matrix. However, it behaves poorly when the number of features p is comparable to, or even much larger than, the sample size n. In this paper, we propose a new iterative thresholding approach for estimating principal subspaces in the setting where the leading eigenvectors are sparse. Under a spiked covariance model, we find that the new approach recovers the principal subspace and leading eigenvectors consistently, and even optimally, in a range of high-dimensional sparse settings. Simulated examples also demonstrate its competitive performance.

Keywords

dimension reduction, high-dimensional statistics, principal component analysis, principal subspace, sparsity, spiked covariance model, thresholding

Disciplines

Statistics and Probability

The Annals of Statistics
2013, Vol. 41, No. 2, 772–801
DOI: 10.1214/13-AOS1097

SPARSE PRINCIPAL COMPONENT ANALYSIS AND ITERATIVE THRESHOLDING

BY ZONGMING MA

University of Pennsylvania

Principal component analysis (PCA) is a classical dimension reduction method which projects data onto the principal subspace spanned by the leading eigenvectors of the covariance matrix. However, it behaves poorly when the number of features p is comparable to, or even much larger than, the sample size n. In this paper, we propose a new iterative thresholding approach for estimating principal subspaces in the setting where the leading eigenvectors are sparse. Under a spiked covariance model, we find that the new approach recovers the principal subspace and leading eigenvectors consistently, and even optimally, in a range of high-dimensional sparse settings. Simulated examples also demonstrate its competitive performance.

1. Introduction. In many contemporary datasets, if we organize the p-dimensional observations $x_1, \ldots, x_n$, into the rows of an $n \times p$ data matrix X, the number of features p is often comparable to, or even much larger than, the sample size n. For example, in biomedical studies, we usually have measurements on the expression levels of tens of thousands of genes, but only for tens or hundreds of individuals. One of the crucial issues in the analysis of such "large p" datasets is dimension reduction of the feature space.

As a classical method, principal component analysis (PCA) [8, 23] reduces dimensionality by projecting the data onto the *principal subspace* spanned by the m leading eigenvectors of the population covariance matrix Σ, which represent the principal modes of variation. In principle, one expects that for some $m < p$, most of the variance in the data is captured by these m modes. Thus, PCA reduces the dimensionality of the feature space while retaining most of the information in data. In addition, projection to a low-dimensional space enables visualization of the data. In practice, Σ is unknown. Classical PCA then estimates the leading population eigenvectors by those of the sample covariance matrix S. It performs well in the traditional data setting where p is small and n is large [2].

In high-dimensional settings, a collection of data can be modeled by a low-rank signal plus noise structure, and PCA can be used to recover the low-rank signal. In particular, each observation vector x_i can be viewed as an independent

Received September 2012; revised January 2013.
MSC2010 subject classifications. Primary 62H12; secondary 62G20, 62H25.
Key words and phrases. Dimension reduction, high-dimensional statistics, principal component analysis, principal subspace, sparsity, spiked covariance model, thresholding.

instantiation of the following generative model:

$$x_i = \mu + Au_i + \sigma z_i. \tag{1.1}$$

Here, μ is the mean vector, A is a $p \times \bar{m}$ deterministic matrix of factor loadings, u_i is an $\bar{m}$-vector of random factors, $\sigma > 0$ is the noise level and z_i is a p-vector of white noise. For instance, in chemometrics, x_i can be a vector of the logarithm of the absorbance or reflectance spectra measured with noise, where the columns of A are characteristic spectral responses of different chemical components, and u_i's the concentration levels of these components [31]. The number of observations are relatively few compared with the number of frequencies at which the spectra are measured. In econometrics, x_i can be the returns for a collection of assets, where the u_i's are the unobservable random factors [29]. The assumption of additive white noise is reasonable for asset returns with low frequencies (e.g., monthly returns of stocks). Here, people usually look at tens or hundreds of assets simultaneously, while the number of observations are also at the scale of tens or hundreds. In addition, model (1.1) represents a big class of signal processing problems [32]. Without loss of generality, we assume $\mu = 0$ from now on.

In this paper, our primary interest lies in PCA of high-dimensional data generated as in (1.1). Let the covariance matrix of u_i be Φ which is of full rank. Suppose that A has full column rank and that u_i and z_i are independent. Then the covariance matrix of x_i becomes

$$\Sigma = A\Phi A' + \sigma^2 I = \sum_{j=1}^{\bar{m}} \lambda_j^2 q_j q_j' + \sigma^2 I. \tag{1.2}$$

Here, $\lambda_1^2 \geq \cdots \geq \lambda_{\bar{m}}^2 > 0$ are the eigenvalues of $A\Phi A'$, with q_j, $j = 1, \ldots, \bar{m}$, the associated eigenvectors. Therefore, the jth eigenvalue of Σ is $\lambda_j^2 + \sigma^2$ for $j = 1, \ldots, \bar{m}$, and σ^2 otherwise. Since there are $\bar{m}$ spikes $(\lambda_1^2, \ldots, \lambda_{\bar{m}}^2)$ in the spectrum of Σ, (1.2) has been called the *spiked covariance model* in the literature [10]. Note that we use λ_j^2 to denote the spikes rather than λ_j used previously in the literature [22]. For data with such a covariance structure, it makes sense to project the data onto the low-dimensional subspaces spanned by the first few q_j's. Here and after, $\bar{m}$ denotes the number of spikes in the model, and m is the target dimension of the principal subspace to be estimated, which is no greater than $\bar{m}$.

Classical PCA encounters both practical and theoretical difficulties in high dimensions. On the practical side, the eigenvectors found by classical PCA involve all the p features, which makes their interpretation challenging. On the theoretical side, the sample eigenvectors are no longer always consistent estimators. Sometimes, they can even be nearly orthogonal to the target direction. When both $n, p \to \infty$ with $n/p \to c \in (0, \infty)$, at different levels of rigor and generality, this phenomenon has been examined by a number of authors [9, 14, 17, 19, 21, 25] under model (1.2). See [13] for similar results when $p \to \infty$ and n is fixed.

In recent years, to facilitate interpretation, researchers have started to develop sparse PCA methodologies, where they seek a set of sparse vectors spanning the low-dimensional subspace that explains most of the variance. See, for example, [3, 12, 27, 30, 34, 36]. These approaches typically start with a certain optimization formulation of PCA and then induce a sparse solution by introducing appropriate penalties or constraints.

On the other hand, when Σ indeed has sparse leading eigenvectors in the current basis (perhaps after transforming the data), it becomes possible to estimate them consistently under high-dimensional settings via new estimation schemes. For example, under normality assumption, when Σ only has a single spike, that is, when $\bar{m} = 1$ in (1.2), Johnstone and Lu [11] proved consistency of PCA obtained on a subset of features with large sample variances when the leading eigenvalue is fixed and $(\log p)/n \to 0$. Under the same single spike model, if in addition the leading eigenvector has exactly k nonzero loadings, Amini and Wainwright [1] studied conditions for recovering the nonzero locations using the methods in [11] and [3], and Shen et al. [26] established conditions for consistency of a sparse PCA method in [27] when $p \to \infty$ and n is fixed. For the more general multiple component case, Paul and Johnstone [22] proposed an augmented sparse PCA method for estimating each of the leading eigenvectors, and showed that their procedure attains near optimal rate of convergence under a range of high-dimensional sparse settings when the leading eigenvalues are comparable and well separated. Notably, these methods all focus on estimating individual eigenvectors.

In this paper, we focus primarily on finding *principal subspaces* of Σ spanned by sparse leading eigenvectors, as opposed to finding each sparse vector individually. One of the reasons is that individual eigenvectors are not identifiable when some leading eigenvalues are identical or close to each other. Moreover, if we view PCA as a dimension reduction technique, it is the low-dimensional subspace onto which we project data that is of the greatest interest.

We propose a new iterative thresholding algorithm to estimate principal subspaces, which is motivated by the orthogonal iteration method in matrix computation. In addition to the usual orthogonal iteration steps, an additional thresholding step is added to seek sparse basis vectors for the subspace. When Σ follows the spiked covariance model and the sparsity of the leading eigenvectors are characterized by the weak-ℓ_r condition (3.5), the algorithm leads to a consistent subspace estimator adaptively over a wide range of high-dimensional sparse settings, and the rates of convergence are derived under an appropriate loss function (2.1). Moreover, for any individual leading eigenvector whose eigenvalue is well separated from the rest of the spectrum, our algorithm also yields an eigenvector estimator which adaptively attains optimal rate of convergence derived in [22] up to a multiplicative log factor. In addition, it has appealing model selection property in the sense that the resulting estimator only involves coordinates with large signal-to-noise ratios.

The contribution of the current paper is threefold. First, we propose to estimate principal subspaces. This is natural for the purpose of dimension reduction and visualization, and avoids the identifiability issue for individual eigenvectors. Second, we construct a new algorithm to estimate the subspaces, which is efficient in computation and easy to implement. Last but not least, we derive convergence rates of the resulting estimator under the spiked covariance model when the eigenvectors are sparse.

The rest of the paper is organized as follows. In Section 2, we frame the principal subspace estimation problem and propose the iterative thresholding algorithm. The statistical properties and computational complexity of the algorithm are examined in Sections 3 and 4 under normality assumption. Simulation results in Section 5 demonstrate its competitive performance. Section 6 presents the proof of the main theorems.

Reproducible code: The MATLAB package `SPCALab` implementing the proposed method and producing the tables and figures of the current paper is available at the author's website.

2. Methodology.

2.1. *Notation*. We say x is a p-vector if $x \in \mathbb{R}^p$, and we use $\|x\|_2$ to denote its Euclidean norm. For an $m \times n$ matrix A, its submatrix with rows indexed by I and columns indexed by J is denoted by A_{IJ}. If I or J includes all the indices, we replace it with a dot. For example, $A_{I\cdot}$ is the submatrix of A with rows in I and all columns. The spectral norm of A is $\|A\| = \max_{\|x\|_2=1} \|Ax\|_2$, and the range, that is, the column subspace, of A is $\operatorname{ran}(A)$. If $m \geq n$, and the columns of A form an orthonormal set in $\mathbb{R}^m$, we say A is orthonormal.

We use C, C_0, C_1, etc. to represent constants, though their values might differ at different occurrences. For real numbers a and b, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We write $a_n = O(b_n)$, if there is a constant C, such that $|a_n| \leq C b_n$ for all n, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Moreover, we write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Throughout the paper, we use ν as the generic index for features, i for observations, j for eigenvalues and eigenvectors and k for iterations in the algorithm to be proposed.

2.2. *Framing the problem*: *Principal subspace estimation*. When the covariance matrix Σ follows model (1.2), its jth largest eigenvalue $\ell_j(\Sigma) = \lambda_j^2 + \sigma^2$ for $j = 1, \ldots, \bar{m}$ and equals σ^2 for all $j > \bar{m}$. Let $\operatorname{span}\{\cdot\}$ denote the linear subspace spanned by the vectors in the curly brackets. If for some $m \leq \bar{m}$, $\ell_m(\Sigma) > \ell_{m+1}(\Sigma)$, the *principal subspace*

$$\mathcal{P}_m = \operatorname{span}\{q_1, \ldots, q_m\}$$

is defined, regardless of the behavior of the other $\ell_j(\Sigma)$'s. Therefore, it is an identifiable object for the purpose of estimation. Note that $\mathcal{P}_{\bar{m}}$ is always identifiable,

because $\ell_{\bar{m}}(\Sigma) > \ell_{\bar{m}+1}(\Sigma)$. The primary goal of this paper is to estimate the principal subspace $\mathcal{P}_m$, for some $m \le \bar{m}$ with $\ell_m(\Sigma) > \ell_{m+1}(\Sigma)$. More precisely, we require the gap $\ell_m(\Sigma) - \ell_{m+1}(\Sigma) = \lambda_m^2 - \lambda_{m+1}^2 \asymp \lambda_1^2$. Note that such an m always exists, for example, the largest $m \le \bar{m}$ such that $\lambda_m^2 \asymp \lambda_1^2$. We allow the case of $m < \bar{m}$ partly because under certain circumstances, one might not be interested in $\mathcal{P}_{\bar{m}}$ directly. For example, to visualize the data, one might want to estimate $\mathcal{P}_2$ or $\mathcal{P}_3$ while $\bar{m}$ could be larger than 3. In addition, sometimes $\mathcal{P}_{\bar{m}}$ might not be consistently estimable while some smaller principal subspace $\mathcal{P}_m$ is. In most part of the paper, we assume that an appropriate m is given for convenience. In Section 3.5, we discuss how to choose m and how to estimate $\bar{m}$ under normality assumption.

To measure the accuracy of an estimator $\widehat{\mathcal{S}}$ for a subspace $\mathcal{S}$, note that each linear subspace is associated with a unique projection matrix onto it. Let P and $\widehat{P}$ be the projection matrices associated with $\mathcal{S}$ and $\widehat{\mathcal{S}}$, respectively. The distance between $\mathcal{S}$ and $\widehat{\mathcal{S}}$ is given by the spectral norm of the difference between P and $\widehat{P}$: $\operatorname{dist}(\mathcal{S}, \widehat{\mathcal{S}}) = \|P - \widehat{P}\|$; see [7], Section 2.6.3. Thus, we can define a loss function by the squared distances between $\mathcal{S}$ and $\widehat{\mathcal{S}}$,

$$L(\mathcal{S}, \widehat{\mathcal{S}}) = \operatorname{dist}^2(\mathcal{S}, \widehat{\mathcal{S}}) = \|P - \widehat{P}\|^2. \tag{2.1}$$

By definition, this loss function measures the maximum possible discrepancy between the projections of any unit vector onto the two subspaces. The loss ranges in $[0, 1]$, and equals zero if and only if $\widehat{\mathcal{S}} = \mathcal{S}$. When $\dim(\widehat{\mathcal{S}}) \ne \dim(\mathcal{S})$, we have $L(\mathcal{S}, \widehat{\mathcal{S}}) = 1$. Geometrically, it equals the squared sine of the largest canonical angle between $\mathcal{S}$ and $\widehat{\mathcal{S}}$ ([28], Theorem 5.5). Throughout the paper, we use the loss function (2.1) for principal subspace estimation.

2.3. *Orthogonal iteration.* Given a positive definite matrix A, a standard technique to compute its leading eigenspace is orthogonal iteration [7]. When only the first eigenvector is sought, it is also known as the power method.

To state the orthogonal iteration method, we note that for any $p \times m$ matrix T, when $p \ge m$, we could decompose it into the product of two matrices $T = QR$, where Q is $p \times m$ orthonormal and R is $m \times m$ upper triangular. This decomposition is called QR factorization and can be computed using Gram–Schmidt orthogonalization and other numerical methods [7]. Suppose A is $p \times p$, and we want to compute its leading eigenspace of dimension m. Starting with a $p \times m$ orthonormal matrix $Q^{(0)}$, orthogonal iteration generates a sequence of $p \times m$ orthonormal matrices $Q^{(k)}$, $k = 1, 2, \ldots$, by alternating the following two steps till convergence:

(1) Multiplication: $T^{(k)} = AQ^{(k-1)}$;
(2) QR factorization: $Q^{(k)}R^{(k)} = T^{(k)}$.

Denote the orthonormal matrix at convergence by $Q^{(\infty)}$. Then its columns are the leading eigenvectors of A, and $\operatorname{ran}(Q^{(\infty)})$ gives the eigenspace. In practice, one terminates the iteration once $\operatorname{ran}(Q^{(k)})$ stabilizes.

When we apply orthogonal iteration directly to the sample covariance matrix S, it gives the classical PCA result, which could be problematic in high dimensions. Observe that all the p features are included in orthogonal iteration. When the dimensionality is high, not only the interpretation is hard, but the variance accumulated across all the features becomes so high that it makes consistent estimation impossible.

If the eigenvectors spanning $\mathcal{P}_m$ are sparse in the current basis, one sensible way to reduce estimation error is to focus only on those features at which the leading eigenvectors have large values, and to estimate other features by zeros. Of course, one introduces bias this way, but hopefully it is much smaller compared to the amount of variance thus reduced.

The above heuristics lead to the estimation scheme in the next subsection which incorporates this feature screening idea in orthogonal iteration.

2.4. *Iterative thresholding algorithm.* Let $S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$ be the sample covariance matrix. An effective way to incorporate feature screening into orthogonal iteration is to "kill" small coordinates of the $T^{(k)}$ matrix after each multiplication step, which leads to the estimation scheme summarized in Algorithm 1. Although the later theoretical study is conducted under normality assumption, Algorithm 1 itself is not confined to normal data.

In addition to the two basic orthogonal iteration steps, Algorithm 1 adds a thresholding step in between them, where we threshold each element of $T^{(k)}$ with a user-specified thresholding function η which satisfies

$$|\eta(t,\gamma) - t| \le \gamma \quad \text{and} \quad \eta(t,\gamma)1_{(|t|\le\gamma)} = 0 \qquad \text{for all } t \text{ and all } \gamma > 0. \tag{2.2}$$

Here, $1_{(E)}$ denotes the indicator function of an event E. We note that both hard-thresholding $\eta_H(t,\gamma) = t1_{(|t|>\gamma)}$ and soft-thresholding $\eta_S(t,\gamma) = \operatorname{sgn}(t)(|t| -$

Algorithm 1: ITSPCA (Iterative thresholding sparse PCA)

Input:
(1) Sample covariance matrix S;
(2) Target subspace dimension m;
(3) Thresholding function η, and threshold levels γ_{nj}, $j = 1, \ldots, m$;
(4) Initial orthonormal matrix $\widehat{Q}^{(0)}$.

Output: Subspace estimator $\widehat{\mathcal{P}}_m = \operatorname{ran}(\widehat{Q}^{(\infty)})$, where $\widehat{Q}^{(\infty)}$ denotes the $\widehat{Q}^{(k)}$ matrix at convergence.

repeat
  Multiplication: $T^{(k)} = (t_{vj}^{(k)}) = S\widehat{Q}^{(k-1)}$;
  Thresholding: $\widehat{T}^{(k)} = (\widehat{t}_{vj}^{(k)})$, with $\widehat{t}_{vj}^{(k)} = \eta(t_{vj}^{(k)}, \gamma_{nj})$;
  QR factorization: $\widehat{Q}^{(k)}\widehat{R}^{(k)} = \widehat{T}^{(k)}$;
until *convergence* ;

$\gamma)_+$ satisfy (2.2). So does any η sandwiched by them, such as that resulting from a SCAD criterion [6]. In $\eta(t, \gamma)$, the parameter γ is called the threshold level. In Algorithm 1, for each column of $T^{(k)}$, a common threshold level γ_{nj} needs to be specified for all its elements, which remains unchanged across iterations. The subscripts of γ_{nj} indicate that it depends on both the size of the problem n and the index j of the column it is applied to.

REMARK 2.1. The ranges of $\widehat{Q}^{(k)}$ and $\widehat{T}^{(k)}$ are the same because QR factorization only amounts to a basis change within the same subspace. However, as in orthogonal iteration, the QR step is essential for numerical stability, and should not be omitted. Moreover, although the algorithm is designed for subspace estimation, the column vectors of $\widehat{Q}^{(\infty)}$ can be used as estimators of leading eigenvectors.

Initialization. Algorithm 1 requires an initial orthonormal matrix $\widehat{Q}^{(0)}$. It can be generated from the "diagonal thresholding" sparse PCA algorithm [11]. Its multiple eigenvector version is summarized in Algorithm 2. Here, for any set I, $\mathrm{card}(I)$ denotes its cardinality. Given the output $\widehat{Q}_B = [\widehat{q}_1, \ldots, \widehat{q}_{\mathrm{card}(B)}]$ of Algorithm 2, we take $\widehat{Q}^{(0)} = [\widehat{q}_1, \ldots, \widehat{q}_m]$. When σ^2 is unknown, we could replace it by an estimator $\widehat{\sigma}^2$ in the definition of B. For example, for normal data, Johnstone and Lu [11] suggested

$$\widehat{\sigma}^2 = \mathrm{median}\left(\frac{1}{n}\sum_{i=1}^{n} x_{iv}^2\right). \tag{2.3}$$

When available, subject knowledge could also be incorporated into the construction of $\widehat{Q}^{(0)}$. Algorithm 1 also requires inputs for the γ_{nj}'s and subspace dimension m. Under normality assumption, we give explicit specification for them in

Algorithm 2: DTSPCA (Diagonal thresholding sparse PCA)

Input:
(1) Sample covariance matrix S;
(2) Diagonal thresholding parameter α_n.

Output: Orthonormal matrix $\widehat{Q}_B$.

1 Variance selection: select the set B of coordinates (which are likely to have "big" signals):

$$B = \{v : s_{vv} \geq \sigma^2(1 + \alpha_n)\};$$

2 Reduced PCA: compute the eigenvectors, $\widehat{q}_1^B, \ldots, \widehat{q}_{\mathrm{card}(B)}^B$, of the submatrix S_{BB};

3 Zero-padding: construct $\widehat{Q}_B = [\widehat{q}_1, \ldots, \widehat{q}_{\mathrm{card}(B)}]$ such that

$$\widehat{q}_{jB} = \widehat{q}_j^B, \qquad \widehat{q}_{jB^c} = 0, \qquad j = 1, \ldots, \mathrm{card}(B).$$

(3.3) and (3.15) later. Under the conditions of the later Section 3, B is nonempty with probability tending to 1, and so $\widehat{Q}^{(0)}$ is well defined.

Convergence. For normal data, to obtain the error rates in later Theorems 3.1 and 3.2, we can terminate Algorithm 1 after K_s iterations with K_s given in (3.4). In practice, one could also stop iterating if the difference between successive iterates becomes sufficiently small, for example, when $L(\mathrm{ran}(\widehat{Q}^{(k)}), \mathrm{ran}(\widehat{Q}^{(k+1)})) \le n^{-2}$. We suggest this empirical stopping rule because n^{-2} typically tends to zero faster than the rates we shall obtain, and so intuitively it should not change the statistical performance of the resulting estimator. In simulation studies reported in Section 5, the difference in numerical performance between the outputs based on this empirical stopping rule and those based on the theoretical rule (3.4) is negligible compared to the estimation errors. Whether Algorithm 1 always converges numerically is an interesting question left for possible future research.

Bibliographical note. When $m = 1$, Algorithm 1 is similar to the algorithms proposed in [27, 34] and [35]. When $m > 1$, all these methods propose to iteratively find the first leading eigenvectors of residual covariance matrices, which becomes different from our approach.

3. Statistical properties. This section is devoted to analyzing the statistical properties of Algorithm 1 under normality assumption. After some preliminaries, we first establish the convergence rates for subspace estimation in a special yet interesting case in Section 3.1. Then we introduce a set of general assumptions in Section 3.2 and a few key quantities in Section 3.3. Section 3.4 states the main results, which include convergence rates for principal subspace estimation under general assumptions and a correct exclusion property. In addition, we derive rates for estimating individual eigenvectors. For conciseness, we first state all the results assuming a suitable target subspace dimension $m \le \bar{m}$ is given. In Section 3.5, we discuss how to choose m and estimate $\bar{m}$ based on data.

We start with some preliminaries. Under normality assumption, $x_1, \ldots, x_n$ are i.i.d. $N_p(0, \Sigma)$ distributed, with Σ following model (1.2). Further assume σ^2 is known—though this assumption could be removed by estimating σ^2 using, say, $\widehat{\sigma}^2$ in (2.3). Since one can always scale the data first, we assume $\sigma^2 = 1$ from now on. Thus, (1.1) reduces to the orthogonal factor form

$$x_i = \sum_{j=1}^{\bar{m}} \lambda_j v_{ij} q_j + z_i, \qquad i = 1, \ldots, n. \tag{3.1}$$

Here, v_{ij} are i.i.d. standard normal random factors, which are independent of the i.i.d. white noise vectors $z_i \sim N_p(0, I)$, and $\{q_j, 1 \le j \le \bar{m}\}$ is a set of leading eigenvectors of Σ. In what follows, we use n to index the size of the problem. So

the dimension $p = p(n)$ and the spikes $\lambda_j^2 = \lambda_j^2(n)$ can be regarded as functions of n, while both $\bar{m}$ and m remain fixed as n grows.

Let $p_n = p \vee n$. We obtain the initial matrix $\widehat{Q}^{(0)}$ in Algorithm 1 by applying Algorithm 2 with

$$\alpha_n = \alpha\left[\frac{\log(p_n)}{n}\right]^{1/2}. \tag{3.2}$$

In Algorithm 1, the threshold levels are set at

$$\gamma_{nj} = \gamma\left[\ell_j^B \frac{\log(p_n)}{n}\right]^{1/2}, \qquad j = 1, \ldots, m. \tag{3.3}$$

Here, α and γ are user specified constants, and $\ell_j^B = \ell_j(S_{BB}) \vee 1$ with $\ell_j(S_{BB})$ the jth largest eigenvalue of S_{BB}, where the set B is obtained in step 1 of Algorithm 2. For theoretical study, we always stop Algorithm 1 after K_s iterations, where for $h(x) = x^2/(x+1)$,

$$K_s = \frac{1.1 \cdot \ell_1^B}{\ell_m^B - \ell_{m+1}^B}\left[\left(1 + \frac{1}{\log 2}\right)\log n + 0 \vee \log h\big(\ell_1^B - 1\big)\right]. \tag{3.4}$$

Under the conditions of Theorem 3.1 or of Theorems 3.2 and 3.3, or when m is defined by (3.15), we have $\ell_m^B \neq \ell_{m+1}^B$ and $K_s < \infty$ with probability 1.

3.1. *A special case.* To facilitate understanding, we first state the convergence rates for principal subspace estimation in a special case.

Consider the asymptotic setting where $n \to \infty$ with $p \geq n$ and $(\log p)/n \to 0$, while the spikes $\lambda_1^2 \geq \cdots \geq \lambda_m^2 > \lambda_{m+1}^2 \geq \cdots \geq \lambda_{\bar{m}}^2 > 0$ remain unchanged. Suppose that the q_j's are sparse in the sense that, for some $r \in (0, 2)$, the ℓ_r norm of the eigenvectors are uniformly bounded by s, that is, $\|q_j\|_r = (\sum_{v=1}^p |q_{vj}|^r)^{1/r} \leq s$, for $j = 1, \ldots, \bar{m}$, where $s \geq 1$ is an absolute constant.

Recall that $h(x) = x^2/(x+1)$. Under the above setup, we have the following upper bound for subspace estimation error.

THEOREM 3.1. *Under the above setup, for sufficiently large constants $\alpha, \gamma > 2\sqrt{3}$ in (3.2) and (3.3), there exist constants C_0, $C_1 = C_1(\gamma, r, m)$ and C_2, such that for sufficiently large n, uniformly over all Σ with $\|q_j\|_r \leq s$ for $1 \leq j \leq \bar{m}$, with probability at least $1 - C_0 p_n^{-2}$, we have $K_s \asymp \log n$ and the subspace estimator $\widehat{\mathcal{P}}_m^{(K_s)} = \operatorname{ran}(\widehat{Q}^{(K_s)})$ of Algorithm 1 satisfies*

$$L\big(\mathcal{P}_m, \widehat{\mathcal{P}}_m^{(K_s)}\big) \leq C_1 \bar{m} s^r \left[\frac{\log p}{n h(\lambda_m^2)}\right]^{1-r/2} + C_2 g_m(\lambda)\frac{\log p}{n},$$

where $g_m(\lambda) = \frac{(\lambda_1^2+1)(\lambda_{m+1}^2+1)}{(\lambda_m^2 - \lambda_{m+1}^2)^2}$.

The upper bound in Theorem 3.1 consists of two terms. The first is a "nonparametric" term, which can be decomposed as the product of two components. The first component, $\bar{m}s^r[nh(\lambda_m^2)/\log p]^{r/2}$, up to a multiplicative constant, bounds the number of coordinates used in estimating the subspace, while the second component, $\log p/[nh(\lambda_m^2)]$, gives the average error per coordinate. The second term in the upper bound, $g_m(\lambda)(\log p)/n$, up to a logarithmic factor, has the same form as the cross-variance term in the "fixed p, large n" asymptotic limit for classical PCA; cf. [2], Theorem 1. We call it a "parametric" error term, because it always arises when we try to separate the first m eigenvectors from the rest, regardless of how sparse they are. Under the current setup, both terms converge to 0 as $n \to \infty$, which establishes the consistency of our estimator.

To better understand the upper bound, we compare it with an existing lower bound. Suppose $\lambda_1^2 > \lambda_2^2$. Consider the simplest case where $m = 1$. Then, estimating $\mathcal{P}_1$ is the same as estimating the first eigenvector q_1. For estimating an individual eigenvector q_j, Paul and Johnstone [22] considered the loss function $l(q_j, \tilde{q}_j) = \|q_j - \operatorname{sgn}(q_j'\tilde{q}_j)\tilde{q}_j\|_2^2$. Here, the λ_j^2's, s and r are fixed and $p \geq n$, so when n is large, $s^r[nh(\lambda_1^2)/\log p]^{r/2} \leq Cp^{1-c}$ for some $c \in (0, 1)$. For this case, Theorem 2 in [22] asserts that for any estimator $\widehat{q}_1$,

$$\sup_{\|q_j\|_r \leq s, \forall j} \mathbb{E} l(q_1, \widehat{q}_1) \geq C_1 s^r \left[\frac{\log p}{nh(\lambda_1^2)}\right]^{1-r/2} + C_2 \frac{g_1(\lambda)}{n}.$$

Let $\widehat{\mathcal{P}}_1 = \operatorname{span}\{\widehat{q}_1\}$. We have $\frac{1}{2} l(q_1, \widehat{q}_1) \leq L(\mathcal{P}_1, \widehat{\mathcal{P}}_1) \leq l(q_1, \widehat{q}_1)$. So the above lower bound also holds for any $\widehat{\mathcal{P}}_1$ and $\mathbb{E}L(\mathcal{P}_1, \widehat{\mathcal{P}}_1)$. Note that in both Theorem 3.1 and the last display, the nonparametric term is dominant, and so both the lower and upper bounds are of order $[(\log p)/n]^{1-r/2}$. Therefore, Theorem 3.1 shows that the estimator from Algorithm 1 is rate optimal.

Since α_n and γ_{nj} and the stopping rule (3.4) do not involve any unknown parameter, the theorem establishes the adaptivity of our estimator: the optimal rate of convergence in Theorem 3.1 is obtained without any knowledge of the power r, the radius s or the spikes λ_j^2. Last but not least, the estimator could be obtained in $O(\log n)$ iterations and holds for all thresholding function η satisfying (2.2).

Later in Section 3.4, Theorem 3.2 establishes analogous convergence rates, but for a much wider range of high-dimensional sparse settings. In particular, the above result will be extended simultaneously along two directions:

(1) the spikes $\lambda_1^2, \ldots, \lambda_{\bar{m}}^2$ will be allowed to scale as $n \to \infty$, and $\lambda_{m+1}^2, \ldots, \lambda_{\bar{m}}^2$ could even be of smaller order as compared to the first m spikes;

(2) each individual eigenvector q_j will be constrained to a weak-ℓ_r ball of radius s_j (which contains the ℓ_r ball of the same radius), and the radii s_j's will be allowed to diverge as $n \to \infty$.

3.2. *Assumptions*. We now state assumptions for the general theoretical results in Section 3.4.

As outlined above, the first extension of the special case is to allow the spikes $\lambda_j^2 = \lambda_j^2(n) > 0$ to change with n, though the dependence will usually not be shown explicitly. Recall that $p_n = p \vee n$; we impose the following growth rate condition on p and the λ_j^2's.

CONDITION GR. *As* $n \to \infty$, *we have*:

(1) *the dimension* p *satisfies* $(\log p)/n = o(1)$;

(2) *the largest spike* λ_1^2 *satisfies* $\lambda_1^2 = O(p_n)$; *the smallest spike* $\lambda_{\bar{m}}^2$ *satisfies* $\log(p_n) = o(n\lambda_{\bar{m}}^4)$; *and their ratio satisfies* $\lambda_1^2/\lambda_{\bar{m}}^2 = O(n[\log(p_n)/n]^{1/2+r/4})$;

(3) $\lim_{n\to\infty} \lambda_1^2/(\lambda_j^2 - \lambda_{j+1}^2) \in [1, \infty]$ *exists for* $j = 1, \ldots, \bar{m}$, *with* $\lambda_{\bar{m}+1}^2 = 0$.

The first part of Condition GR requires the dimension to grow at a subexponential rate of the sample size. The second part ensures that the spikes grow at most at linear rate with p_n, and are all of larger magnitude than $\sqrt{\log(p_n)/n}$. In addition, the condition on the ratio $\lambda_1^2/\lambda_{\bar{m}}^2$ allows us to deal with the interesting cases where the first several spikes scale at a faster rate with n than the others. This is more flexible than the assumption previously made in [22] that all the spikes grow at the same rate. The third part requires $\lim_{n\to\infty} \lambda_1^2/(\lambda_j^2 - \lambda_{j+1}^2)$ to exist for each $1 \le j \le \bar{m}$, but the limit can be infinity.

Turn to the sparsity assumption on the q_j's. We first make a mild extension from ℓ_r ball to weak-ℓ_r ball [5]. To this end, for any p-vector u, order its coordinates by magnitude as $|u|_{(1)} \ge \cdots \ge |u|_{(p)}$. We say that u belongs to the weak-ℓ_r ball of radius s, denoted by $u \in w\ell_r(s)$, if

$$|u|_{(\nu)} \le s\nu^{-1/r} \qquad \text{for all } \nu. \tag{3.5}$$

For $r \in (0, 2)$, the above condition implies rapid decay of the ordered coefficients of u, and thus describes its sparsity. For instance, consider $u = (1/\sqrt{k}, \ldots, 1/\sqrt{k}, 0, \ldots, 0)'$ with exactly k nonzero entries all equal to $1/\sqrt{k}$. Then, for fixed $r \in (0, 2)$, we have $u \in w\ell_r(k^{1/r-1/2})$. In particular, when $k = 1$, $u \in w\ell_r(1)$. Note that weak-ℓ_r ball extends ℓ_r ball, because $\|u\|_r \le s$, that is, $u \in \ell_r(s)$, implies $u \in w\ell_r(s)$.

In what follows, we assume that for some fixed $r \in (0, 2)$ and all $j \le \bar{m}$, $q_j \in w\ell_r(s_j)$ for some $s_j > 1$. We choose to use the notion of "weak-ℓ_r decay," because it provides a unified framework for several different notions of sparsity, which is convenient for analyzing a statistical estimation problem from a minimax point of view [5]. Hence, at any fixed n, we will consider whether Algorithm 1 performs uniformly well on n i.i.d. observations x_i generated by (3.1) whose covariance matrix Σ belongs to the following uniformity class:

$$\mathcal{F}_n = \left\{ \Sigma_{p\times p} = \sum_{j=1}^{\bar{m}} \lambda_j^2 q_j q_j' + I : q_j \in w\ell_r(s_j), \forall j \right\}.$$

For general results, we allow the radii s_j's to depend on or even diverge with n, though we require that they do not grow too rapidly, so the leading eigenvectors are indeed sparse. This leads to the following sparsity condition.

CONDITION SP. *As $n \to \infty$, the radius s_j of the weak-ℓ_r ball satisfies $s_j \geq 1$ and*

$$s_j^r \left[\frac{\log(p_n)}{n\lambda_j^4} \right]^{1/2-r/4} = o(1 \wedge \lambda_1^4) \qquad \text{for } j = 1, \ldots, \bar{m}.$$

This type of condition also appeared in a previous study of individual eigenvector estimation in the multiple component spiked covariance model [22]. The condition is, for example, satisfied if Condition GR holds and the largest spike λ_1^2 is bounded away from zero while the radii s_j's are all bounded above by an arbitrarily large constant. That is, if there exists a constant $C > 0$, such that $\lambda_1^2 \geq 1/C$ and $s_j \leq C$ for all $j \leq \bar{m}$ and all n.

It is straightforward to verify that Conditions GR and SP are satisfied by the special case in Section 3.1. We conclude this part with an example.

EXAMPLE. When each x_i collects noisy measurements of an underlying random function on a regular grid, model (3.1) becomes discretization of a functional PCA model [24], and the q_j's are discretized eigenfunctions. When the eigenfunctions are smooth or have isolated singularities either in themselves or in their derivatives, their wavelet coefficients belong to some weak ℓ_r ball [5]. So do the discrete wavelet transform of the q_j's. Moreover, the radii of the weak ℓ_r balls are determined by the underlying eigenfunctions and are thus uniformly bounded as the size of the grid p gets larger. In this case, Condition SP is satisfied when Condition GR holds and λ_1^2 is bounded away from zero. So, for functional data of this type, we could always first transform to the wavelet domain and then apply Algorithm 1.

3.3. *Key quantities.* We now introduce a few key quantities which appear later in the general theoretical results.

The first quantity gives the rate at which we distinguish high from low signal coordinates. Recall that $h(x) = x^2/(x+1)$. For $j = 1, \ldots, \bar{m}$, define

$$\tau_{nj} = \sqrt{\frac{\log(p_n)}{nh(\lambda_j^2)}}. \tag{3.6}$$

According to [20], up to a logarithmic factor, τ_{nj}^2 can be interpreted as the average error per coordinate in estimating an eigenvector with eigenvalue $\lambda_j^2 + 1$. Thus, a coordinate can be regarded as of high signal if at least one of the leading eigenvectors is of larger magnitude on this coordinate compared to τ_{nj}. Otherwise,

we call it a low signal coordinate. We define $H(\beta)$ to be the set of high signal coordinates

$$H = H(\beta) = \{v : |q_{vj}| \geq \beta \tau_{nj}, \text{ for some } 1 \leq j \leq \bar{m}\}. \tag{3.7}$$

Here, β is a constant not depending on n, the actual value of which will be specified in Theorem 3.2. If $\bar{m} = 1$ and q_1 has k nonzero entries all equal to $1/\sqrt{k}$, then H contains exactly these k coordinates when $k < nh(\lambda_1^2)/[\beta^2 \log(p_n)]$, which is guaranteed under Condition SP. In addition, let $L = \{1, \ldots, p\} \setminus H$ be the complement of H. Here, H stands for "high," and L for "low" (also recall B in Algorithm 2, where B stands for "big"). The dependence of H, L and B on n is suppressed for notational convenience.

To understand the convergence rate of the subspace estimator stated later in (3.11), it is important to have an upper bound for $\text{card}(H)$, the cardinality of H. To this end, define

$$M_n = p \wedge \sum_{j=1}^{\bar{m}} \frac{s_j^r}{\tau_{nj}^r}. \tag{3.8}$$

The following lemma shows that a constant multiple of M_n bounds $\text{card}(H)$. The proof of the lemma is given in [15]. Thus, in the general result, M_n plays the same role as the term $\bar{m} s^r [nh(\lambda^2)/\log p]^{r/2}$ has played in Theorem 3.1.

LEMMA 3.1. *For sufficiently large n, the cardinality of $H = H(\beta)$ satisfies $\bar{m} \leq \text{card}(H) \leq C M_n$ for a constant C depending on β and r.*

The last quantity we introduce is related to the "parametric" term in the convergence rate. Let $\lambda_{\bar{m}+1}^2 = 0$. For $j = 1, \ldots, \bar{m}$, define

$$\varepsilon_{nj}^2 = \frac{(\lambda_1^2 + 1)(\lambda_{j+1}^2 + 1)}{(\lambda_j^2 - \lambda_{j+1}^2)^2} \frac{\log(p_n)}{n}. \tag{3.9}$$

So the second term of the upper bound in Theorem 3.1 is $C_2 \varepsilon_{nm}^2$. For the interpretation of this quantity, we refer to the discussion after Theorem 3.1.

3.4. *Main results.* We turn to the statement of main theoretical results.

A key condition for the results is the asymptotic distinguishability (AD) condition introduced below. Recall that all the spikes λ_j^2 (hence all the leading eigenvalues) are allowed to depend on n. The condition AD will guarantee that the largest few eigenvalues are asymptotically well separated from the rest of the spectrum, and so the corresponding principal subspace is distinguishable.

DEFINITION. We say that *condition* $\text{AD}(j, \kappa)$ *is satisfied with constant* κ, if there exists a numeric constant $\kappa \geq 1$, such that for sufficiently large n, the gap between the jth and the $(j+1)$th eigenvalues satisfies

$$\lambda_j^2 - \lambda_{j+1}^2 \geq \lambda_1^2/\kappa.$$

We define $\mathrm{AD}(0,\kappa)$ and $\mathrm{AD}(\bar{m},\kappa)$ by letting $\lambda_0^2=\infty$, and $\lambda_{\bar{m}+1}^2=0$. So $\mathrm{AD}(0,\kappa)$ holds for any $\kappa\geq 1$. Note that there is always some $1\leq j\leq \bar{m}$ such that condition $\mathrm{AD}(j,\kappa)$ is satisfied. For instance, $\mathrm{AD}(j,\kappa)$ is satisfied with some κ for the largest j such that $\lambda_j^2\asymp\lambda_1^2$. When the spikes do not change with n, condition $\mathrm{AD}(\bar{m},\kappa)$ is satisfied with any constant $\kappa\geq\lambda_1^2/\lambda_{\bar{m}}^2$.

Rates of convergence for principal subspace estimation. Recall definitions (3.2)–(3.4) and (3.6)–(3.9). The following theorem establishes the rate of convergence of the principal subspace estimator obtained via Algorithm 1 under relaxed assumptions, which generalizes Theorem 3.1.

THEOREM 3.2. *Suppose Conditions* GR *and* SP *hold, and condition* $\mathrm{AD}(m,\kappa)$ *is satisfied with some constant* $\kappa\geq 1$ *for the given subspace dimension* m. *Let the constants* $\alpha,\gamma>2\sqrt{3}$ *in* (3.2) *and* (3.3), *and for* $c=0.9(\gamma-2\sqrt{3})$, *let* $\beta=c/\sqrt{m}$ *in* H (3.7). *Then, there exist constants* C_0, $C_1=C_1(\gamma,r,m,\kappa)$ *and* C_2, *such that for sufficiently large* n, *uniformly over* $\mathcal{F}_n$, *with probability at least* $1-C_0p_n^{-2}$, $K_s\in[K,2K]$ *for*

$$K=\frac{\lambda_1^2+1}{\lambda_m^2-\lambda_{m+1}^2}\left[\left(1+\frac{1}{\log 2}\right)\log n+0\vee\log h\big(\lambda_1^2\big)\right], \tag{3.10}$$

and the subspace estimator $\widehat{\mathcal{P}}_m^{(K_s)}=\mathrm{ran}(\widehat{Q}^{(K_s)})$ *satisfies*

$$L\big(\mathcal{P}_m,\widehat{\mathcal{P}}_m^{(K_s)}\big)\leq C_1M_n\tau_{nm}^2+C_2\varepsilon_{nm}^2=o(1). \tag{3.11}$$

Theorem 3.2 states that for appropriately chosen threshold levels and all thresholding function satisfying (2.2), after enough iterations, Algorithm 1 yields principal subspace estimators whose errors are, with high probability, uniformly bounded over $\mathcal{F}_n$ by a sequence of asymptotically vanishing constants as $n\to\infty$. In addition, the probability that the estimation error is not well controlled vanishes polynomially fast. Therefore, the subspace estimators are uniformly consistent over $\mathcal{F}_n$.

The interpretation of the two terms in the error bound (3.11) is similar to those in Theorem 3.1. Having introduced those quantities in Section 3.3, we could elaborate a little more on the first, that is, the "nonparametric" term. By Theorem 3.3 below, when estimating $\mathcal{P}_m$, Algorithm 1 focuses only on the coordinates in H, whose cardinality is $\mathrm{card}(H)=O(M_n)$. Though H does not appear explicitly in the rates, the rates depend crucially on its cardinality which is further upper bounded by M_n. Since τ_{nm}^2 can be interpreted as the average error per coordinate, the total estimation error accumulated over all coordinates in H is thus of order $O(M_n\tau_{nm}^2)$. Moreover, as we will show later, the squared bias induced by focusing only on H is also of order $O(M_n\tau_{nm}^2)$. Thus, this term indeed comes from the bias-variance tradeoff of the nonparametric estimation procedure. The meaning of the

second, that is, the "parametric," term is the same as in Theorem 3.1. Finally, we note that both terms vanish as $n \to \infty$ under Conditions GR, SP and AD(m, κ).

The threshold levels α_n and γ_{nj} in (3.2) and (3.3) as well as K_s in (3.4) do not depend on unknown parameters. So the estimation procedure achieves the rates adaptively over a wide range of high-dimensional sparse settings.

In addition, (3.10) implies that Algorithm 1 only needs a relatively small number of iterations to yield the desired estimator. In particular, when the largest spike λ_1^2 is bounded away from zero, (3.10) shows that it suffices to have $K_s \asymp \log n$ iterations. We remark that it is not critical to run precisely K_s iterations. The result holds when we stop anywhere between K and $2K$.

Theorem 3.2 could also be extended to an upper bound for the risk. Note that $p_n^{-2} = o(\tau_{nm}^2 \vee \varepsilon_{nm}^2)$, and that the loss function (2.1) is always bounded above by 1. The following result is a direct consequence of Theorem 3.2.

COROLLARY 3.1. *Under the setup of Theorem 3.2, we have*

$$\sup_{\mathcal{F}_n} \mathbb{E} L\big(\mathcal{P}_m, \widehat{\mathcal{P}}_m^{(K_s)}\big) \le C_1 M_n \tau_{nm}^2 + C_2 \varepsilon_{nm}^2.$$

Correct exclusion property. We now switch to the model selection property of Algorithm 1. By the discussion in Section 2, an important motivation for the iterative thresholding procedure is to trade bias for variance by keeping low signal coordinates out of the orthogonal iterations. More specifically, it is desirable to restrict our effort to estimating those coordinates in H and simply estimating those coordinates in L with zeros.

By construction, Algorithm 2 yields an initial matrix with a lot of zeros, but Algorithm 1 is at liberty to introduce new nonzero coordinates. The following result shows that with high probability all the nonzero coordinates introduced are in the set H.

THEOREM 3.3. *Under the setup of Theorem 3.2, uniformly over $\mathcal{F}_n$, with probability at least $1 - C_0 p_n^{-2}$, for all $k = 0, \ldots, K_s$, the orthonormal matrix $\widehat{Q}^{(k)}$ has zeros in all its rows indexed by L, that is, $\widehat{Q}_{L\cdot}^{(k)} = 0$.*

We call the property in Theorem 3.3 "correct exclusion," because it ensures that all the low signal coordinates in L are correctly excluded from iterations. In addition, Theorem 3.3 shows that the principal subspace estimator is indeed spanned by a set of sparse loading vectors, where all loadings in L are exactly zero.

Note that the initial matrix $\widehat{Q}^{(0)}$ has all its nonzero coordinates in B, which, with high probability, only selects "big" coefficients in the leading eigenvectors, whose magnitudes are no less than $O([\log p_n/(n\lambda_m^4)]^{1/4})$. On the other hand, the set H includes all coordinates with magnitude no less than $O([\log p_n/(nh(\lambda_m^2))]^{1/2})$.

Thus, the minimum signal strength for H is of smaller order than that for B. So, with high probability, B is a subset of H consisting only of its coordinates with "big" signals. Thus, though $\widehat{Q}^{(0)}$ excludes all the coordinates in L, it only includes "big" coordinates in H and fails to pick those medium sized ones which are crucial for obtaining the convergence rate (3.11). Algorithm 1 helps to include more coordinates in H along iterations and hence achieves (3.11).

Rates of convergence for individual eigenvector estimation. The primary focus of this paper is on estimating principal subspaces. However, when an individual eigenvector, say q_j, is identifiable, it is also of interest to see whether Algorithm 1 can estimate it well. The following result shows that for K_s in (3.4), the jth column of $\widehat{Q}^{(K_s)}$ estimates q_j well, provided that the jth eigenvalue is well separated from the rest of the spectrum.

COROLLARY 3.2. *Under the setup of Theorem 3.2, suppose for some* $j \le m$, *both conditions* $\mathrm{AD}(j-1,\kappa')$ *and* $\mathrm{AD}(j,\kappa')$ *are satisfied for some constant* $\kappa' < \lim_{n\to\infty} \lambda_1^2/(\lambda_m^2 - \lambda_{m+1}^2)$. *Then uniformly over* $\mathcal{F}_n$, *with probability at least* $1 - C_0 p_n^{-2}$, $\widehat{q}_j^{(K_s)}$, *the* j*th column of* $\widehat{Q}^{(K_s)}$, *satisfies*

$$L\big(\mathrm{span}\{q_j\}, \mathrm{span}\{\widehat{q}_j^{(K_s)}\}\big) \le C_1 M_n \tau_{nj}^2 + C_2\big(\varepsilon_{n,j-1}^2 \vee \varepsilon_{nj}^2\big).$$

Moreover, $\sup_{\mathcal{F}_n} \mathbb{E}L(\mathrm{span}\{q_j\}, \mathrm{span}\{\widehat{q}_j^{(K_s)}\})$, *the supremum risk over* $\mathcal{F}_n$, *is also bounded by the right-hand side of the above inequality.*

Corollary 3.2 connects closely to the previous investigation [22] on estimating individual sparse leading eigenvectors. Recall their loss function $l(q_j, \tilde{q}_j) = \|q_j - \mathrm{sgn}(q_j'\tilde{q}_j)\tilde{q}_j\|_2^2$. Since $\frac{1}{2}l(q_j,\tilde{q}_j) \le L(\mathrm{span}\{q_j\}, \mathrm{span}\{\tilde{q}_j\}) \le l(q_j,\tilde{q}_j)$, l is equivalent to the restriction of the loss function (2.1) to one-dimensional subspaces. Thus, Corollary 3.2 implies that

$$\sup_{\mathcal{F}_n} \mathbb{E}l\big(q_j, \widehat{q}_j^{(K_s)}\big) \le C_1 M_n \tau_{nj}^2 + C_2 \frac{(\lambda_1^2+1)(\lambda_j^2+1)}{[(\lambda_{j-1}^2 - \lambda_j^2) \wedge (\lambda_j^2 - \lambda_{j+1}^2)]^2} \frac{\log(p_n)}{n}.$$

When the radii of the weak-ℓ_r balls grow at the same rate, that is, $\max_j s_j \asymp \min_j s_j$, the upper bound in the last display matches the lower bound in Theorem 2 of [22] up to a logarithmic factor. Thus, when the jth eigenvalue is well separated from the rest of the spectrum, Algorithm 1 yields a near optimal estimator of q_j in the adaptive rate minimax sense.

3.5. *Choice of* m. The main results in this section are stated with the assumption that the subspace dimension m is given. In what follows, we discuss how to choose m and also how to estimate $\bar{m}$ based on data.

Recall ℓ_j^B defined after (3.3) and the set B in step 1 of Algorithm 2. Let

$$\widehat{\bar{m}} = \max\{j : \ell_j^B > 1 + \delta_{\mathrm{card}(B)}\} \tag{3.12}$$

be an estimator for $\bar{m}$, where for any positive integer k,

$$\delta_k = 2(\sqrt{k/n} + t_k) + (\sqrt{k/n} + t_k)^2 \tag{3.13}$$

with

$$t_k^2 = \frac{6\log p_n}{n} + \frac{2k(\log p_n + 1)}{n}. \tag{3.14}$$

Then in Algorithm 1, for a large constant $\bar{\kappa}$, we define

$$m = \max\left\{ j : 1 \le j \le \widehat{\bar{m}} \text{ and } \frac{\ell_1^B - 1}{\ell_j^B - \ell_{j+1}^B} \le \bar{\kappa} \right\}. \tag{3.15}$$

Setting $\bar{\kappa} = 15$ works well in simulation. For a given dataset, such a choice of m is intended to lead us to estimate the largest principal subspace such that its eigenvalues maintain a considerable gap from the rest of the spectrum. Note that (3.15) can be readily incorporated into Algorithm 2: we could compute the eigenvalues ℓ_j^B of S_{BB} in step 2, and then obtain $\widehat{\bar{m}}$ and m.

For $\widehat{\bar{m}}$ in (3.12), we have the following results.

PROPOSITION 3.1. *Suppose Conditions GR and SP hold. Let $\widehat{\bar{m}}$ be defined in (3.12) with B obtained by Algorithm 2 with α_n specified by (3.2) for some $\alpha > 2\sqrt{3}$. Then, uniformly over $\mathcal{F}_n$, with probability at least $1 - C_0 p_n^{-2}$:*

(1) $\widehat{\bar{m}} \le \bar{m}$;

(2) *for any m such that the condition $\mathrm{AD}(m, \kappa)$ is satisfied with some constant κ, $m \le \widehat{\bar{m}}$ when n is sufficiently large*;

(3) *if the condition $\mathrm{AD}(\bar{m}, \kappa)$ is satisfied with some constant κ, $\widehat{\bar{m}} = \bar{m}$ when n is sufficiently large.*

By claim (1), any $m \le \widehat{\bar{m}}$ satisfies $m \le \bar{m}$ with high probability. In addition, claim (2) shows that, for sufficiently large n, any m such that $\mathrm{AD}(m, \kappa)$ holds is no greater than $\widehat{\bar{m}}$. Thus, when restricting to those $m \le \widehat{\bar{m}}$, we do not miss any m such that Theorems 3.2 and 3.3 hold for estimating $\mathcal{P}_m$. These two claims jointly ensure that we do not need to consider any target dimension beyond $\widehat{\bar{m}}$. Finally, claim (3) shows that we recover the exact number of spikes with high probability for large samples when $\mathrm{AD}(\bar{m}, \kappa)$ is satisfied, that is, when $\lambda_1^2 \asymp \lambda_{\bar{m}}^2$. Note that this assumption was made in [22].

Turn to the justification of (3.15). We show later in Corollary 6.1 that for $1 \le j \le \bar{m}$, $(\ell_1^B - 1)/(\ell_j^B - \ell_{j+1}^B)$ estimates $\lambda_1^2/(\lambda_j^2 - \lambda_{j+1}^2)$ consistently under Conditions GR and SP. [It is important that the condition $\mathrm{AD}(m, \kappa)$ is not needed

for this result!] This implies that for m in (3.15), we have $\lambda_1^2/(\lambda_m^2 - \lambda_{m+1}^2) \le 1.1\bar{\kappa}$ when n is sufficiently large. Hence, the condition AD(m, κ) is satisfied with the constant $\kappa = 1.1\bar{\kappa}$. Therefore, the main theoretical results in Section 3.4 remain valid when we set m by (3.15) in Algorithm 1.

4. Computational complexity. We now study the computational complexity of Algorithm 1. Throughout, we assume the same setup as in Section 3, and restrict the calculation to the high probability event on which the conclusions of Theorems 3.2 and 3.3 hold. For any matrix A, we use supp$\{A\}$ to denote the index set of the nonzero rows of A.

Consider a single iteration, say, the kth. In the multiplication step, the (ν, j)th element of $T^{(k)}$, $t_{\nu j}^{(k)}$, comes from the inner product of the νth row of S and the jth column of $\widehat{Q}^{(k-1)}$. Though both are p-vectors, Theorem 3.3 asserts that for any column of $\widehat{Q}^{(k-1)}$, at most card(H) of its entries are nonzero. So if we know supp$\{\widehat{Q}^{(k-1)}\}$, then $t_{\nu j}^{(k)}$ can be calculated in $O(\text{card}(H))$ flops, and $T^{(k)}$ in $O(mp\,\text{card}(H))$ flops. Since supp$\{\widehat{Q}^{(k-1)}\}$ can be obtained in $O(mp)$ flops, the multiplication step can be completed in $O(mp\,\text{card}(H))$ flops. Next, the thresholding step performs elementwise operation on $T^{(k)}$, and hence can be completed in $O(mp)$ flops. Turn to the QR step. First, we can obtain supp$\{\widehat{T}^{(k)}\}$ in $O(mp)$ flops. Then QR factorization can be performed on the reduced matrix which only includes the rows in supp$\{\widehat{T}^{(k)}\}$. Since Theorem 3.3 implies supp$\{\widehat{T}^{(k)}\} =$ supp$\{\widehat{Q}^{(k)}\} \subset H$, the complexity of this step is $O(m^2\,\text{card}(H))$. Since $m = O(p)$, the complexity of the multiplication step dominates, and so the complexity of each iteration is $O(mp\,\text{card}(H))$. Theorem 3.2 shows that K_s iteration is enough. Therefore, the overall complexity of Algorithm 1 is $O(K_s mp\,\text{card}(H))$.

When the true eigenvectors are sparse, card(H) is of manageable size. In many realistic situations, λ_1^2 is bounded away from 0 and so $K_s \asymp \log n$. For these cases, Algorithm 1 is scalable to very high dimensions.

We conclude the section with a brief discussion on parallel implementation of Algorithm 1. In the kth iteration, both matrix multiplication and elementwise thresholding can be computed in parallel. For QR factorization, one needs only to communicate the rows of $\widehat{T}^{(k)}$ with nonzero elements, the number of which is no greater than card(H). Thus, the overhead from communication is $O(m\,\text{card}(H))$ for each iteration, and $O(K_s m\,\text{card}(H))$ in total. When the leading eigenvectors are sparse, card(H) is manageable, and parallel computing of Algorithm 1 is feasible.

5. Numerical experiments.

5.1. *Single spike settings.* We first consider the case where each x_i is generated by (3.1) with $\bar{m} = 1$. Motivated by functional data with localized features, four test vectors q_1 are considered, where $q_1 = (f(1/p), \ldots, f(p/p))'$, with f one of

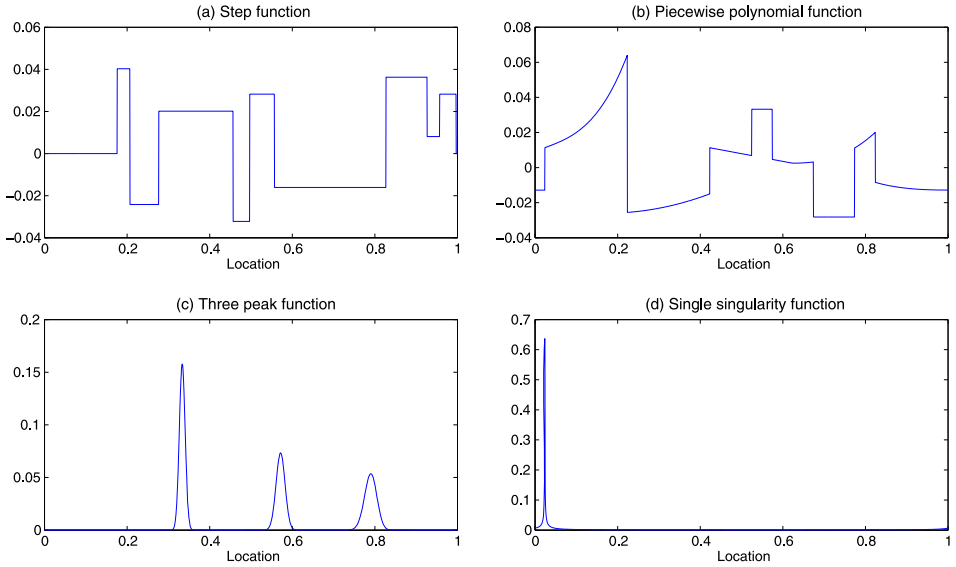

FIG. 1. *Four test vectors in the original domain*: *values at* $p = 2048$ *equispaced points on* $[0, 1]$ *of four test functions*. (a) `step`: *step function*, (b) `poly`: *piecewise polynomial function*, (c) `peak`: *three-peak function and* (d) `sing`: *single singularity function*.

the four functions in Figure 1. For each test vector, the dimension $p = 2048$, the sample size $n = 1024$ and λ_1^2 ranges in $\{100, 25, 10, 5, 2\}$.

Before applying any sparse PCA method, we transform the observed data vectors into the wavelet domain using the Symmlet 8 basis [16], and scale all the observations by $\widehat{\sigma}$ with $\widehat{\sigma}^2$ given in (2.3). The multi-resolution plots of wavelet coefficients of the test vectors are shown in Figure 2. In the wavelet domain, the four vectors exhibits different levels of sparsity, with `step` the least sparse, and `sing` the most.

Table 1 compares the average loss of subspace estimation over 100 runs for each spike value and each test vector by Algorithm 1 (ITSPCA) with several existing methods: augmented sparse PCA (AUGSPCA) [22], correlation augmented sparse PCA (CORSPCA) [18] and diagonal thresholding sparse PCA (DTSPCA) given in Algorithm 2. For ITSPCA, we computed $\widehat{Q}^{(0)}$ by Algorithm 2. α_n and γ_{n1} are specified by (3.2) and (3.3) with $\alpha = 3$ and $\gamma = 1.5$. These values are smaller than those in theoretical results, but lead to better numerical performance. We stop iterating once $L(\mathrm{ran}(\widehat{Q}^{(k)}), \mathrm{ran}(\widehat{Q}^{(k+1)})) \le n^{-2}$. Parameters in competing algorithms are all set to the values recommended by their authors.

From Table 1, ITSPCA and CORSPCA outperform the other two methods in all settings. Between the two, CORSPCA only wins by small margins when the spike values are large. Otherwise, ITSPCA wins, sometimes with large margins. For the same algorithm at the same spike value, the sparser the signal, the smaller the estimation error.

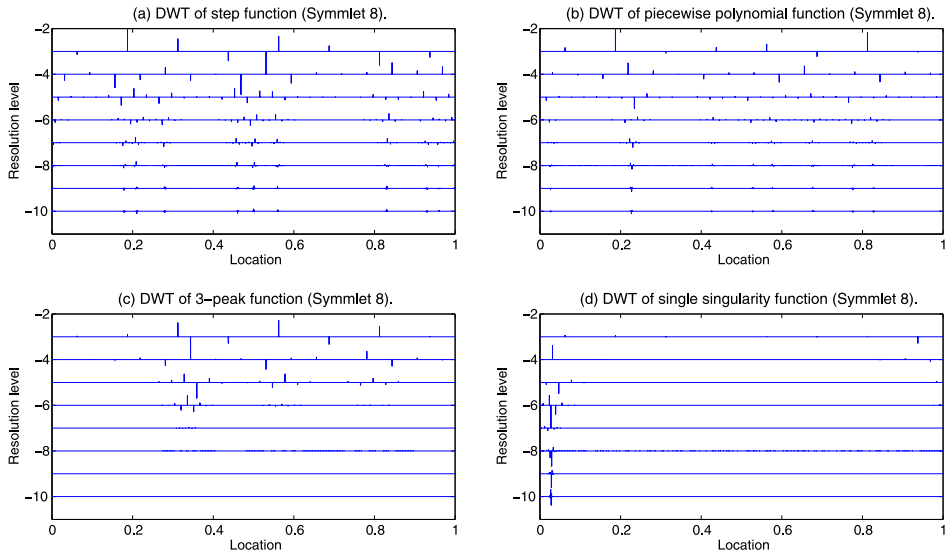

FIG. 2. *Discrete wavelet transform of the four test vectors in Figure 1. In each plot, the length of each stick is proportional to the magnitude of the* `Symmlet 8` *wavelet coefficient at the given location and resolution level.*

Table 1 also presents the average sizes of the sets of selected coordinates. While all methods yield sparse PC loadings, AUGSPCA and DTSPCA seem to select too few coordinates, and thus introduce too much bias. ITSPCA and CORSPCA apparently result in a better bias-variance tradeoff.

5.2. *Multiple spike settings.* Next, we simulated data vectors using model (3.1) with $\bar{m} = 4$. The q_j vectors are taken to be the four test vectors used in single spike settings, in the same order as in Figure 1, up to orthonormalization.[1] We tried four different configurations of the spike values $(\lambda_1^2, \ldots, \lambda_4^2)$, as specified in the first column of Table 2. For each configuration of spike values, the dimension is $p = 2048$, and the sample size is $n = 1024$.

For each simulated dataset, we estimate $\mathcal{P}_m$ for $m = 1, 2, 3$ and 4. The last four columns of Table 2 present the losses in estimating subspaces, averaged over 100 runs, using the same sparse PCA methods as in single spike settings. For ITSPCA, we set the thresholds $\{\gamma_{nj}, j = 1, \ldots, 4\}$ as in (3.3) with $\gamma = 1.5$. All other implementation details are the same. Again, we used recommended values for parameters in all other competing methods.

The simulation results reveal two interesting phenomena. First, when the spikes are relatively well separated (the first and the last blocks of Table 2), all methods

[1]The four test vectors are shifted such that the inner product of any pair is close to 0. So the vectors after orthonormalization are visually indistinguishable from those in Figure 1.

TABLE 1
Comparison of sparse PCA methods in single spike settings: average loss in estimation and size of selected feature set

		ITSPCA		AUGSPCA		CORSPCA		DTSPCA	
Test vector	λ_1^2	**Loss**	**Size**	**Loss**	**Size**	**Loss**	**Size**	**Loss**	**Size**
`Step`	100	0.0061	114.2	0.0096	96.5	0.0055	120.1	0.0275	66.6
	25	0.0224	76.3	0.0362	55.4	0.0236	73.9	0.0777	38.3
	10	0.0470	53.4	0.0710	37.4	0.0551	45.9	0.1494	24.1
	5	0.0786	45.5	0.1370	23.7	0.1119	28.7	0.2203	17.1
	2	0.1921	25.4	0.3107	11.4	0.3846	15.2	0.4518	9.7
`Poly`	100	0.0060	83.1	0.0088	66.5	0.0051	92.0	0.0191	49.2
	25	0.0175	52.4	0.0254	41.4	0.0173	53.1	0.0540	28.7
	10	0.0346	38.7	0.0527	27.5	0.0404	34.0	0.0959	20.5
	5	0.0588	30.7	0.0844	20.2	0.0684	24.6	0.1778	14.0
	2	0.1317	20.0	0.2300	10.3	0.2155	16.3	0.3370	8.1
`Peak`	100	0.0019	45.7	0.0032	39.6	0.0016	51.2	0.0075	32.8
	25	0.0071	34.1	0.0099	29.9	0.0069	35.2	0.0226	24.3
	10	0.0158	28.0	0.0222	23.8	0.0165	27.3	0.0592	18.6
	5	0.0283	24.7	0.0449	19.6	0.0320	22.5	0.1161	14.1
	2	0.0927	20.8	0.1887	9.9	0.1176	14.6	0.2702	8.8
`Sing`	100	0.0016	38.0	0.0025	33.2	0.0014	43.6	0.0070	26.3
	25	0.0068	27.1	0.0095	23.1	0.0060	31.8	0.0237	17.5
	10	0.0161	20.3	0.0233	16.6	0.0154	20.9	0.0377	13.6
	5	0.0279	17.3	0.0372	13.2	0.0313	15.2	0.0547	12.7
	2	0.0631	15.2	0.0792	10.9	0.0652	13.0	0.2025	8.8

yield decent estimators of $\mathcal{P}_m$ for all values of m, which implies that the individual eigenvectors are also estimated well. In this case, ITSPCA always outperforms the other three competing methods. Second, when the spikes are not so well separated (the middle two blocks, with $m = 1, 2$ or 3), no method leads to decent subspace estimator. However, all methods give reasonable estimators for $\mathcal{P}_4$ because λ_4^2 in both cases are well above 0. This implies that, under such settings, we fail to recover individual eigenvectors, but we can still estimate $\mathcal{P}_4$ well. ITSPCA again gives the smallest average losses. In all configurations, the estimated number of spikes $\widehat{\widehat{m}}$ in (3.12) and the data-based choice of m in (3.15) with $\bar{\kappa} = 15$ consistently picked $m = \widehat{\widehat{m}} = 4$ in all simulated datasets. Therefore, we are always led to estimating the "right" subspace $\mathcal{P}_4$, and ITSPCA performs favorably over the competing methods.

In summary, simulations under multiple spike settings not only demonstrate the competitiveness of Algorithm 1, but also suggest:

TABLE 2
Comparison of sparse PCA methods in multiple spike settings: *average loss in estimation*

		$L(\mathcal{P}_m, \widehat{\mathcal{P}}_m)$			
$(\lambda_1^2, \lambda_2^2, \lambda_3^2, \lambda_4^2)$	m	**ITSPCA**	**AUGSPCA**	**CORSPCA**	**DTSPCA**
(100, 75, 50, 25)	1	0.0216	0.0260	0.0240	0.0378
	2	0.0180	0.0213	0.0214	0.0308
	3	0.0094	0.0129	0.0126	0.0234
	4	0.0087	0.0122	0.0181	0.0235
(60, 55, 50, 45)	1	0.3100	0.2588	0.2548	0.2831
	2	0.2675	0.2045	0.2095	0.2349
	3	0.1844	0.1878	0.1872	0.1968
	4	0.0157	0.0203	0.0178	0.0333
(30, 27, 25, 22)	1	0.3290	0.2464	0.2495	0.2937
	2	0.3147	0.2655	0.2882	0.3218
	3	0.1740	0.1662	0.1708	0.1821
	4	0.0270	0.0342	0.0338	0.0573
(30, 20, 10, 5)	1	0.0268	0.0392	0.0380	0.0658
	2	0.0237	0.0353	0.0391	0.0605
	3	0.0223	0.0336	0.0372	0.0599
	4	0.0298	0.0414	0.0717	0.0638

(1) The quality of principal subspace estimation depends on the gap between successive eigenvalues, in addition to the sparsity of eigenvectors;

(2) Focusing on individual eigenvectors can be misleading for the purpose of finding low-dimensional projections.

6. Proof. This section is devoted to the proofs of Theorems 3.2 and 3.3. We state the main ideas in Section 6.1 and divide the proof into three major steps, which are then completed in sequel in Sections 6.2–6.4. Others results in Section 3.4 are proved in the supplementary material [15].

6.1. *Main ideas and outline of proof.* The proof is based on an *oracle sequence approach*, the main ideas of which are as follows. First, assuming oracle knowledge of the set H, we construct a sequence of $p \times m$ orthonormal matrices $\{\widehat{Q}^{(k),\mathrm{o}}, k \geq 0\}$. Then we study how fast the sequence converges, and how well each associated column subspace approximates the principal subspace $\mathcal{P}_m$ of interest. Finally, we show that, with high probability, the first K_s terms of the oracle sequence is exactly the sequence $\{\widehat{Q}^{(k)}, 0 \leq k \leq K_s\}$ obtained by Algorithm 1. The actual estimating sequence thus inherits from the oracle sequence various properties in terms of estimation error and number of steps needed to achieve the desired

error rate. The actual sequence mimics the oracle because the thresholding step forces it to only consider the high signal coordinates in H.

In what follows, we first construct the oracle sequence and then lay out a road map of the proof. Here and after, we use an extra superscript "o" to indicate oracle quantities. For example, $\widehat{Q}^{(k),\mathrm{o}}$ denotes the kth orthonormal matrix in the oracle sequence.

Construction of the oracle sequence. First, we construct $\widehat{Q}^{(0),\mathrm{o}}$ using an oracle version of Algorithm 2, where the set B is replaced by its oracle version $B^{\mathrm{o}} = B \cap H$. This ensures that $\widehat{Q}^{(0),\mathrm{o}}_{L\cdot} = 0$.

To construct the rest of the sequence, suppose that the p features are organized (after reordering) in such a way that those in H always have smaller indices than those in L, and that within H, those in B^{o} precede those not. Define the oracle sample covariance matrix

$$S^{\mathrm{o}} = \begin{bmatrix} S_{HH} & 0 \\ 0 & I_{LL} \end{bmatrix}. \tag{6.1}$$

Here, I_{LL} is the identity matrix of dimension $\mathrm{card}(L)$. Then, the matrices $\{\widehat{Q}^{(k),\mathrm{o}}, k \geq 0\}$ are obtained via an oracle version of Algorithm 1, in which the initial matrix is $\widehat{Q}^{(0),\mathrm{o}}$, and S is replaced by S^{o}.

REMARK 6.1. This formal construction does not guarantee that $\widehat{Q}^{(k),\mathrm{o}}$ has full column rank or that $\widehat{Q}^{(k),\mathrm{o}}_{L\cdot} = 0$ for all k. Later, Lemma 6.3, Proposition 6.1 and Lemma 6.4 show that these statements are true with high probability for all $k \leq K_s$.

Major steps of the proof. In the kth iteration of the oracle Algorithm 1, denote the matrices obtained after multiplication and thresholding by

$$\begin{aligned} &T^{(k),\mathrm{o}} = S^{\mathrm{o}}\widehat{Q}^{(k-1),\mathrm{o}} = \bigl(t^{(k),\mathrm{o}}_{vj}\bigr) \quad \text{and} \\ &\widehat{T}^{(k),\mathrm{o}} = \bigl(\widehat{t}^{(k),\mathrm{o}}_{vj}\bigr) \qquad \text{with } \widehat{t}^{(k),\mathrm{o}}_{vj} = \eta\bigl(t^{(k),\mathrm{o}}_{vj}, \gamma_{nj}\bigr). \end{aligned} \tag{6.2}$$

Further denote the QR factorization of $\widehat{T}^{(k),\mathrm{o}}$ by $\widehat{T}^{(k),\mathrm{o}} = \widehat{Q}^{(k),\mathrm{o}}\widehat{R}^{(k),\mathrm{o}}$. Last but not least, let $\widehat{\mathcal{P}}^{(k),\mathrm{o}}_m = \mathrm{ran}(\widehat{Q}^{(k),\mathrm{o}})$.

A joint proof of Theorems 3.2 and 3.3 can then be completed by the following three major steps:

(1) show that the principal subspace of S^{o} with dimension m, denoted by $\widehat{\mathcal{P}}^{\mathrm{o}}_m$, satisfies the error bound in (3.11) for estimating $\mathcal{P}_m$;

(2) show that for K in (3.10), $K_s \in [K, 2K]$ and that the approximation error of $\widehat{\mathcal{P}}^{(k),\mathrm{o}}_m$ to $\widehat{\mathcal{P}}^{\mathrm{o}}_m$ for all $k \geq K$ also satisfies the bound in (3.11);

(3) show that $\widehat{Q}^{(k),\mathrm{o}}_{L\cdot} = 0$ for all $k \leq 2K$, and that the oracle and the actual estimating sequences are identical up to $2K$ iterations.

In each step, we only need the result to hold with high probability. By the triangle inequality, steps 1 and 2 imply that the error of $\widehat{\mathcal{P}}_m^{(K_s),\mathrm{o}}$ in estimating $\mathcal{P}_m$ satisfies (3.11). Step 3 shows this is also the case for the actual estimator $\widehat{\mathcal{P}}_m^{(K_s)}$. It also implies the correct exclusion property in Theorem 3.3.

In what follows, we complete the three steps in Sections 6.2–6.4.

6.2. *Principal subspace of S^{o}.* To study how well the principal subspace of S^{o} approximates $\mathcal{P}_m$, we break into a "bias" part and a "variance" part.

Consider the "bias" part first. Define the oracle covariance matrix

$$\Sigma^{\mathrm{o}} = \begin{bmatrix} \Sigma_{HH} & 0 \\ 0 & I_{LL} \end{bmatrix}, \tag{6.3}$$

which is the expected value of S^{o}. The following lemma gives the error of the principal subspace of Σ^{o} in approximating $\mathcal{P}_m$, which could be regarded as the "squared bias" induced by feature selection.

LEMMA 6.1. *Let the eigenvalues of Σ^{o} be $\ell_1^{\mathrm{o}} \geq \cdots \geq \ell_{\bar{m}}^{\mathrm{o}} \geq \cdots \geq 0$ and $\{q_1^{\mathrm{o}}, \ldots, q_{\bar{m}}^{\mathrm{o}}\}$ be a set of first $\bar{m}$ eigenvectors. Denote $Q^{\mathrm{o}} = [q_1^{\mathrm{o}}, \ldots, q_m^{\mathrm{o}}]$. Then, uniformly over $\mathcal{F}_n$:*

(1) $|\ell_j^{\mathrm{o}} - (\lambda_j^2 + 1)|/\lambda_1^2 \to 0$ *as* $n \to \infty$, *for* $j = 1, \ldots, \bar{m}+1$, *with* $\lambda_{\bar{m}+1}^2 = 0$;

(2) *for sufficiently large n, $Q_{L\cdot}^{\mathrm{o}} = 0$ and for $\mathcal{P}_m^{\mathrm{o}} = \mathrm{ran}(Q^{\mathrm{o}})$, there exists a constant $C = C(m, r, \kappa)$, s.t. $L(\mathcal{P}_m, \mathcal{P}_m^{\mathrm{o}}) \leq C M_n \tau_{nm}^2$.*

A proof is given in the supplementary material [15]. Weyl's theorem ([28], Corollary 4.4.10) and Davis–Kahn's $\sin\theta$ theorem [4] are the key ingredients in the proof here, and also in the proofs of Lemmas 6.2 and 6.3. Here, claim (1) only requires Conditions GR and SP, but not the condition $\mathrm{AD}(m, \kappa)$.

Turn to the "variance" part. We check how well the principal subspace of S^{o} estimates $\mathcal{P}^{\mathrm{o}}$. Since $\Sigma^{\mathrm{o}} = \mathbb{E}[S^{\mathrm{o}}]$, the error here is analogous to "variance."

LEMMA 6.2. *Let the eigenvalues of S^{o} be $\widehat{\ell}_1^{\mathrm{o}} \geq \cdots \geq \widehat{\ell}_{\bar{m}}^{\mathrm{o}} \geq \cdots \geq 0$ and $\{\widehat{q}_1^{\mathrm{o}}, \ldots, \widehat{q}_{\bar{m}}^{\mathrm{o}}\}$ be a set of first $\bar{m}$ eigenvectors. Denote $\widehat{Q}^{\mathrm{o}} = [\widehat{q}_1^{\mathrm{o}}, \ldots, \widehat{q}_m^{\mathrm{o}}]$. Then, uniformly over $\mathcal{F}_n$, with probability at least $1 - C_0 p_n^{-2}$:*

(1) $|\widehat{\ell}_j^{\mathrm{o}} - \ell_j^{\mathrm{o}}|/\lambda_1^2 \to 0$ *as* $n \to \infty$, *for* $j = 1, \ldots, \bar{m}+1$;

(2) *for sufficiently large n, $\widehat{Q}_{L\cdot}^{\mathrm{o}} = 0$, and for $\widehat{\mathcal{P}}_m^{\mathrm{o}} = \mathrm{ran}(\widehat{Q}^{\mathrm{o}})$, there exist constants $C_1 = C_1(m, r, \kappa)$ and C_2, s.t.*

$$L(\mathcal{P}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{\mathrm{o}}) \leq C_1 M_n \tau_{nm}^2 / \log(p_n) + C_2 \varepsilon_{nm}^2.$$

A proof is given in the supplementary material [15]. Again, claim (1) does not require the condition $\mathrm{AD}(m, \kappa)$. By the triangle inequality, the above two lemmas imply the error in estimating $\mathcal{P}_m$ with $\widehat{\mathcal{P}}_m^{\mathrm{o}}$ satisfies the bound in (3.11).

6.3. *Properties of the oracle sequence.* In step 2, we study properties of the oracle sequence. For K in (3.10), the goal is to show that, with high probability, for all $k \geq K$, the error of the oracle subspace estimator $\widehat{\mathcal{P}}_m^{(k),\mathrm{o}}$ in approximating $\widehat{\mathcal{P}}_m^{\mathrm{o}}$ satisfies in (3.11). To this end, characterization of the oracle sequence evolution in Proposition 6.1 below plays the key role.

The initial point. We start with the initial point. Let

$$\rho = \widehat{\ell}_{m+1}^{\mathrm{o}} / \widehat{\ell}_m^{\mathrm{o}} \tag{6.4}$$

denote the ratio between the $(m+1)$th and the mth largest eigenvalues of S^{o}. The following lemma shows that $\widehat{Q}^{(0),\mathrm{o}}$ is orthonormal and is a good initial point for (oracle) Algorithm 1.

LEMMA 6.3. *Uniformly over $\mathcal{F}_n$, with probability at least $1 - C_0 p_n^{-2}$:*

(1) $B^{\mathrm{o}} = B$;
(2) $|\ell_j(S_{B^{\mathrm{o}}B^{\mathrm{o}}}) \vee 1 - \widehat{\ell}_j^{\mathrm{o}}|/\lambda_1^2 \to 0$ *as* $n \to \infty$, *for* $j = 1, \ldots, \bar{m}+1$;
(3) *for sufficiently large* n, $\widehat{Q}^{(0),\mathrm{o}}$ *has full column rank, and* $L(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(0),\mathrm{o}}) \leq (1-\rho)^2/5$;
(4) *for sufficiently large* n, $K_s \in [K, 2K]$.

A proof is given in the supplementary material [15]. Here, claims (1) and (2) do not require the condition $\mathrm{AD}(m, \kappa)$. In claim (3), the bound $(1-\rho)^2/5$ is much larger than that in (3.11). For instance, if $\lambda_m^2 + 1 \asymp \lambda_m^2 - \lambda_{m+1}^2$, Lemmas 6.1 and 6.2 imply that $(1-\rho^2)/5 \asymp 1$ with high probability.

Claims (1) and (2) here, together with claims (1) of Lemmas 6.1 and 6.2, lead to the following result on consistent estimation of $\lambda_1^2/(\lambda_j^2 - \lambda_{j+1}^2)$ and λ_j^2, the proof of which is given in the supplementary material [15].

COROLLARY 6.1. *Suppose Conditions* GR *and* SP *hold, and let* $\ell_j^B = \ell_j(S_{BB}) \vee 1$. *For* $1 \leq j \leq \bar{m}$, *if* $\lim_{n\to\infty} \lambda_1^2/(\lambda_j^2 - \lambda_{j+1}^2) < \infty$, *then*

$$\lim_{n\to\infty} \frac{(\ell_1^B - 1)/(\ell_j^B - \ell_{j+1}^B)}{\lambda_1^2/(\lambda_j^2 - \lambda_{j+1}^2)} = 1 \qquad a.s.$$

Otherwise, $\lim_{n\to\infty}(\ell_1^B - 1)/(\ell_j^B - \ell_{j+1}^B) = \lim_{n\to\infty} \lambda_1^2/(\lambda_j^2 - \lambda_{j+1}^2) = \infty$, *a.s.*

If further the condition $\mathrm{AD}(m, \kappa)$ *holds for some* $m \leq \bar{m}$ *and* $\kappa > 0$, *then* $\lim_{n\to\infty}(\ell_j^B - 1)/\lambda_j^2 = 1$, *a.s., for* $1 \leq j \leq m$.

Evolution of the oracle sequence. Next, we study the evolution of the oracle sequence. Let $\theta^{(k)} \in [0, \pi/2]$ be the largest canonical angle between the subspaces $\widehat{\mathcal{P}}_m^{\mathrm{o}}$ and $\widehat{\mathcal{P}}_m^{(k),\mathrm{o}}$. By the discussion after (2.1), we have

$$\sin^2 \theta^{(k)} = L\big(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(k),\mathrm{o}}\big). \tag{6.5}$$

The following proposition describes the evolution of $\theta^{(k)}$ over iterations.

PROPOSITION 6.1. *Let n be sufficiently large. On the event such that the conclusions of Lemmas 6.1–6.3 hold, uniformly over $\mathcal{F}_n$, for all $k \geq 1$:*

(1) *$\widehat{Q}^{(k),\mathrm{o}}$ is orthonormal, and $\theta^{(k)}$ satisfies*

$$\sin \theta^{(k)} \leq \rho \tan \theta^{(k-1)} + \omega \sec \theta^{(k-1)}, \tag{6.6}$$

where $\omega = (\widehat{\ell}_m^{\mathrm{o}})^{-1}[\mathrm{card}(H) \sum_{j=1}^m \gamma_{nj}^2]^{1/2}$;

(2) *for any $a \in (0, 1/2]$, if*

$$\sin^2 \theta^{(k-1)} \leq 1.01(1-a)^{-2}\omega^2(1-\rho)^{-2}, \tag{6.7}$$

then so is $\sin^2 \theta^{(k)}$. Otherwise,

$$\sin^2 \theta^{(k)} / \sin^2 \theta^{(k-1)} \leq \big[1 - a(1-\rho)\big]^2. \tag{6.8}$$

A proof is given in the supplementary material [15], the key ingredient of which is Wedin's $\sin\theta$ theorem for singular subspaces [33]. The recursive inequality (6.6) characterizes the evolution of the angles $\theta^{(k)}$, and hence of the oracle subspace $\widehat{\mathcal{P}}_m^{(k),\mathrm{o}}$. It is the foundation of claim (2) in the current proposition and of Proposition 6.2 below.

By (6.5), inequality (6.8) gives the rate at which the approximation error $L(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(k),\mathrm{o}})$ decreases. For a given $a \in (0, 1/2]$, the rate is maintained until the error becomes smaller than $1.01(1-a)^{-2}\omega^2(1-\rho)^{-2}$. Then the error continues to decrease, but at a slower rate, say, with a replaced by $a/2$ in (6.8), until (6.7) is satisfied with a replaced by $a/2$. The decrease continues at slower and slower rate in this fashion until the approximation error falls into the interval $[0, 1.01\omega^2/(1-\rho)^2]$, and remains inside thereafter.

Together with Lemma 6.3, Proposition 6.1 also justifies the previous claim that elements of the oracle sequence are orthonormal with high probability.

Convergence. Finally, we study how fast the oracle sequence converges to a stable subspace estimator, and how good this estimator is.

To define convergence of the subspace sequence $\{\widehat{\mathcal{P}}_m^{(k),\mathrm{o}}, k \geq 0\}$, we first note that $1.01\omega^2/(1-\rho)^2$ is almost the smallest possible value of $L(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(k),\mathrm{o}})$ that

(6.6) could imply. Indeed, when $\sin\theta^{(k)}$ converges and is small, we have $\sin\theta^{(k)} \approx \sin\theta^{(k-1)}$, and $\cos\theta^{(k)} \approx 1$. Consequently, (6.6) reduces to

$$\sin\theta^{(k)} \le \bigl(\rho\sin\theta^{(k)} + \omega\bigr)\bigl(1 + o(1)\bigr).$$

So, $L(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(k),\mathrm{o}}) = \sin^2\theta^{(k)} \le (1 + o(1))\omega^2/(1-\rho)^2$. In addition, Lemma 6.2 suggests that we can stop the iteration as soon as $L(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(k),\mathrm{o}})$ becomes smaller than a constant multiple of ε_{nm}^2, for we always get an error of order $O(\varepsilon_{nm}^2)$ for estimating $\mathcal{P}_m$, even if we use $\widehat{\mathcal{P}}_m^{\mathrm{o}}$ directly. In observation of both aspects, we say that $\widehat{\mathcal{P}}_m^{(k),\mathrm{o}}$ has *converged* if

$$L\bigl(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(k),\mathrm{o}}\bigr) \le \max\left\{\frac{1.01}{(1-n^{-1})^2}\frac{\omega^2}{(1-\rho)^2}, \varepsilon_{nm}^2\right\}. \tag{6.9}$$

On the event that conclusions of Lemmas 6.1–6.3 hold, we have $\omega^2/(1-\rho)^2 = O(M_n\tau_{nm}^2)$. Under definition (6.9), for K in (3.10), the following proposition shows that it takes K iterations for the oracle sequence to converge, and for all $k \ge K$, the error of approximating $\widehat{\mathcal{P}}_m^{\mathrm{o}}$ by $\widehat{\mathcal{P}}_m^{(k),\mathrm{o}}$ satisfies (3.11).

PROPOSITION 6.2. *For sufficiently large n, on the event such that the conclusions of Lemmas 6.1–6.3 hold, uniformly over $\mathcal{F}_n$, it takes at most K steps for the oracle sequence to converge. In addition, there exist constants $C_1 = C_1(\gamma, r, m, \kappa)$ and C_2, such that for all $k \ge K$,*

$$\sup_{\mathcal{F}_n} L\bigl(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(k),\mathrm{o}}\bigr) \le C_1 M_n\tau_{nm}^2 + C_2\varepsilon_{nm}^2. \tag{6.10}$$

A proof is given in the supplementary material [15], and this completes step 2.

6.4. *Proof of main results.* We now prove the properties of the actual estimating sequence. The proof relies on the following lemma, which shows the actual and the oracle sequences are identical up to $2K$ iterations.

LEMMA 6.4. *For sufficiently large n, with probability at least $1 - C_0 p_n^{-2}$, for all $k \le 2K$, we have $\widehat{Q}_{L\cdot}^{(k),\mathrm{o}} = 0$, $\widehat{Q}^{(k)} = \widehat{Q}^{(k),\mathrm{o}}$, and hence $\widehat{\mathcal{P}}_m^{(k)} = \widehat{\mathcal{P}}_m^{(k),\mathrm{o}}$.*

A proof is given in the supplementary material [15], and this completes step 3.

We now prove Theorems 3.2 and 3.3 by showing that the actual sequence inherits the desired properties from the oracle sequence. Since Theorem 3.1 is a special case of Theorem 3.2, we do not give a separate proof.

PROOF OF THEOREM 3.2. Note that the event on which the conclusions of Lemmas 6.1–6.4 hold has probability at least $1 - C_0 p_n^{-2}$. On this event,

$$\begin{aligned} L(\mathcal{P}_m, \widehat{\mathcal{P}}_m^{(K_s)}) &= L(\mathcal{P}_m, \widehat{\mathcal{P}}_m^{(K_s),\mathrm{o}}) \\ &\le \bigl[L^{1/2}(\mathcal{P}_m, \mathcal{P}_m^{\mathrm{o}}) + L^{1/2}(\mathcal{P}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{\mathrm{o}}) + L^{1/2}(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(K_s),\mathrm{o}})\bigr]^2 \\ &\le C\bigl[L(\mathcal{P}_m, \mathcal{P}_m^{\mathrm{o}}) + L(\mathcal{P}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{\mathrm{o}}) + L(\widehat{\mathcal{P}}_m^{\mathrm{o}}, \widehat{\mathcal{P}}_m^{(K_s),\mathrm{o}})\bigr] \\ &\le C_1 M_n \tau_{nm}^2 + C_2 \varepsilon_{nm}^2. \end{aligned}$$

Here, the first equality comes from Lemma 6.4. The first two inequalities result from the triangle inequality and Jensen's inequality, respectively. Finally, the last inequality is obtained by noting that $K_s \in [K, 2K]$ and by replacing all the error terms by their corresponding bounds in Lemmas 6.1, 6.2 and Proposition 6.2. □

PROOF OF THEOREM 3.3. Again, we consider the event on which the conclusions of Lemmas 6.1–6.4 hold. Then Lemma 6.4 directly leads to the conclusion that $\widehat{Q}_{L\cdot}^{(k)} = \widehat{Q}_{L\cdot}^{(k),\mathrm{o}} = 0$, for all $0 \le k \le K_s \le 2K$. □

Acknowledgment. The author would like to thank Iain Johnstone for many helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "Sparse principal component analysis and iterative thresholding" (DOI: 10.1214/13-AOS1097SUPP; .pdf). We give in the supplement proofs to Corollaries 3.1 and 3.2, Proposition 3.1 and all the claims in Section 6.

REFERENCES

[1] AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

[2] ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** 122–148. MR0145620

[3] D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448 (electronic). MR2353806

[4] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450

[5] DONOHO, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.* **1** 100–115. MR1256530

[6] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[7] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

[8] HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24** 417–441, 498–520.

[9] HOYLE, D. C. and RATTRAY, M. (2004). Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Phys. Rev. E* (3) **69** 026124.
[10] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961
[11] 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. MR2751448
[12] 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. MR2002634
[13] JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454
[14] LU, A. Y. (2002). Sparse principal component analysis for functional data. Ph.D. thesis, Stanford Univ., Stanford, CA. MR2703298
[15] MA, Z. (2013). Supplement to "Sparse principal component analysis and iterative thresholding." DOI:10.1214/13-AOS1097SUPP.
[16] MALLAT, S. (2009). *A Wavelet Tour of Signal Processing*: *The Sparse Way*. Academic Press, New York.
[17] NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013
[18] NADLER, B. (2009). Discussion of "On consistency and sparsity for principal components analysis in high dimensions," by I. M. Johnstone and A. Y. Lu. *J. Amer. Statist. Assoc.* **104** 694–697. MR2751449
[19] ONATSKI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* **168** 244–258. MR2923766
[20] PAUL, D. (2005). Nonparametric estimation of principal components. Ph.D. thesis, Stanford Univ. MR2707156
[21] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865
[22] PAUL, D. and JOHNSTONE, I. M. (2007). Augmented sparse principal component analysis for high dimensional data. Available at arXiv:1202.1242v1.
[23] PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag. Ser.* 6 **2** 559–572.
[24] RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993
[25] REIMANN, P., VAN DEN BROECK, C. and BEX, G. J. (1996). A Gaussian scenario for unsupervised learning. *J. Phys. A* **29** 3521–3535.
[26] SHEN, D., SHEN, H. and MARRON, J. S. (2011). Consistency of sparse PCA in high dimension, low sample size contexts.
[27] SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336
[28] STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA. MR1061154
[29] TSAY, R. S. (2005). *Analysis of Financial Time Series*, 2nd ed. Wiley, Hoboken, NJ. MR2162112
[30] ULFARSSON, M. O. and SOLO, V. (2008). Sparse variable PCA using geodesic steepest descent. *IEEE Trans. Signal Process.* **56** 5823–5832. MR2518261
[31] VARMUZA, K. and FILZMOSER, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Boca Raton, FL.
[32] WAX, M. and KAILATH, T. (1985). Detection of signals by information theoretic criteria. *IEEE Trans. Acoust. Speech Signal Process.* **33** 387–392. MR0788604
[33] WEDIN, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *Nordisk Tidskr. Informationsbehandling* (*BIT*) **12** 99–111. MR0309968

[34] 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.

[35] YUAN, X. T. and ZHANG, T. (2011). Truncated power method for sparse eigenvalue problems. Available at arXiv:1112.2679v1.

[36] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: zongming@wharton.upenn.edu