Sparse Principal Component Analysis and Iterative Thresholding

Zongming Ma*

University of Pennsylvania

September 2, 2011

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

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) \cite{25, 10} reduces dimensionality by projecting the data onto the principal subspace spanned by the $m$ leading eigenvectors of the population covariance matrix $\Sigma$, 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, $\Sigma$ 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 \cite{2}.

\* Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. Email: zongming@wharton.upenn.edu
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 instantiation of the following generative model:

$$x_i = \mu + Au_i + \sigma z_i. \quad (1.1)$$

Here, $\mu$ 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 [33]. Here, 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 [31]. 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 typically also at the scale of tens or hundreds. In addition, model (1.1) also represents a big class of signal processing problems [34]. 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 $\Phi$, and suppose that $A$ has full column rank and $u_i$ and $z_i$ are independent. Then, the covariance matrix of $x_i$ becomes

$$\Sigma = AA' + \sigma^2 I = \sum_{j=1}^{\bar{m}} \lambda_j^2 q_j q_j' + \sigma^2 I. \quad (1.2)$$

Here, $\lambda_1^2 \geq \cdots \geq \lambda_{\bar{m}}^2 > 0$ are the eigenvalues of $AA'$, with $q_j$, $j = 1, \ldots, \bar{m}$, the associated eigenvectors. Therefore, the $j$-th eigenvalue of $\Sigma$ is $\lambda_j^2 + \sigma^2$ for $j = 1, \ldots, \bar{m}$, and $\sigma^2$ otherwise. Since there are $\bar{m}$ spikes ($\lambda_1^2, \ldots, \lambda_{\bar{m}}^2$) in the spectrum of $\Sigma$, (1.2) has been called the spiked covariance model in the literature [12]. 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.

Classical PCA encounters both practical and theoretical difficulties in high dimensions. On the practical side, the eigenvectors found by classical PCA typically involve all the $p$ features, which makes their interpretation challenging. On the theoretical side, the sample eigenvectors are no longer always consistent estimators of their population counterparts. 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 [27, 17, 11, 23, 19, 21] under model (1.2). See [16] for similar results in the case where $p \to \infty$ 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, [15, 37, 6, 29, 32, 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 $\Sigma$ indeed has sparse leading eigenvectors (perhaps in some transform domain), it becomes possible to estimate them consistently under high-dimensional settings via
new estimation schemes. For example, under normality assumption, when $\Sigma$ only has a single spike, i.e., when $\bar{m} = 1$ in (1.2), Johnstone and Lu [1] 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 $k$ non-zero loadings, Amini and Wainwright [3] studied conditions for recovering the non-zero locations using the methods in [14] and [3], and Shen, et al [28] established conditions for consistency of a sparse PCA method in [29] when $p \to \infty$ and $n$ is fixed. For the more general multiple component case, Paul and Johnstone [24] 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 $\Sigma$ 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 of the leading eigenvalues are identical or close to each other. In addition, 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 the matrix computation literature. In addition to the usual steps of orthogonal iteration, an additional thresholding step is added to seek a sparse basis for the subspace. When $\Sigma$ follows the spiked covariance model, the algorithm is shown to yield a uniformly consistent subspace estimator over a wide range of high-dimensional sparse settings, and the rate of convergence is given under an appropriate loss function. 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 attains near optimal rate of convergence. In addition, the algorithm also has appealing model selection property and computational efficiency.

The contribution of the current paper is twofold. 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 theoretically justified under an informative model, efficient in computation, and easy to implement.

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 theoretical results.

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 use $\|x\|_2$ to denote the Euclidean norm of a vector $x$. 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 $\|A\| = \max_{\|x\|=1} \|Ax\|_2$, and the range, i.e., the column subspace, of $A$ is $\text{ran}(A)$. 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 $|a_n| \leq Cb_n$ for all $n$, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Throughout the paper, we use $\nu$ 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 $\Sigma$ follows the spiked covariance model \([12]\), its $j$-th largest eigenvalue $\ell_j(\Sigma) = \lambda_j^2 + \sigma^2$ for $j = 1, \ldots, \bar{m}$, and equals $\sigma^2$ for all $j > \bar{m}$. Let $\text{span}\{\cdot\}$ 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

$$P_m = \text{span}\{q_1, \ldots, q_m\}$$

is well 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 $P_{\bar{m}}$ 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 $P_m$, when the target dimension $m$ is given.

We do not always aim at $P_{\bar{m}}$ directly for two reasons. First, the number of factors $\bar{m}$ is usually not known a priori and needs to be estimated. So, always aiming for $P_{\bar{m}}$ can be too ambitious. Second, under certain circumstances, one might only be interested in the first several principal subspaces. For example, to visualize the data, one typically only need $P_2$ or $P_3$.

To measure the accuracy of an estimator for a subspace, consider a subspace $S$ and an estimator $\hat{S}$ such that $\text{dim}(\hat{S}) = \text{dim}(S)$. Note that for each linear subspace, there is a unique orthogonal projection matrix with that subspace as its range. Let $P$ and $\hat{P}$ be the projection matrices associated with $S$ and $\hat{S}$, respectively. The distance between $S$ and $\hat{S}$ is given by the spectral norm of the difference between $P$ and $\hat{P}$: $\text{dist}(S, \hat{S}) = \|P - \hat{P}\|$. See, for example, Sec. 2.6.3 in \[9\]. Thus, we can define a loss function by the squared distances between $S$ and $\hat{S}$:

$$L(S, \hat{S}) = \text{dist}^2(S, \hat{S}) = \|P - \hat{P}\|^2. \quad (2.1)$$

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 $\hat{S} = S$. Geometrically, it equals the squared sine of the largest canonical angle between $S$ and $\hat{S}$ \[30\, Theorem 5.5\]. 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 low-dimensional eigenspace is orthogonal iteration \[9\]. In the special case where only the first eigenvector is sought, it is also known as the power method.

Suppose that $A$ 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:
Algorithm 1: ITSPCA (Iterative thresholding sparse PCA)

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

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

1. repeat
2. Multiplication: $T^{(k)} = (t^{(k)}_{\nu j}) = S\hat{Q}^{(k-1)}$;
3. Thresholding: $\hat{T}^{(k)} = (\hat{t}^{(k)}_{\nu j})$, with $\hat{t}^{(k)}_{\nu j} = \eta(t^{(k)}_{\nu j}; \gamma_{nj})$;
4. QR factorization: $\hat{Q}^{(k)}\hat{R}^{(k)} = \hat{T}^{(k)}$;
5. until 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 $\text{ran}(Q^{(\infty)})$ gives the eigenspace. In practice, one terminates the iteration once $\text{ran}(Q^{(k)})$ stabilizes.

Orthogonal iteration is useful for computing low-dimensional eigenspaces of high-dimensional matrices. However, when applied directly to a sample covariance matrix $S$, it gives the classical PCA result, which could be problematic in high dimensions. Observe that all the coordinates are included in orthogonal iteration. When the number of coordinates is large, not only the interpretation is hard, but the variance accumulated across all the coordinates becomes so high that it makes consistent estimation impossible.

If the eigenvectors spanning $P_m$ are sparse, one sensible way to reduce estimation error is to focus only on those coordinates at which the leading eigenvectors have large values, and to estimate other coordinates 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 motivate us to develop in the next subsection an estimation scheme which incorporates this coordinate screening idea, while at the same time retains the simplicity of orthogonal iteration.

### 2.4 Iterative Thresholding Algorithm

An effective way to incorporate coordinate screening into orthogonal iteration is to ‘kill’ small coordinates of the $T^{(k)}$ matrix after each multiplication step, which leads to the iterative thresholding 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 steps in orthogonal iteration, Algorithm 1 adds a thresholding step in between them, where we threshold each element of $T^{(k)}$ with a user-specified thresholding
Algorithm 2: DTSPCA (Diagonal thresholding sparse PCA)

Input:
1. Sample covariance matrix $S$;
2. Target subspace dimension $m$;
3. Diagonal thresholding parameter $\alpha_n$.

Output: Orthonormal matrix $\hat{Q}^{(0)}$.

1. Variance selection: select the set $B$ of coordinates (which are likely to have “big” signals):
   $$B = \{ \nu : s_{\nu\nu} \geq \sigma^2(1 + \alpha_n) \};$$

2. Reduced PCA: compute $m$ leading eigenvectors, $\hat{\mathbf{q}}_1^B, \ldots, \hat{\mathbf{q}}_m^B$, of the submatrix $S_{BB}$;

3. Zero-padding: construct $\hat{Q}^{(0)} = [\hat{\mathbf{q}}_1^{(0)}, \ldots, \hat{\mathbf{q}}_m^{(0)}]$ such that,
   $$\hat{\mathbf{q}}_j^B = \hat{\mathbf{q}}_j^0, \quad \hat{\mathbf{q}}_j^B = 0, \quad j = 1, \ldots, m.$$

function $\eta$ which satisfies for all $t$ and all $\gamma > 0$,

$$|\eta(t, \gamma) - t| \leq \gamma, \quad \text{and} \quad \eta(t, \gamma)1_{(|t| \leq \gamma)} = 0. \quad (2.2)$$

Here, $1_{(E)}$ denotes the indicator function of an event $E$. We note that both the hard-thresholding function $\eta_H(t, \gamma) = t 1_{(|t| > \gamma)}$ and the soft-thresholding function $\eta_S(t, \gamma) = \text{sgn}(t)(|t| - \gamma)_+$ satisfy condition (2.2). So does any $\eta$ that is sandwiched by them, such as the thresholding function resulting from a SCAD criterion [7]. In $\eta(t, \gamma)$, the parameter $\gamma$ is called the threshold level. In Algorithm 1, for each column of $T^{(k)}$, a common threshold level $\gamma_{nj}$ needs to be specified for all its elements. These threshold levels remain unchanged across iterations.

Remark 2.1. The ranges of $\hat{Q}^{(k)}$ and $\hat{T}^{(k)}$ 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 $\hat{Q}^{(\infty)}$ can be used as estimators of leading eigenvectors under sparsity assumption.

Construction of $\hat{Q}^{(0)}$ Note that Algorithm 1 requires an initial orthonormal matrix $\hat{Q}^{(0)}$. It can be generated by the ‘diagonal thresholding’ sparse PCA algorithm proposed by Johnstone and Lu [14]. The multiple eigenvector version of their proposal is summarized in Algorithm 2. When $\sigma^2$ is unknown, we could replace it by an estimator $\hat{\sigma}^2$ in the definition of $B$. For example, Johnstone and Lu [14] suggested

$$\hat{\sigma}^2 = \text{median}\left(\frac{1}{n} \sum_{i=1}^n x_{i\nu}^2\right) \quad (2.3)$$

for normal data. When available, subject knowledge could also be incorporated into the construction of $\hat{Q}^{(0)}$. 

6
Bibliographical note In the special case of \( m = 1 \), Algorithm 1 is similar to the sPCA-rSVD algorithm by Shen and Huang [29] and the SPC algorithm by Witten, et al [36]. When \( m > 1 \), both the sPCA-rSVD and the SPC 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. The main results include rate of convergence for principal subspace estimation and a correct exclusion property. We also investigate the convergence rate for individual eigenvector estimation under appropriate conditions.

Assume that the observation vectors \( x_1, \ldots, x_n \) are i.i.d. \( N_p(0, \Sigma) \) distributed, with \( \Sigma \) following the spiked covariance model (1.2). Further assume that \( \sigma^2 \) is known – though this assumption could be removed by estimating \( \hat{\sigma}^2 \) using, say, \( \hat{\sigma}^2 \) in (2.3). Since one could always scale the data first, we assume \( \sigma^2 = 1 \) from now on. Thus, (1.1) reduces to the following orthogonal factor form

\[
x_i = \sum_{j=1}^m \lambda_j v_{ij} q_j + z_i, \quad i = 1, \ldots, n.
\]  

(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 \( \{q_j, 1 \leq j \leq \bar{m}\} \) is a set of leading eigenvectors of \( \Sigma \).

Throughout the section, the target subspace dimension is assumed to be pre-specified by some \( m \leq \bar{m} \), and both \( m \) and \( \bar{m} \) remain unchanged as \( p \) grows. Here and after, let \( p_n = p^\vee n \), because we are mostly concerned with settings where \( p \) is greater than \( n \) and scales at certain rate as \( n \to \infty \). For concreteness, the initial orthonormal matrix \( \hat{Q}^{(0)} \) is always obtained via Algorithm 2 with

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

(3.2)

for constructing the set \( B \) in step 1. In addition, the threshold levels in Algorithm 2 are set at

\[
\gamma_{nj} = \gamma \left[ \ell_j(S_{BB}) \frac{\log(p_n)}{n} \right]^{1/2}, \quad j = 1, \ldots, m.
\]

(3.3)

Here, \( \alpha \) and \( \gamma \) are user specified constants, and \( \ell_j(S_{BB}) \) is the \( j \)-th largest eigenvalue of \( S_{BB} \), where the set \( B \) is obtained in step 1 of Algorithm 2.

3.1 A Special Case

To facilitate understanding, we first state the rate of convergence result 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 \( \lambda_1^2 \geq \cdots \geq \lambda_m^2 > \lambda_{m+1}^2 \geq \cdots \lambda_{\bar{m}}^2 > 0 \) remain unchanged. Suppose that the \( q_j \)'s are sparse in the sense that, for some \( r \in (0, 2) \), the \( \ell_r \) norm of the eigenvectors are uniformly bounded by \( s \), i.e.,

\[
\|q_j\|_r = (\sum_{j=1}^p |q_{\nu_j}|^r)^{1/r} \leq s, \quad \text{for } j = 1, \ldots, \bar{m}.
\]

Here, \( s \geq 1 \) is a constant not depending on \( p \).

Under the above setup, for \( h(x) = x^2/(x + 1) \), we have the following upper bound on the rate of convergence for subspace estimation error.
Theorem 3.1. Under the above setup, for sufficiently large constants $\alpha$ and $\gamma$ in (2.2) and (2.3), there exist constants $C_0$, $C_1 = C_1(\gamma, r, m)$ and $C_2$, such that for sufficiently large $n$, with probability at least $1 - C_0p_n^{-2}$, the subspace estimator $\hat{P}_m = \text{ran}(\hat{Q}(k))$ in the $k$-th iteration of Algorithm 1 satisfies

$$\sup_{\|q_j\| \leq s, \forall j} L(P_m, \hat{P}_m(k)) \leq C_1 \sqrt{s} \left[ \frac{\log p}{n h(\lambda_m^2)} \right]^{1-r/2} + C_2 g_m(\lambda) \frac{\log p}{n},$$

for all $k = K, \ldots, 2K$, where $K = O(\log n)$ and $g_m(\lambda) = \frac{(\lambda_m^2 + 1)(\lambda_{m+1}^2 + 1)}{(\lambda_m^2 - \lambda_{m+1}^2)^2}$.

The upper bound in Theorem 3.1 consists of two terms. The first term is a “nonparametric” term, which can be decomposed as the product of two components. The first component, $\sqrt{s} [\lambda_m^2 / \log p]^{r/2}$, up to a multiplicative constant, bounds the number of coordinates used in estimating the subspace, while the second component, $\log p / [n h(\lambda_m^2)]$, 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, large $n$” asymptotic limit for classical PCA estimator (cf., [2, Theorem 1]). We could 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 in the upper bound converge to $0$ as $n \to \infty$, which establishes the consistency of our estimator.

To better understand the upper bound, we compare it with some lower bound result. Suppose $\lambda_1^2 > \lambda_2^2$. Consider the simplest case where $m = 1$. Then, estimating $P_1$ is the same as estimating the first eigenvector $q_1$. For estimating any individual eigenvector $q_j$, Paul and Johnstone [24] considered the loss function $l(q_j, \hat{q}_j) = \|q_j - \text{sgn}(\hat{q}_j^T q_j) \hat{q}_j\|_2^2$. In the special case here, the $\lambda_j^2$'s, $s$ and $r \in (0, 2)$ are fixed and $p \geq n$, so when $n$ is large, $s [\lambda_j^2 / \log p]^{r/2} \leq C p^{1-c}$ for some $c \in (0, 1)$. For this case, Theorem 2 in [24] asserts that for any estimator $\hat{q}_1$,

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

Let $\hat{P}_1 = \text{span}\{\hat{q}_1\}$. We have $\frac{1}{n} l(q_1, \hat{q}_1) \leq L(P_1, \hat{P}_1) \leq l(q_1, \hat{q}_1)$. So the above lower bound also holds for any $\hat{P}_1$ and $\mathbb{E} L(P_1, \hat{P}_1)$. Therefore, Theorem 3.1 shows that the estimator from Algorithm 1 is near optimal, since it attains the minimax lower bound within a logarithmic factor.

The theorem also states that the estimator could be obtained in $O(\log n)$ iterations, and the upper bound holds for all thresholding function $\eta$ satisfying (2.2). Last but not least, since the choices of $\alpha_n$ and $\gamma_{mj}$ do not involve any unknown parameter, the theorem establishes the adaptivity of our estimator: the near optimal rate of convergence is obtained without any knowledge of the power $r$, the radius $s$ or the spikes $\lambda_j^2$.

Remark 3.1. Although special, the case considered here is still reasonable for a number of settings. For example, the settings of numerical experiments in Section 5.

Later in Section 3.3, Theorem 3.2 establishes analogous rate of convergence result, 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 $\lambda_1^2, \ldots, \lambda_m^2$ will be allowed to scale as $n \to \infty$, and $\lambda_1^2, \ldots, \lambda_m^2$ 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-$\ell_r$ ball of radius $s_j$ (which contains the $\ell_r$ ball of the same radius), and the radii $s_j$’s will be allowed to diverge as $n \to \infty$.

In what follows, we first state precise assumptions for the aforementioned extension in Section 3.2, followed by the introduction of a few key quantities in Section 3.3. Finally, the main results are stated in Section 3.4.

3.2 Assumptions

Here, we 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 $\lambda^2_j = \lambda^2_j(n) > 0$ to depend on $n$, though the dependence will usually not be shown explicitly. Recall that $p_n = p \lor n$, we impose the following growth rate condition on $p$ and the $\lambda^2_j$’s.

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

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

2. the largest spike $\lambda^2_1$ satisfies $\log(\lambda^2_1) = O(\log(p_n))$; the smallest spike $\lambda^2_\tilde{m}$ satisfies $\log(p_n) = o(n\lambda^4_\tilde{m})$; and their ratio satisfies $\lambda^2_1/\lambda^2_\tilde{m} = O(n\log(p_n)/n^{1/2+r/4})$.

The first part of condition GR requires the dimension to grow at a sub-exponential rate of the sample size. The second part ensures that the spikes grow at most at a polynomial rate with $p_n$, and are all of larger magnitude compared to the sample size. The second part ensures that the spikes grow at most at a polynomial rate with $p_n$, and are all of larger magnitude compared to $\sqrt{\log(p_n)/n}$. In addition, the condition on the ratio $\lambda^2_1/\lambda^2_\tilde{m}$ 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 [24] that all the spikes grow at the same rate.

Turn to the sparsity assumption on the $q_j$’s. We first make a mild extension from $\ell_r$ ball to weak-$\ell_r$ ball [6]. To this end, for any $p$-vector $u$, order its coordinates by magnitude as $|u|^1 \geq \cdots \geq |u|^p$. We say that $u$ belongs to the weak-$\ell_r$ ball of radius $s$, denoted by $u \in w\ell_r(s)$, if

$$|u|^\nu \leq \frac{s}{p^{1/\nu}}, \quad \text{for all } \nu.$$ 

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 $k$ non-zero 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-$\ell_r$ ball extends $\ell_r$ ball, because $|u|_r \leq s$, i.e., $u \in \ell_r(s)$, implies $u \in w\ell_r(s)$.

In what follows, we assume that for some fixed $r \in (0, 2)$ and all $j \leq \tilde{m}$, $q_j \in w\ell_r(s_j)$ for some $s_j \geq 1$. We choose to use the notion of “weak-$\ell_r$ 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 [6]. 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 $\Sigma$ belongs to the following uniformity class

$$\mathcal{F}_n = \left\{ \Sigma_{p \times p} = \sum_{j=1}^m \lambda^2_j q_j q_j' + I : q_j \in w\ell_r(s_j), \forall j \right\}.$$
For our general results, we allow the radii $s_j$'s to depend on or even diverge with $n$, though we need to assume 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-$\ell_r$ ball satisfies $s_j \geq 1$ and

$$s_j \left[ \frac{\log(p_n)}{n \lambda_2^j} \right]^{1/2 - r/4} = o(1 \land \lambda_2^j), \quad \text{for } j = 1, \ldots, \tilde{m}.$$ 

We note that this type of condition also appeared in previous investigation on estimating individual eigenvectors in the multiple component spiked covariance model [24]. The condition is, for example, satisfied if condition GR holds and the largest spike $\lambda_2^j$ is bounded away from zero while the radii $s_j$'s are all bounded above by an arbitrarily large constant. That is, if there exist a constant $C > 0$, such that $\lambda_2^j \geq 1/C$ and $s_j \leq C$ for all $j \leq \tilde{m}$ 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 regularly spaced grid, model (3.1) becomes discretization of a functional PCA model [26], 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 $\ell_r$ ball [6]. So do the discrete wavelet transform of the $q_j$'s. Moreover, the radii of the weak $\ell_r$ 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 $\lambda_2^j$ 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 Preliminaries

We now introduce a few 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 $j = 1, \ldots, \tilde{m}$, define

$$\tau_{nj} = \sqrt{\log(p_n)} \left[ h(\lambda_2^j) \right]. \quad (3.4)$$

According to Paul [22], up to a logarithmic factor, $\tau_{nj}^2$ can be interpreted as the average error per coordinate in estimating an eigenvector with eigenvalue $\lambda_2^j + 1$. 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 $\tau_{nj}$. Otherwise, we call it a low signal coordinate.

Given the cutoff rate $\tau_{nj}$ between high and low signals, we can define the set of high signal coordinates

$$H = H(\beta) = \{ \nu : |q_{\nu j}| \geq \beta \tau_{nj}, \text{ for some } 1 \leq j \leq \tilde{m} \}. \quad (3.5)$$

Here, $\beta$ is a constant that does not depend on $n$, the actual value of which will be specified later. In addition, we let $L = \{1, \ldots, p\} \setminus H$ be the complement of $H$. Here, $H$ stands for “high”, and $L$ for “low” (also recall the set $B$ 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 our subspace estimator, it is important to have an upper bound for \( \text{card}(H) \), the cardinality of \( H \). As we shall show later, the upper bound is a constant multiple of

\[
M_n = p \land \sum_{j=1}^{\bar{m}} \frac{s_j^r}{r_j^{nm}}.
\]  

(3.6)

Here we use capital letter to indicate the order \( m \leq \bar{m} \leq M_n \). In the general result, \( M_n \) plays the same role as \( \bar{m} s^r [nh(\lambda^2)/\log p]^{r/2} \) plays in Theorem 3.1. See the discussion following Theorem 3.1.

The last quantity we introduce is related to the “parametric” term in the convergence rate. For \( j = 1, \ldots, \bar{m} \), define

\[
\epsilon_n^2 = \frac{(\lambda_j^2 + 1)(\lambda_{j+1}^2 + 1) \log(p_n)}{\left(\lambda_j^2 - \lambda_{j+1}^2\right)^2 n},
\]  

(3.7)

where \( \lambda_{\bar{m}+1}^2 = 0 \). So the second term of the upper bound in Theorem 3.1 is then \( C_2 \epsilon_{nm}^2 \). For the interpretation of this quantity, see the discussion following Theorem 3.1.

### 3.4 Main Results

We turn to the statement of main theoretical results.

A key condition for our results is the asymptotic distinguishability (AD) condition introduced below. Recall that all the spikes \( \lambda_j^2 \) (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 AD\((j, \kappa)\) is satisfied if there exists a numeric constant \( \kappa \), such that the gap between the \( j \)-th and the \((j + 1)\)-th eigenvalues satisfies

\[
\lambda_j^2 - \lambda_{j+1}^2 \geq \frac{\lambda_j^2}{\kappa}, \quad \text{for all } n.
\]

In Corollary 3.2 below, we will need a pair of conditions AD\((j - 1, \kappa)\) and AD\((j, \kappa)\) for some \( j \leq \bar{m} \), hence we define AD\((0, \kappa)\) and AD\((\bar{m}, \kappa)\) by letting \( \lambda_0^2 = \infty \), and \( \lambda_{\bar{m}+1}^2 = 0 \). Therefore, AD\((0, \kappa)\) always holds trivially.

In the special case of Section 3.1, the spikes are all constants with \( \lambda_m^2 > \lambda_{m+1}^2 \). So condition AD\((m, \kappa)\) is satisfied for any constant \( \kappa \geq \lambda_{m+1}^2 / (\lambda_m^2 - \lambda_{m+1}^2) \).

**Rate of Convergence for Principal Subspace Estimation**

Recall that the thresholding parameters \( \alpha_n \) and \( \gamma_{nj} \) are specified in (3.2) and (3.3). 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 AD\((m, \kappa)\) holds for the given subspace dimension \( m \) and some \( \kappa \). Let \( \beta = c / \sqrt{m} \) in the definition (3.5) of \( H \), and let \( \alpha \) and \( \gamma \geq \gamma_0(c) \) in (3.2) and (3.3) be sufficiently large. Then, there exist constants \( C_0, C_1 = C_1(\gamma, r, m, \kappa) \)
and $C_2$, such that for sufficiently large $n$, with probability at least $1 - C_0 p_n^{-2}$, the subspace estimator $\hat{P}_m^{(k)} = \text{ran}(\hat{Q}^{(k)})$ in the $k$-th iteration of Algorithm 1 satisfies
\begin{equation}
\sup_{F_n} L(P_m, \hat{P}_m^{(k)}) \leq C_1 M_n \tau_{nm}^2 + C_2 \epsilon_{nm}^2 = o(1),
\end{equation}
for all $k = K, \ldots, 2K$, where
\begin{equation}
K = 1.01 \kappa \left( 1 + \frac{1}{\lambda_1^2} \right) \left[ \left( 1 + \frac{1}{\log 2} \right) \log n + 0 \lor \log h(\lambda_1^2) \right].
\end{equation}

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 $F_n$ 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 $F_n$.

The interpretation of the two terms in the error bound (3.8) 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, i.e., the “nonparametric”, term. By Theorem 3.3 below, when estimating $P_m$, Algorithm 1 focuses only on the coordinates in the set $H$, the cardinality of which is $\text{card}(H) = O(M_n)$. Since $\tau_{nm}^2$ can be interpreted as the average error per coordinate, the total estimation error accumulated over all the 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, i.e., the “parametric”, term is exactly the same as in Theorem 3.1. Finally, we note that both terms vanish as $n \to \infty$ under conditions GR, SP and AD($m, \kappa$).

The expression in (3.9) implies that Algorithm 1 only needs a relatively small number of iterations to yield the desired estimator. In particular, when the largest spike $\lambda_1^2$ is bounded away from zero, (3.9) implies that $K = O(\log n)$ iterations suffice.

The threshold levels $\alpha_n$ and $\gamma_{nj}$ in (3.2) and (3.3) does not depend on unknown parameters. Thus, similar to the special case, the estimation procedure is adaptive to a wide range of high-dimensional sparse settings. In the simulation studies, fixing $\alpha = 3$ in (3.2) and $\gamma = 1.5$ in (3.3) usually leads to satisfactory results. For details, see Section 5.

The result in Theorem 3.2 could also be extended to an upper bound for the risk. Note that $p_n^{-2} = o(\tau_{nm}^2 \lor \epsilon_{nm}^2)$, 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
\begin{equation}
\sup_{F_n} \mathbb{E} L(P_m, \hat{P}_m^{(k)}) \leq C_1 M_n \tau_{nm}^2 + C_2 \epsilon_{nm}^2.
\end{equation}

**Correct Exclusion Property**

We now switch to the model selection property of Algorithm 1. By our 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 orthonormal matrix with a lot of zeros, but Algorithm 1 is at liberty to introduce new non-zero coordinates. The following result shows that with high probability all the non-zero coordinates introduced are in the set $H$.

**Theorem 3.3.** Under the setup of Theorem 3.2, with probability at least $1 - C_0 p_n^{-2}$, for all $k = 0, \ldots, 2K$, the orthonormal matrix $\hat{Q}^{(k)}$ has zeros in all its rows indexed by $L$, i.e., $\hat{Q}^{(k)}_L = 0$.

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

**Rate of Convergence for Individual Eigenvector Estimation**

The primary focus of the current paper is on estimating principal subspaces. However, when an individual eigenvector, say $q_j$, is identifiable, it may also be of interest to see whether Algorithm 1 could estimate it well. The following result shows that for large enough $k$, the $j$-th column of $\hat{Q}^{(k)}$ estimates $q_j$ well, provided that the $j$-th eigenvalue is well separated from the rest of the spectrum.

**Corollary 3.2.** Under the setup of Theorem 3.2, if in addition for some $j \leq m$, both conditions $AD(j - 1, \kappa)$ and $AD(j, \kappa)$ hold, then with probability at least $1 - C_0 p_n^{-2}$, $\hat{q}_j^{(k)}$, the $j$-th column of $\hat{Q}^{(k)}$, satisfies

$$
\sup_{\mathcal{F}_n} L(\text{span}\{q_j\}, \text{span}\{\hat{q}_j^{(k)}\}) \leq C_1 M_n \tau_{nj}^2 + C_2 (\epsilon_{n,j-1}^2 + \epsilon_{nj}^2).
$$

Moreover, $\sup_{\mathcal{F}_n} \mathbb{E} L(\text{span}\{q_j\}, \text{span}\{\hat{q}_j^{(k)}\})$, the supremum risk over $\mathcal{F}_n$, is also bounded by the right side of the above inequality.

Corollary 3.2 connects closely to the previous investigation by Paul and Johnstone [24] on estimating individual sparse leading eigenvectors. Recall their loss function $l(q_j, \hat{q}_j) = \|q_j - \text{sgn}(q_j^\prime \hat{q}_j) \hat{q}_j\|_2^2$, which is equivalent to the restriction of the loss function (2.1) to one-dimensional subspaces, since $\frac{1}{2} l(q_j, \hat{q}_j) \leq L(\text{span}\{q_j\}, \text{span}\{\hat{q}_j\}) \leq l(q_j, \hat{q}_j)$. Then, Corollary 3.2 implies that

$$
\sup_{\mathcal{F}_n} \mathbb{E} l(q_j, \hat{q}_j^{(k)}) \leq C_1 M_n \tau_{nj}^2 + C_2 \frac{(\lambda_j^2 + 1)(\lambda_j^2 + 1)}{[(\lambda_j^2 - \lambda_j^2 - \lambda_j^2)^2] \wedge (\lambda_j^2 - \lambda_j^2 + 1)^2} \log(p_n) n.
$$

When the radii of the weak-$\ell_r$ balls grow at the same rate, i.e., $\max_j s_j / \min_j s_j = O(1)$, the upper bound in the last display matches the lower bound in Theorem 2 of [24] up to a logarithmic factor. Thus, when the $j$-th 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, since the thresholds $\alpha_n$ and $\gamma_{nj}$ do not depend on the unknown parameters.
4 Computational Complexity

In this section, we study the computational complexity of Algorithm 1. Throughout, we assume the same setup as in Section 3. In addition, we restrict the calculation to the high probability event on which the conclusions of Theorems 3.2 and 3.3 hold. In what follows, for any matrix \( A \), we use \( \text{supp}\{A\} \) to denote the index set of the rows of \( A \) which has at least one non-zero entry.

Consider a single iteration, say, the \( k \)-th. In the multiplication step, the \((\nu, j)\)-th element of \( T^{(k)} \), \( t_{\nu j}^{(k)} \), comes from the inner product of the \( \nu \)-th row of \( S \) and the \( j \)-th column of \( \hat{Q}^{(k-1)} \). Though both are \( p \)-vectors, Theorem 3.3 asserts that for any column of \( \hat{Q}^{(k-1)} \), at most \( \text{card}(H) \) of its entries are non-zero. Thus, if we know \( \text{supp}\{\hat{Q}^{(k-1)}\} \), then \( t_{\nu j}^{(k)} \) could be calculated in \( O(\text{card}(H)) \) flops, and \( T^{(k)} \) in \( O(mp\text{card}(H)) \) flops. Since \( \text{supp}\{\hat{Q}^{(k-1)}\} \) 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. Finally, consider the QR step. First, we can obtain \( \text{supp}\{\hat{T}^{(k)}\} \) in \( O(mp) \) flops. Then, QR factorization could be performed on the reduced matrix which only includes the rows in \( \text{supp}\{\hat{T}^{(k)}\} \). Since Theorem 3.3 implies \( \text{supp}\{\hat{T}^{(k)}\} = \text{supp}\{\hat{Q}^{(k)}\} \subset H \), the complexity of this step is \( O(m^2\text{card}(H)) \).

Since \( m = O(p) \), the complexity of the multiplication step dominates the other two steps, and so the complexity of each iteration is \( O(mp\text{card}(H)) \). Theorem 3.2 shows that \( K \) iteration is enough to achieve the error bound in (3.8). Therefore, the overall complexity of Algorithm 1 is \( O(Kmp\text{card}(H)) \).

When the true eigenvectors are sparse, \( \text{card}(H) \) is of manageable size. Moreover, in many realistic situations, \( \lambda_2^2 \) is bounded away from 0 and so \( K = O(\log n) \). 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 \( k \)-th iteration, both matrix multiplication and elementwise thresholding can be computed in parallel. For QR factorization, one needs only to communicate the rows of \( \hat{T}^{(k)} \) with non-zero elements, the number of which is no greater than \( \text{card}(H) \). Thus, the overhead from communication is \( O(m\text{card}(H)) \) for each iteration, and \( O(Km\text{card}(H)) \) in total. When the leading eigenvectors are sparse, \( \text{card}(H) \) is manageable, and parallel computing of Algorithm 1 is feasible.

5 Numerical Experiments

5.1 Single Spike Settings

We first consider single spike settings, where each observed data vector \( x_i \) is generated by (5.1) with \( \tilde{m} = 1 \). Motivated by functional data with localized features, four different test vectors \( q_k \) are considered, where \( q_1 = (f(1/p), \ldots, f(p/p))' \), with \( f \) one of the four functions in Fig. 1. For each test vector, the dimension \( p = 2048 \), the number of observations \( n = 1024 \), and the spike value \( \lambda_2^2 \) 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 \( \hat{\psi} \), and scale all the observations by \( \hat{\sigma} \) with \( \hat{\sigma}^2 \) given in (2.3). The multi-resolution plots of wavelet coefficients of the test vectors are shown in Fig. 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 by Algorithm 1 (ITSPCA) with...
Figure 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.

several existing methods in the literature, including augmented sparse PCA (AUGSPCA) by Paul and Johnstone \[24\], correlation augmented sparse PCA (CORSPCA) by Nadler \[20\], and diagonal thresholding sparse PCA (DTSPCA) given in Algorithm 2. For ITSPCA, we computed \( \hat{Q}^{(0)} \) using Algorithm 2. The thresholds \( \alpha_n \) and \( \gamma_{n1} \) are specified by (3.2) and (3.3) with \( \alpha = 3 \) and \( \gamma = 1.5 \). Moreover, we stopped the iteration once \( L(\text{ran}(\hat{Q}^{(k)}), \text{ran}(\hat{Q}^{(k+1)})) \leq 10^{-6} \). To make fair comparison, parameters in the competing algorithms are all set to the values recommended by their authors.

From Table 1, ITSPCA and CORSPCA outperform the other two methods for all test vectors and across different spike values. Between the two, CORSPCA only wins by small margin when the spike values are large. Otherwise, ITSPCA wins, sometimes with large margin. Moreover, at the same spike value and for the same algorithm, the sparser the signal, the smaller the estimation error.

Table 1 also presents the average sizes of the sets of selected coordinates. While all the four 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 Fig. 1, up to orthonormalization.\[1\] We tried four different configurations of the spike values \( (\lambda_2^1, \ldots, \lambda_2^4) \), as specified in the first column of Table 2. For each configuration of spike values, the dimension of data vectors is \( p = 2048 \), and the number of observations is \( n = 1024 \).

---

\[1\]These four test vectors are carefully shifted such that the inner product of any pair is close to 0. Thus, the vectors after orthonormalization are visually indistinguishable from those in Fig. 1.
Figure 2: Discrete wavelet transform of the four test vectors in Fig. 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.

| Test vector | ITSPCA | AUGSPCA | CORSPCA | DTSPCA |
|-------------|--------|---------|---------|--------|
|              | 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    |

Table 1: Comparison of sparse PCA methods in single spike settings: average loss in estimation and size of selected feature set.
For each simulated dataset, we estimate $P_m$ 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 yield decent estimators of $P_m$ for all values of $m$, which implies that the corresponding 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. On the other hand, all methods give reasonable estimators for $P_4$, for $\lambda_4^2$ in both cases are well above 0. Here, ITSPCA again gives the smallest average loss. This implies that, under such settings, we fail to recover individual eigenvectors, but can still estimate $P_4$ well.

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

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.

### Table 2: Comparison of sparse PCA methods in multiple spike settings: average loss in estimation.

| $(\lambda_1^2, \lambda_2^2, \lambda_3^2, \lambda_4^2)$ | $m$ | 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 |

6 Proof

This section is dedicated to the proof of the main results in Section 3.4. We first summarize 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.
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 high signal feature set \( H \), we construct a sequence of \( p \times m \) orthonormal matrices \( \{ \hat{Q}^{(k)}, k \geq 0 \} \). Then, for this sequence, we study how fast it converges, and how well each associated column subspace approximates the principal subspace \( P_m \) of interest. Finally, we show that, with high probability, this oracle sequence is exactly the sequence \( \{ \hat{Q}^{(k)}, k \geq 0 \} \) obtained by Algorithm 1. Therefore, the actual estimating sequence inherits from the oracle sequence various properties in terms of estimation error and number of steps needed to achieve the desired error rate. Here, the actual sequence mimics the oracle one because the thresholding step forces it to only consider the high signal coordinates.

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, \( \hat{Q}^{(k)} \), denotes the \( k \)-th orthonormal matrix in the oracle sequence.

Construction of The Oracle Sequence

First, we construct \( \hat{Q}^{(0)} \) using an oracle version of Algorithm 2, where the set \( B \) is replaced by its oracle version \( B^o = B \cap H \). This ensures that \( \hat{Q}^{(0)} \), \( L \cdot = 0 \). Here, \( \alpha_n \) is given by (3.2).

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

\[
S^o = \begin{bmatrix} S_{HH} & 0 \\ 0 & I_{LL} \end{bmatrix}.
\] (6.1)

Here, \( I_{LL} \) is the identity matrix of dimension \( \text{card}(L) \). Then, the matrices \( \{ \hat{Q}^{(k)}, k \geq 0 \} \) are obtained via an oracle version of Algorithm 1, in which the initial matrix is \( \hat{Q}^{(0)} \), and \( S \) is replaced by \( S^o \). Here, the \( \gamma_{nj} \)'s are specified as in (3.3).

Remark 6.1. The above formal construction does not guarantee that the \( \hat{Q}^{(k)} \) matrices have full column rank or that \( \hat{Q}^{(k)}_L = 0 \) for all \( k \). Later, we show in Lemma 6.3, Proposition 6.1 and Lemma 6.4 that these statements are true with high probability for all \( k \leq 2K \).

Major Steps of The Proof

We first introduce some notation. In the \( k \)-th iteration of the oracle Algorithm 1 denote the matrices obtained after the multiplication and the thresholding steps by

\[
\begin{align*}
T^{(k)} &= S^o \hat{Q}^{(k-1)}, \\
\hat{T}^{(k)} &= (\hat{t}^{(k)}_{j}), \quad \text{with} \quad \hat{t}^{(k)}_{j} = \eta(\hat{t}^{(k)}_{j}, \gamma_{nj}).
\end{align*}
\] (6.2)

Moreover, denote the QR factorization of \( \hat{T}^{(k)} \) by \( \hat{T}^{(k)} = \hat{Q}^{(k)} \hat{R}^{(k)} \). Last but not least, let \( \hat{P}^{(k)}_m = \text{ran}(\hat{Q}^{(k)}) \).

With the oracle sequence, a joint proof of Theorems 3.2 and 3.3 can be completed by the following three major steps:
1. Show that the principal subspace of $S^0$ with dimension $m$, denoted by $\hat{P}_m^o$, satisfies the error bound in (3.8) for estimating $P_m$;

2. Show that for $K$ in (3.9), after $K$ steps, the approximation error of $\hat{P}_m^{(k),o}$ to $\hat{P}_m^o$ also satisfies the bound in (3.8);

3. Show that the oracle sequence satisfies $\hat{Q}_L^{(k),o} = 0$ for all $k \leq 2K$, and that the oracle sequence and the actual estimating sequence 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 $\hat{P}_m^{(k),o}$ in estimating $P_m$ satisfies the bound in (3.8). Step 3 further shows this is also the case for the actual estimator $\hat{P}_m^{(k),m}$. In addition, it implies the correct exclusion property.

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

### 6.2 The Principal Subspace of $S^0$

To study how well the principal subspace of $S^0$ approximates $P_m$, we divide into a “bias” part and a “variance” part.

Consider the “bias” part first. Define the oracle covariance matrix

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

which is the expected value of $S^o$. The following lemma gives the error of the principal subspace of $\Sigma^o$ in approximating $P_m$, which could be regarded as the “squared bias” induced by feature selection.

**Lemma 6.1.** Let the leading eigenvalues of $\Sigma^o$ be $\ell_1^o \geq \cdots \geq \ell_m^o$, and $\{q_1^o, \ldots, q_m^o\}$ be a set of corresponding eigenvectors. Denote $Q^o = [q_1^o, \ldots, q_m^o]$. Then, uniformly over $F_n$,

1. $|\ell_j^o - (\lambda_j^2 + 1)/\lambda_1^2| \to 0$ as $n \to \infty$, for $j = 1, \ldots, m$;
2. For sufficiently large $n$, $Q_L^o = 0$, and for $P_m^o = \text{ran}(Q_o)$, there exists a constant $C = C(m, r, \kappa)$, s.t.

$$L(P_m, P_m^o) \leq CM_n \tau_{nm}^2.$$

A proof is given in the appendix. Weyl’s theorem [30, Corollary 4.4.10] and Davis-Kahn’s sin $\theta$ theorem [12] play the key roles in the proof here, and also in those for Lemmas 6.2 and 6.3.

Turn to the “variance” part. We look at how well the principal subspace of $S^0$ estimates that of $\Sigma^o$. Since $\Sigma^o = E[S^0]$, the error here is analogous to “variance”.

**Lemma 6.2.** Let the leading eigenvalues of $S^0$ be $\ell_1^o \geq \cdots \geq \ell_m^o$, and $\{\tilde{q}_1^o, \ldots, \tilde{q}_m^o\}$ be a set of corresponding eigenvectors. Denote $\tilde{Q}^o = [\tilde{q}_1^o, \ldots, \tilde{q}_m^o]$. Then, uniformly over $F_n$, with probability at least $1 - C_0 p_n^2$,

1. $|\ell_j^o - \ell_j^o|/\lambda_1^2 \to 0$ as $n \to \infty$, for $j = 1, \ldots, m$;
(2) For sufficiently large \( n \), \( \hat{Q}_L^0 = 0 \), and for \( \hat{P}_m^0 = \text{ran}(\hat{Q}^o) \), there exist constants \( C_1 = C_1(m,r,\kappa) \) and \( C_2 \), s.t.

\[
L(\hat{P}_m^0, \hat{P}_m^0) \leq \frac{C_1}{\log(pn)} M_n \tau_n^2 + C_2 \epsilon_{nm}^2.
\]

A proof is given in the appendix.

By the triangle inequality, the above two lemmas imply that the error in estimating \( \hat{P}_m \) with \( \hat{P}_m^0 \) satisfies the bound in (3.8), which completes step 1.

### 6.3 Properties of The Oracle Sequence

In step 2, we investigate the properties of the oracle sequence. The goal is to show that, with high probability, for all \( k \geq K \), the error of the oracle subspace estimator \( \hat{P}_m^{(k),o} \) in approximating \( \hat{P}_m^0 \) satisfies the bound in (3.8). To this end, characterization of the evolution of the oracle sequence in Proposition 6.1 plays the key role.

#### The Initial Point

We start with the initial point of the oracle sequence. The following lemma shows that \( \hat{Q}^{(0),o} \) 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. \( B^o = B \);
2. \( |\ell_j(S_{B^o B^o}) - \hat{\ell}_j^o|/\lambda_j^2 \to 0 \) as \( n \to \infty \), for \( j = 1, \ldots, \hat{m} \);
3. For sufficiently large \( n \), \( \hat{Q}^{(0),o} \) has full column rank, and \( L(\hat{P}_m^0, \hat{P}_m^{(0),o}) \leq (1 - \rho)^2/5 \).

A proof is given in the appendix. Claims (1) and (2) here, together with Lemmas 6.1 and 6.2, also imply that \( \ell_j(S_{BB})/(\lambda_j^2 + 1) \to 1 \) as \( n \to \infty \). So, the \( \ell_j(S_{BB})'s \) are good estimators of the leading eigenvalues of \( \Sigma \).

#### 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 \( \hat{P}_m^o \) and \( \hat{P}_m^{(k),o} \). By the discussion after (2.1), we have

\[
\sin^2 \theta^{(k)} = L(\hat{P}_m^0, \hat{P}_m^{(k),o}).
\]

In addition, let

\[
\rho = \frac{\hat{\ell}_{m+1}^o}{\hat{\ell}_m^o}
\]

denote the ratio between the \((m+1)\)-th and the \( m \)-th largest eigenvalues of \( S^o \). 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 \( k \geq 1 \),
of both aspects, we say that $\hat{\theta}(k)$ decreases. For a given $a$ below.

$$L \text{ is almost the smallest possible value of } L(\hat{\mathcal{P}}_m, \hat{\mathcal{P}}_m^{(k), o}) \text{.}$$

Under this definition, the following proposition shows that it takes some iterations for the oracle sequence to converge, and for all $k \geq K$, the error of approximating $\hat{\mathcal{P}}_m^o$ by $\hat{\mathcal{P}}_m^{(k), o}$ satisfies $\leq \epsilon_{nm}^2$.

A proof is given in the appendix, the key ingredient of which is Wedin’s sin $\theta$ theorem for singular subspaces.

The recursive inequality (6.6) characterizes the evolution of the angles $\theta(k)$, and hence of the oracle subspace $\hat{\mathcal{P}}_m^{(k), o}$. It is the foundation of claim (2) in the current proposition and of Proposition 6.2 below.

By (6.4), the inequality (6.8) gives the rate at which the approximation error $L(\hat{\mathcal{P}}_m^o, \hat{\mathcal{P}}_m^{(k), o})$ decreases. For a given $a \in (0, 1/2]$, the decrease 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.

Remark 6.2. 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 $\{\hat{\mathcal{P}}_m^{(k), o}, k \geq 0\}$, we first note that $1.01\omega^2/(1 - \rho)^2$ is almost the smallest possible value of $L(\hat{\mathcal{P}}_m^o, \hat{\mathcal{P}}_m^{(k), o})$ 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

$$\sin \theta(k) \leq (\rho \sin \theta(k) + \omega)(1 + o(1)).$$

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

$$L(\hat{\mathcal{P}}_m^o, \hat{\mathcal{P}}_m^{(k), o}) \leq \max\left\{\frac{1.01}{(1 - n^{-1})^2}, \frac{\omega^2}{(1 - \rho)^2}, \epsilon_{nm}^2\right\}. \quad (6.9)$$

Under this definition, the following proposition shows that it takes $K$ iterations for the oracle sequence to converge, and for all $k \geq K$, the error of approximating $\hat{\mathcal{P}}_m^o$ by $\hat{\mathcal{P}}_m^{(k), o}$ satisfies $\leq \epsilon_{nm}^2$.

---

2Here, the introduction of the multiplier $(1 - n^{-1})^{-2}$ is only for mathematical convenience.
Proposition 6.2. Let $K$ be specified in (3.9). For sufficiently large $n$, on the event such that the conclusions of Lemmas 6.1–6.4 hold, uniformly over $F_n$, 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 $k \geq K$,

$$\sup_{F_n} L(\hat{P}^0_m, \hat{P}^{(k)}_m) \leq C_1 M_n \tau_{nm}^2 + C_2 \varepsilon_{nm}^2. \quad (6.10)$$

A proof is given in the appendix, and this completes step 2.

6.4 Proof of Main Results

Here we come back to prove the properties of the actual estimating sequence. The proof relies on the following lemma, which establishes the equivalence of the actual sequence to the oracle one up to $2K$ iterations.

Lemma 6.4. Let $\beta = c/\sqrt{m}$ in (3.5). There exist a constant $\gamma_0 = \gamma_0(c)$, such that if we set $\gamma_{nj}$ as in (3.3) for some $\gamma \geq \gamma_0$, then for sufficiently large $n$, with probability at least $1 - C_0 p_n^{-2}$, for $k = 0, 1, \ldots, 2K$, we have $\hat{Q}^{(k), o}_L = 0$, $\hat{Q}^{(k)} = \hat{Q}^{(k), o}$, and hence $\hat{P}^{(k)}_m = \hat{P}^{(k), o}_m$.

A proof is given in the appendix, and this completes step 3.

Proof of Theorem 3.2. Note that the event on which the conclusions of Lemmas 6.1–6.4 all hold has probability at least $1 - C_0 p_n^{-2}$. On this event, we have, for $k = K, \ldots, 2K$,

$$L(P_m, \hat{P}^{(k)}_m) = L(P_m, \hat{P}^{(k), o}_m) \leq \left[ L^{1/2}(P_m, P^o_m) + L^{1/2}(P^o_m, \hat{P}^o_m) + L^{1/2}(\hat{P}^o_m, \hat{P}^{(k), o}_m) \right]^2 \leq C \left[ L(P_m, P^o_m) + L(P^o_m, \hat{P}^o_m) + L(\hat{P}^o_m, \hat{P}^{(k), o}_m) \right] \leq C_1 M_n \tau_{nm}^2 + C_2 \varepsilon_{nm}^2.$$

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 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 conclusion of Lemmas 6.1–6.4 all hold. Then Lemma 6.4 directly leads to the conclusion that $\hat{Q}^{(k)}_{L^c} = \hat{Q}^{(k), o}_{L^c} = 0$.

We conclude the section by proving Corollaries 3.1 and 3.2.

Proof of Corollary 3.1. Under conditions GR, SP and AD$(m, \kappa)$, we have $p_n^{-2} = o(\tau_{nm}^2 \vee \varepsilon_{nm}^2)$, and so $p_n^{-2} = o(C_1 M_n \tau_{nm}^2 + C_2 \varepsilon_{nm}^2)$. In addition, we note that the loss function (2.1) is always bounded above by 1. Let $E$ denote the event on which the conclusions of Lemmas 6.1–6.4 hold. Then we have

$$EL(P_m, \hat{P}^{(k)}_m) = EL(P_m, \hat{P}^{(k)}_m) 1_E + EL(P_m, \hat{P}^{(k)}_m) 1_{E^c} \leq \sup_{E} L(P_m, \hat{P}^{(k)}_m) P(E) + P(E^c) \leq C_1 M_n \tau_{nm}^2 + C_2 \varepsilon_{nm}^2.$$

This completes the proof.

22
Lemma A.1. The quantities $\|q_j\|_2^2$ and card$(H)$ satisfy the following bounds:

1. $\|q_j\|_2^2 \leq \frac{2}{\tau_{nj}^2} s_j^2 (\beta \tau_{nj})^{-r} \wedge p (\beta \tau_{nj})^2$, $j = 1, \ldots, m$;
2. When $n$ is sufficiently large, $\bar{m} \leq \text{card}(H) \leq C M_n$, where $C = C(m, r)$.

A.1 Proof of Lemma 6.1

We use Weyl’s theorem (Corollary 4.4.10 in [30]) to prove claim (1), and Davis-Kahn’s sinθ theorem (Theorem 3.11) to prove claim (2). In addition, the following lemma is useful.

Lemma A.1. The quantities $\|q_j\|_2^2$ and card$(H)$ satisfy the following bounds:

1. $\|q_j\|_2^2 \leq \frac{2}{\tau_{nj}^2} s_j^2 (\beta \tau_{nj})^{-r} \wedge p (\beta \tau_{nj})^2$, $j = 1, \ldots, m$;
2. When $n$ is sufficiently large, $\bar{m} \leq \text{card}(H) \leq C M_n$, where $C = C(m, r)$.
Proof. For claim (1), we apply the “ideal risk calculation trick” in [14]. Let \( t_j = t_j(n) \) be the solution of \( s_j t^{-1/2} = \beta \tau_{nj} \). We have
\[
\|q_jL\|_2^2 = \sum_{\nu \in L} q_{\nu j}^2 = \sum_{\nu \in L} \min\{q_{\nu j}^2, \beta^2 \tau_{nj}^2\} = \sum_{\nu \in L} \min\{|q_j|_{(\nu)}, \beta^2 \tau_{nj}^2\}
\leq \int_0^\infty \min\{\beta^2 \tau_{nj}^2, s_j^2 t^{-2r}\} dt \leq t_j \beta^2 \tau_{nj}^2 + \frac{r}{2 - r} s_j^2 t_j^{-1/2} = \frac{2}{2 - r} s_j^2 (\beta \tau_{nj})^{2-r}.
\]
On the other hand, by the definition of \( L \), we always have the bound \( \|q_jL\|_2^2 \leq \text{card}(L) \cdot \max_{\nu \in L} \{q_{\nu j}^2\} \leq p(\beta \tau_{nj})^2 \). Recall the definition of \( M_n \), we complete the proof of (1).

For claim (2), we first show that \( \text{card}(H) \leq CM_n \). To this end, for \( j = 1, \ldots, \tilde{m} \), let
\[
H_j = \{\nu : |q_{\nu j}| \geq \beta \tau_{nj}\}.
\]
Note that \( H = \bigcup_{j=1}^{\tilde{m}} H_j \). So, \( \text{card}(H) \leq \sum_{j=1}^{\tilde{m}} \text{card}(H_j) \). For each \( H_j \), the weak-\( \ell_r \) constraint implies that \( \text{card}(H_j) \leq s_j^2 (\beta \tau_{nj})^{-r} \). So, \( \text{card}(H) \leq p \wedge \sum_{j=1}^{\tilde{m}} s_j^2 (\beta \tau_{nj})^{-r} \leq C M_n \). Since \( \beta = c/\sqrt{m} \), we get \( C = C(m, r) \).

To show \( \text{card}(H) \geq \tilde{m} \), it suffices to show that \( \bar{Q}_H \) has full column rank. Note that \( \bar{Q} \) is orthonormal. So, this is equivalent to showing \( \|Q_L\| < 1 \). For \( \|Q_L\| \), claim (1) implies that
\[
\|Q_L\| \leq \sum_{j=1}^{\tilde{m}} \|q_jL\|_2^2 \leq \sum_{j=1}^{\tilde{m}} \left[ p \wedge \frac{2}{2 - r} s_j^2 (\beta \tau_{nj})^{-r} \right] (\beta \tau_{nj})^2 = o(1).
\]
Here, the last equality comes from condition SP. So, when \( n \) is large, \( \|Q_L\| < 1 \), which completes the proof of (2).

Proof of claim (1) Weyl’s theorem states that \( |\ell_j^o - (\lambda_j^2 + 1)| \leq \|\Sigma^o - \Sigma\| \), for all \( j \). Note that \( \Sigma^o - \Sigma = \begin{bmatrix} 0 & -\Sigma_{HL} \\ -\Sigma_{LH} & I_{LL} - \Sigma_{LL} \end{bmatrix} \), where \( \Sigma_{HL} = Q_H \Lambda_2 \bar{Q}_L^t \) and \( \Sigma_{LL} - I_{LL} = \bar{Q}_L \Lambda_2 \bar{Q}_L^t \). Thus, \( \|\Sigma^o - \Sigma\| \leq \|\Sigma_{HL}\| + \|\Sigma_{LL} - I_{LL}\| \). Note that \( \|q_jH\|, \|q_jL\| \leq 1 \), and that Lemma A.1 together with condition SP, implies that \( \sum_{j=1}^{\tilde{m}} \|q_jL\|_2 = o(1) \). Thus, we obtain
\[
\|\Sigma_{HL}\|, \|\Sigma_{LL} - I_{LL}\| \leq \lambda_j^2 \sum_{j=1}^{\tilde{m}} \|q_jL\|_2 = o(\lambda_j^2).
\]
Therefore, \( |\ell_j^o - (\lambda_j^2 + 1)| \leq \|\Sigma^o - \Sigma\| = o(\lambda_j^2) \), for \( j = 1, \ldots, \tilde{m} \).

Proof of claim (2) First, claim (1) implies that, when \( n \) is large, \( \ell_m^o \geq \lambda_m^2 + 1 - o(\lambda_m^2) > 1 \). Thus, the \( m \) leading eigenvectors of \( \Sigma^o \) could be constructed by first taking the \( m \) leading eigenvectors of \( \Sigma_{HH} \), and then augmenting all the coordinates in \( L \) with zeros. Consequently, \( Q_0^o = 0 \).

To further obtain the bound on \( L(P_m, P_m^o) \), we apply Davis-Kahn’s \( \sin \theta \) theorem (Theorem B.1). In Theorem B.1 let \( A = \Sigma, A + E = \Sigma^o, G_0 = Q, \) and \( F_0 = Q^o \). Since \( G_0 \) is orthonormal, the theorem states that
\[
L(P_m, P_m^o) \leq \frac{\|Q_0^o \Sigma^o - \Sigma\|^2}{(\lambda_m^2 + 1 - \ell_m^o)^2}, \quad (A.2)
\]
where $Q^o_c$ is a $p \times (p - m)$ orthonormal matrix whose columns are orthogonal to those in $Q^o$. The subscript 'c' is used to indicate the range of $Q^o_c$ is the orthogonal complement to that of $Q^o$.

Consider the right side of (A.2). For the denominator, claim (1) and condition AD($m, \kappa$) imply

$$\left| (\lambda_n^2 + 1) - \epsilon_{m+1} \right| \geq \lambda_n^2 - \lambda_{m+1}^2 - o(\lambda_n^2) = (\lambda_n^2 - \lambda_{m+1}^2)(1 + o(1)).$$

(A.3)

For the numerator, since $Q^o_c$ is orthonormal, we can bound it as $\| (Q^o_c)'(\Sigma^o - \Sigma)Q \| \leq \| (\Sigma^o - \Sigma)Q \|$. Furthermore, we have

$$\langle (\Sigma^o - \Sigma)Q \rangle = \begin{bmatrix} E_{0,H}^1 \bar{E}_{0,H}^2 \end{bmatrix},$$

where $E_{0,H} = -Q_H \Lambda^2 Q_H, Q_L, \bar{E}_{0,H}^1 = -Q_L \Lambda^2 Q_H, Q_H$ and $E_{0,L}^2 = -Q_L \Lambda^2 Q_L, Q_L$. Thus, we obtain $\| (\Sigma^o - \Sigma)Q \| \leq \| E_{0,H}^1 \| + \| E_{0,L}^1 \|$. In what follows, we bound each of the three terms on the right side. Take $E_{0,L}^1$ for example. Let

$$A^2_0 = \text{diag}(\lambda_1^2, \ldots, \lambda_{m}^2), \quad A^2_2 = \text{diag}(\lambda_{m+1}^2, \ldots, \lambda_{m}^2), \quad Q_1 = [q_{m+1}, \ldots, q_m].$$

(A.4)

Then $-E_{0,L}^1 = Q_L A^2_0 Q_H, Q_H + Q_{1,L} A^2_2 Q_{1,H}, Q_H = Q_{1,L} A^2_0 Q_{1,H}, Q_H - Q_{1,L} A^2_2 Q_{1,L}, Q_L$. Here, the second equality comes from the identity $Q_{1,L}, Q_H + Q_{1,L}, Q_L = Q_{1,L}$. So,

$$\| E_{0,L}^1 \| \leq \| Q_L \| \| A^2_0 \| \| Q_H \| \leq \| Q_L \| \| A^2_0 \| (1 + o(1)) \leq \lambda_n^2 \| Q_L \|_{F}(1 + o(1)).$$

Similar arguments show that $\| E_{0,H}^1 \|, \| E_{0,L}^2 \| = o(\lambda_n^2 \| Q_L \|_{F})$. Thus, we obtain that

$$\| (Q^o_c)'(\Sigma^o - \Sigma)Q \| \leq \| (\Sigma^o - \Sigma)Q \| \leq \lambda_n^2 \| Q_L \|_{F}(1 + o(1)).$$

(A.5)

By (A.2), (A.3), we obtain that

$$L(P_m, \rho^o_m) \leq \frac{\lambda_1^2 \| Q_L \|_{F}^2(1 + o(1))}{(\lambda_1^2 - \lambda_{m+1}^2)^2(1 + o(1))} \leq \left( \frac{\lambda_1^2}{\lambda_1^2 - \lambda_{m+1}^2} \right)^2 \sum_{j=1}^{m} \| q_{jL} \|_{2}^2 (1 + o(1)).$$

Since $\lambda_n^2 / (\lambda_n^2 - \lambda_{m+1}^2) \leq \kappa$ under condition AD($m, \kappa$), we complete the proof by bounding the $\| q_{jL} \|_{2}^2$’s using Lemma A.1.

### A.2 Proof of Lemma 6.2

We prove Lemma 6.2 by a similar approach to that for Lemma 6.1. The major difference here is that we are dealing with a random matrix $S^o$ instead of a deterministic one.

Many of the bounds to appear in the proof involve the following two quantities:

$$\zeta(n, l) = 2\sqrt{\frac{l}{n} + \frac{l}{n} + 6 \sqrt{\frac{2 \log(p_n)}{n}}},$$

(A.6)

$$\xi(n, l, m; b) = \sqrt{1 + \frac{4 \log(p_n)}{n}} \left( \sqrt{\frac{l}{n} + \frac{m}{n} + b \sqrt{\frac{\log(p_n)}{n}}} \right).$$

(A.7)

Here, $l, m, n \in \mathbb{N}$ and $b > 0$. By Proposition B.1, $\zeta(n, l)$ gives a probabilistic bound on the spectral norm of the difference between a random (scaled) Wishart matrix $A$ and its mean $I$. On the other hand, Proposition B.2 shows that $\xi(n, l, m; b)$ is a probabilistic bound on the spectral norm of $Y'Z/n$, where $Y$ and $Z$ are independent $n \times l$ and $n \times m$ matrices with i.i.d. $N(0, 1)$ entries. For controlling these quantities, the following lemma will be used multiple times.
**Lemma A.2.** There exist constants $C_1(\beta, r, \kappa), C'_1(\beta, r, \kappa), C_2$ and $C'_2$, such that when $n$ is large,

$$
\zeta(n, \text{card}(H)) \leq C_1 \frac{\lambda^2 M_n^{1/2} \tau_{nm}}{\sqrt{\log(p_n)}} + C_2 \sqrt{\frac{\log(p_n)}{n}},
$$

$$
\xi(n, \bar{m}, \text{card}(H); 2) \leq C'_1 \frac{\lambda^2 M_n^{1/2} \tau_{nm}}{\sqrt{\lambda_1^2 + 1} \log(p_n)} + C'_2 \frac{\log(p_n)}{n}.
$$

**Proof.** By Lemma A.1, we have

$$
\frac{\text{card}(H)}{n} \leq \frac{CM_n}{n} = CM_n \frac{\lambda^2_m}{\sqrt{\log(p_n)}} \leq C M_n \tau_{nm}^2 \frac{\lambda^4_1}{(\lambda_1^2 + 1) \log(p_n)},
$$

Here, the second inequality comes from condition AD($m, \kappa$). Therefore, we obtain

$$
\zeta(n, \text{card}(H)) \leq \frac{2C^{1/2} \lambda^2 M_n^{1/2} \tau_{nm}}{\sqrt{\log(p_n)}} + C M_n \tau_{nm}^2 \frac{\lambda^4_1}{(\lambda_1^2 + 1) \log(p_n)} + C_2 \sqrt{\frac{\log(p_n)}{n}}\tag{A.8}
$$

$$
= \frac{C^{1/2} \lambda^2 M_n^{1/2} \tau_{nm}}{\sqrt{\log(p_n)}} \left( 2 + \frac{\lambda^2_1}{(\lambda_1^2 + 1) \sqrt{\log(p_n)}} \right) + C_2 \frac{\log(p_n)}{n}.
$$

The bound on $\xi(n, \bar{m}, \text{card}(H); 2)$ follows from similar arguments. \hfill \square

To facilitate the arguments, note that the difference $S^o - \Sigma^o = \begin{bmatrix} E_{HH} & 0 \\ 0 & 0 \end{bmatrix}$, and we decompose $E_{HH}$ as $E_{HH} = \sum_{k=1}^3 E_k$, where

$$
E_1 = \bar{Q}_H \Lambda \left( \frac{1}{n} V' V - I \right) \Lambda Q'_H, \quad E_2 = \frac{1}{n} Z'_H Z_H - I_{HH},
$$

$$
E_3 = \frac{1}{n} (\bar{Q}_H \Lambda V' Z_H + Z'_H V \Lambda Q'_H).
$$

**Proof of claim (1)** As before, Weyl’s theorem implies $|\hat{\ell}_j^o - \ell_j^o| \leq ||S^o - \Sigma^o|| = ||E_{HH}|| \leq \sum_{k=1}^3 ||E_k||$. In what follows, we derive bounds for each of the $||E_k||$’s.

Consider $E_1$ first. Since $||\bar{Q}_H|| \leq 1$ and $||\Lambda|| \leq \lambda_1$, on the event

$$
\left\{ ||\frac{1}{n} V' V - I || \leq \zeta(n, \bar{m}) \right\},
$$

we obtain $||E_1|| \leq ||\bar{Q}_H||^2 ||\Lambda||^2 \frac{1}{n} ||V' V - I|| \leq \lambda^2_1 \zeta(n, \bar{m})$. Next, for $E_2$, we immediately get the bound $||E_2|| \leq \zeta(n, \text{card}(H))$ on the event

$$
\left\{ ||\frac{1}{n} Z'_H Z_H - I_{HH} || \leq \zeta(n, \text{card}(H)) \right\}.
$$

Finally, for $E_3$, since $||\bar{Q}_H|| \leq 1$ and $||\Lambda|| \leq \lambda_1$, on the event

$$
\left\{ ||V' Z_H || \leq n \zeta(n, \bar{m}, \text{card}(H); 2) \right\},
$$

26
we have \( \|E_3\| \leq \frac{2}{n} \|\tilde{Q}_H\| \|\Lambda\| \|V'Z_H\| \leq 2\lambda_1(n, m, \text{card}(H); 2) \). Combine these bounds and apply Lemma A.2. We obtain

\[
|\hat{\theta}_j^0 - \theta_j^0| \leq \lambda_1^2 \left( C_1 \frac{M_1^{1/2} \tau_{nm}}{\sqrt{\log(p_n)}} + C_2 \sqrt{\frac{\log(p_n)}{n(\lambda_1^2 \wedge 1)^2}} \right) = o(\lambda_1^2).
\]

Here, the last equality comes from conditions GR and SP. Furthermore, observe that Propositions B.1 and B.2 imply the events in (A.9), (A.10) and (A.11) occur with probability at least \( 1 - C_0 p_n^{-2} \), and so does the intersection of them. This completes the proof of claim (1).

**Proof of claim (2)** We use Theorem B.1 again. In the theorem, let \( A = \Sigma^o \), \( A + E = S^o \), \( \Lambda_o = Q^o \), \( F_0 = \tilde{Q}^o \). Then \( F_1 = \tilde{Q}^o_c \), where \( \tilde{Q}^o_c \) is a \( p \times (p - m) \) orthonormal matrices whose columns are orthogonal to those in \( \tilde{Q}^o \). As before, the subscript ‘c’ is used to indicate the range of \( \tilde{Q}^o_c \) is the orthogonal complement to that of \( \tilde{Q}^o \). With this setup, Theorem B.1 asserts that

\[
L(P^o_m, \tilde{P}^o_m) \leq \frac{\|\tilde{Q}^o_c \|^2 (S^o - \Sigma^o) Q^o}{(\tilde{t}^o_m - \tilde{t}^o_{m+1})^2}.
\]

Consider the right side of (A.12). For the denominator, claim (1), Lemma 6.1 and condition AD\((m, \kappa)\) imply that \( |\tilde{t}^o_m - \tilde{t}^o_{m+1}| \geq \lambda_2^2 - \lambda_1^2 - o(\lambda_1) = (\lambda_2^2 - \lambda_1^2)(1 + o(1)) \). For the numerator, (A.8) leads to the decomposition:

\[
(\tilde{Q}^o_c)^\prime (S^o - \Sigma^o) Q^o = (\tilde{Q}^o_c)^\prime E_{HH} \tilde{Q}^o_H. = \sum_{k=1}^{3} (\tilde{Q}^o_c)^\prime E_k \tilde{Q}^o_H. = \sum_{k=1}^{3} E_k^o.
\]

Here, \( \tilde{Q}^o_{c,H} \) denotes the submatrix of \( \tilde{Q}^o_c \) that contains all its rows in \( H \). Thus, the triangle inequality leads to an upper bound of the numerator if we bound each \( \|E_k^o\| \), which we carry out below.

For \( E_k^o \), note that \( E_k^o = (\tilde{Q}^o_c)^\prime \tilde{Q}^o_{H,A} (\frac{1}{n} V'V - I) \Lambda \tilde{Q}^o_{H,A} \). Since \( \zeta(n, \bar{m}) \leq C \sqrt{\log(p_n)/n} \), \( \|\Lambda\| \leq \lambda_1 \) and \( \|\tilde{Q}^o_H\|, \|\tilde{Q}^o_{H,A}\| \leq 1 \), on the event \( (A.9) \), we have

\[
\|E_k^o\| \leq \frac{\|\tilde{Q}^o_c\|}{n} \frac{\|\tilde{Q}^o_{H,A}\|}{n} \frac{1}{n} \|V'V - I\| \|\Lambda\| \|\tilde{Q}^o_H\| \|\tilde{Q}^o_{H,A}\|
\]

\[
\leq C_\lambda \sqrt{\log(p_n)/n} \|\tilde{Q}^o_{c,H}\| \|\tilde{Q}^o_{H,A}\|.
\]

To further control the last term on the right side, we apply an “orthogonal decomposition” trick. Observe that \( \tilde{Q}^o_H \) is a card\((H) \times m \) orthonormal matrix, because \( Q^o \) is \( p \times m \) orthonormal with \( Q^o_L = 0 \). Define \( P^o_H \) as a card\((H) \times (\text{card}(H) - m) \) orthonormal matrix whose columns are orthogonal to those in \( Q^o_H \), we obtain the decomposition

\[
I_{HH} = Q^o_H (Q^o_{H})' + P^o_H (P^o_{H})'.
\]

Thus \( (\tilde{Q}^o_c)^\prime \tilde{Q}^o_{H,A} = (\tilde{Q}^o_c)^\prime Q^o_{H,A} (Q^o_{H,A})' \tilde{Q}^o_{H,A} + (\tilde{Q}^o_c)^\prime P^o_H (P^o_{H})' \tilde{Q}^o_{H,A} \). The first term on the right side can be bounded as

\[
\| (\tilde{Q}^o_c)^\prime Q^o_{H,A} (Q_{H,A})' \| \leq \lambda_1 \| (\tilde{Q}^o_c)^\prime Q^o_{H,A} \| \leq \lambda_1 \| (\tilde{Q}^o_c)^\prime Q^o \|.
\]
Here, the first inequality holds because $\|Q^0_H\|, \|\hat{Q}_H\| \leq 1$ and $\|A\| \leq \lambda_1$, and the last equality comes from the observation that $Q^0_L = 0$. By (A.4), the second term in the last display could be further decomposed as $(\hat{Q}^0_{c,H})'P^0_H(P^0_H)'Q_H\Lambda_0 + (\hat{Q}^0_{c,H})'P^0_H(Q^0_{1,H}\Lambda_1$. Note that $\|P^0_H\|Q_H\| \leq \|(Q^0_c)'Q\|, \|A_0\| \leq \lambda_1, \|A_1\| \leq \lambda_{m+1}$, and the spectrum norms of all the other matrices are uniformly bounded above by 1. Thus, we obtain that

$$\|(\hat{Q}^0_{c,H})'P^0_H(P^0_H)'\hat{Q}_H\Lambda\| \leq \lambda_1\|(Q^0_c)'Q\| + \lambda_{m+1}. \tag{A.15}$$

Combining the parts leads to

$$\|(\hat{Q}^0_{c,H})'\hat{Q}_H\Lambda\| \leq \lambda_{m+1} + \lambda_1\left(\|\hat{Q}^0_{c,H}'\| + \|(Q^0_c)'Q\|\right), \tag{A.15}$$

which in turn implies $E_0^1 \leq C\left[\lambda_1\lambda_{m+1} + \lambda_1^2\left(\|\hat{Q}^0_{c,H}'\| + \|(Q^0_c)'Q\|\right)\right]\sqrt{\log(p_n)/n}$. Turn to $E_0^2$. (A.14) leads the decomposition $E_0^2 = E_{21}^0 + E_{22}^0$, where

$$E_{21}^0 = (\hat{Q}^0_{c,H})'Q_H^0\left[\frac{1}{n}(Z_HQ_H^0)'(Z_HQ_H^0) - I_m\right], \quad E_{22}^0 = (\hat{Q}^0_{c,H})'P^0_H(P^0_H)'\frac{1}{n}Z_H(Z_HQ_H^0).$$

For $E_{21}^0$, recall that $Q^0_L = 0$. Thus, $\|(\hat{Q}^0_{c,H})'Q^0_H\| = \|(\hat{Q}^0_{c,H})'Q^0\|$. So, on the event

$$\{\|\frac{1}{n}(Z_HQ_H^0)'(Z_HQ_H^0) - I_m\| \leq \zeta(n,m)\}, \tag{A.16}$$

we obtain $E_{21}^0 \leq \|(\hat{Q}^0_{c,H})'Q^0\|\zeta(n,m) \leq C\|(\hat{Q}^0_{c,H})'Q^0\|\sqrt{\log(p_n)/n}$. For $E_{22}^0$, on the event

$$\{\||Z_HP^0_H)'(Z_HQ^0_H)\| \leq n\xi(n,\bar{m},\text{card}(H);2)\}, \tag{A.17}$$

we obtain $E_{22}^0 \leq \xi(n,\bar{m},\text{card}(H);2) \leq \xi(n,\bar{m},\text{card}(H);2)$. Combine the two parts and apply Lemma (A.2) We obtain

$$\|E_2^0\| \leq C_1\frac{\lambda_1^2M_1^{1/2}m}{\sqrt{(\lambda_1^2 + 1)\log(p_n)}} + C_2\left(1 + \|\hat{Q}^0_{c,H}'\|\right)\sqrt{\log(p_n)/n}. \tag{A.18}$$

For $E_3^0$, (A.8) and (A.13) lead to $E_3^0 = E_{31}^0 + E_{32}^0$, where

$$E_{31}^0 = \frac{1}{n}(\hat{Q}^0_{c,H})'V^*Z_HQ^0_H, \quad E_{32}^0 = \frac{1}{n}(\hat{Q}^0_{c,H})'Z_H^*V\Lambda\hat{Q}^0_HQ^0_H.$$ 

For $E_{31}^0$, on the event

$$\{\|V'Z_HQ^0_H\| \leq n\xi(n,\bar{m},m;2)\}, \tag{A.18}$$

(A.15) and (A.7) imply

$$\|E_{31}^0\| \leq \frac{1}{n}\|(\hat{Q}^0_{c,H})'\hat{Q}_H\Lambda\||V'Z_HQ^0_H|| \leq \lambda_{m+1} + \lambda_1\left(\|\hat{Q}^0_{c,H}'\| + \|(Q^0_c)'Q\|\right)\xi(n,\bar{m},m;2) \leq C\left[\lambda_{m+1} + \lambda_1\left(\|\hat{Q}^0_{c,H}'\| + \|(Q^0_c)'Q\|\right)\right]\sqrt{\log(p_n)/n}. \tag{A.18}$$
In addition, on the event \( \{ \text{A.11} \} \), we obtain \( \| E_{32}^0 \| \leq \lambda_1 \xi(n, \hat{m}, \text{card}(H); 2) \). Then the triangle inequality and Lemma \[ \text{A.2} \] lead to
\[
\| E_{32}^0 \| \leq C_1 \frac{\lambda_1^3 M_n^{1/2} \tau_{nm}}{\sqrt{(\lambda_1^2 + 1) \log(p_n)}} + \left[ C_2 (\lambda_1 + \lambda_{m+1}) + C_3 \lambda_1 \left( \left\| (\hat{Q}_c^0)' Q^0 \right\| + \left\| (Q_c^0)' Q \right\| \right) \right] \sqrt{\frac{\log(p_n)}{n}}.
\]

Summing up the above bounds for the \( \| E_k^0 \| \)'s leads to
\[
L^{1/2}(P_m^0, \hat{P}_m^0) = \left\| (\hat{Q}_c^0)' Q^0 \right\| \leq \sum_{k=1}^3 \| E_k^0 \| \leq C_1 \frac{\lambda_1^2 (\lambda_1 + 1)M_n^{1/2} \tau_{nm}}{\lambda_m^2 - \lambda_{m+1}^2 \sqrt{(\lambda_1^2 + 1) \log(p_n)}} + C_2 \frac{\lambda_1 \lambda_{m+1} + \lambda_1 + \lambda_{m+1} + 1}{\lambda_m^2 - \lambda_{m+1}^2} \sqrt{\frac{\log(p_n)}{n}} + \alpha \left( \left\| (\hat{Q}_c^0)' Q^0 \right\| \right).
\]

Condition AD(\( m, \kappa \)) implies \( \lambda_1^2 / (\lambda_m^2 - \lambda_{m+1}^2) \leq \kappa \). We collect terms and apply Jensen’s inequality to obtain
\[
L(P_m^0, \hat{P}_m^0) = \left\| (\hat{Q}_c^0)' Q^0 \right\|^2 \leq C_1 M_n^{1/2} \tau_{nm} \log(p_n) + C_2 \frac{(\lambda_1^2 + 1)(\lambda_m^2 + 1) \log(p_n)}{n}.
\]

Since \( Z_H Q_H^0 \) is a \( n \times m \) matrix with i.i.d. \( N(0, 1) \) entries, independent of the matrices \( Z_H P_m^0 \) and \( V \) which also have i.i.d. \( N(0, 1) \) entries, Propositions \[ \text{A.1} \] and \[ \text{A.2} \] thus imply that the events in \( \{ \text{A.16} \}, \{ \text{A.17} \} \) and \( \{ \text{A.18} \} \) occur with probability at least \( 1 - C_0 p_n^{-2} \), and so does the intersection of them. This completes the proof.

### A.3 Proof of Lemma \[ \text{6.3} \]
We first introduce some preliminary results. The first result shows that, with high probability, the set \( B \) selects all the “big” coefficients in the leading eigenvectors up to order \( O(\sqrt{\log(p_n)/(n\lambda_1^2)}) \).

Indeed, for \( 0 < a_- < 1 < a_+ \), define
\[
B_\pm = \left\{ \nu : \sum_{j=1}^{\hat{m}} \lambda_j^2 q_{\nu j}^2 > \alpha \sqrt{\frac{\log(p_n)}{n}} \right\}.
\]

**Lemma A.3.** Let \( B_\pm \) be defined as above. For appropriately chosen \( \alpha \) and \( a_\pm \), when \( n \) is sufficiently large, with probability at least \( 1 - C_0 p_n^{-2} \), \( B_- \subset B \subset B_+ \subset H \).

**Proof.** Note that \( S_{\nu \nu} \sim \sigma_{\nu}^2 \chi_n^2 / n \), where \( \sigma_{\nu}^2 = 1 + \sum_{j=1}^{\hat{m}} \lambda_j^2 q_{\nu j}^2 \). Consider the event \( \{ B_- \subset B \} \).

We have
\[
P \{ B_- \not\subset B \} = P \left\{ \bigcup_{\nu \in B_-} \{ S_{\nu \nu} < 1 + a_+ \sqrt{\log(p_n)/n} \} \right\} \leq \sum_{\nu \in B_-} P \{ S_{\nu \nu} < 1 + a_+ \sqrt{\log(p_n)/n} \}
\]
\[
\leq \sum_{\nu \in B_-} P \left\{ \frac{S_{\nu \nu}}{\sigma_{\nu}^2} \leq \frac{1 + a_+ \sqrt{\log(p_n)/n}}{1 + a_+ \sqrt{\log(p_n)/n}} \right\} \leq \sum_{\nu \in B_-} P \left\{ \frac{\lambda_{\nu}^2}{n} - 1 \leq - \frac{(a_+ - 1)^2 \alpha^2 \log(p_n)}{4(1 + a_+ \sqrt{\log(p_n)/n})^2} \right\}
\]
\[
\leq p_n^{1-(a_+ - 1)^2 \alpha^2 (1-o(1))/4}.
\]
Here, the last inequality comes from Lemma B.2. On the other hand, we have

$$\mathbb{P}\{B \not\subseteq B_+\} = \mathbb{P}\left\{ \bigcup_{\nu \in B_+^c} \{ S_{\nu \nu} \geq 1 + \alpha \sqrt{\log(p_n)/n} \} \right\} \leq \sum_{\nu \in B_+^c} \mathbb{P}\{ S_{\nu \nu} \geq 1 + \alpha \sqrt{\log(p_n)/n} \}$$

$$\leq \sum_{\nu \in B_+^c} \mathbb{P}\left\{ \frac{S_{\nu \nu}}{\sigma_{\nu \nu}^2} \geq 1 + \alpha \sqrt{\log(p_n)/n} \cdot \frac{1}{1 + a_{-\alpha} \sqrt{\log(p_n)/n}} \right\} \leq \sum_{\nu \in B_+^c} \mathbb{P}\left\{ \lambda_{\nu}^2 \geq 1 + \frac{(1 - a_{-\alpha}) \alpha \sqrt{\log(p_n)/n}}{1 + a_{-\alpha} \sqrt{\log(p_n)/n}} \right\}$$

$$\leq |B_+^c| \sqrt{2(1 + a_{-\alpha} \sqrt{\log(p_n)/n}) \frac{1}{(1 - a_{-\alpha}) \alpha \sqrt{\log(p_n)/n}}} \exp\left\{ - \frac{(1 - a_{-\alpha})^2 \alpha^2 \log(p_n)}{4(1 + a_{-\alpha} \sqrt{\log(p_n)/n})^2} \right\}$$

$$\leq p_n^{1-(a_{-\alpha})^2 \alpha^2(1-a(1))/4}.$$ 

Here, the second last inequality comes from Lemma B.2. From the above bounds, if we choose $\alpha$ and $a_{\pm}$ properly, then $B_- \subset B \subset B_+$ holds with probability at least $1 - C_0 p_n^{-2}$.

To show $B_+ \subset H$, observe that for any $\nu \in B_+$, there exists some $j \in \{1, \ldots, \bar{m}\}$, s.t. $\lambda_{j\nu}^2 > \frac{a_{-\alpha}}{\bar{m}} \sqrt{\log(p_n)/n}$. By the definition of $H$, it then suffices to show that for large enough $n$,

$$\frac{a_{-\alpha}}{\bar{m}} \lambda_j^2 \sqrt{\frac{\log(p_n)}{n}} > \beta^2 r_{n,j}^2 = \beta^2 \lambda_j^2 \frac{1}{\bar{m}} \frac{\log(p_n)}{n}, \quad j = 1, \ldots, \bar{m}.$$

The inequality holds because, under conditions GR and SP, both $\log(p_n)/n \lambda_{j\nu}^2$ and $\log(p_n)/n$ converge to 0 as $n \to \infty$. 

**Lemma A.4.** Let $e_n^2 = \sum_{i=1}^{\bar{m}} s_i^2 \log(p_n)/(n \lambda_i^2)^{1/2-r/4}$. Then $\|\hat{Q}_{B_-^c}\|_F^2 = \sum_{j=1}^{m} \|q_j B_-\|_F^2 \leq C e_n^2$, and $\lambda_i \sqrt{\text{card}(H)/n}$, $\lambda_i \sqrt{\log(p_n)/n} = o(e_n)$ for $i = 1, 2$.

**Proof.** The bound on $\|\hat{Q}_{B_-^c}\|_F^2$ is obtained by following the lines of the proof of Lemma A.1 but with $H$ replaced by $B_-$. In addition, conditions GR and SP, together with Lemma A.1 imply the bounds on $\lambda_i \sqrt{\text{card}(H)/n}$ and $\lambda_i \sqrt{\log(p_n)/n}$.

We now prove Lemma 6.3. Throughout, we restrict our attention to the high probability event on which the conclusion of Lemma A.3 holds. Let $M = H \setminus B^o$, and define

$$S_0^o = \begin{bmatrix} S_{0,HH} & 0 \\ 0 & I_{LL} \end{bmatrix}, \quad \text{with} \quad S_{0,HH} = \begin{bmatrix} S_{B^oB^o} & 0 \\ 0 & I_{MM} \end{bmatrix}. \quad (A.19)$$

**Proof of claim (1)** First, claim (1) is a direct consequence of Lemma A.3. Recall that $B^o = B \cap H$. On the event such that the conclusion of Lemma A.3 holds, we get $B \subset H$, and so $B^o = B$.

**Proof of claim (2)** Here, we first apply Weyl’s theorem to obtain

$$|\ell_j(S_0^o) - \hat{\ell}_j^o| \leq \|S^o - S_0^o\| \leq \|S_{HH} - S_{0,HH}\| \leq 2\|S_{MB^o}\| + \|S_{MM} - I_{MM}\|. \quad (A.20)$$

To bound $\|S_{MB^o}\|$, observe that $S_{MB^o} = \sum_{i=1}^{4} D_i$, where

$$D_1 = \frac{1}{n} \hat{Q}_{M} \Lambda V'V \Lambda \hat{Q}_{B^o}, \quad D_2 = \frac{1}{n} Z_M Z_{B^o}, \quad D_3 = \frac{1}{n} \hat{Q}_M \Lambda V'Z_{B^o}, \quad D_4 = \frac{1}{n} Z_M V \Lambda Q_{B^o}.' $$

30
Note that $\|\tilde{Q}_B.\| \leq 1$ and that $M \subset B^c$. So, on event $\{A.9\}$, Lemma $A.4$ leads to
\[
\|D_1\| \leq \|\mathbf{A}\|^2 \frac{1}{n} V'V\|\|\tilde{Q}_M.\|_F \leq \lambda_1^2 \frac{1}{n} V'V\|\|\tilde{Q}_B.\|_F \leq C\lambda_1^2 e_n.
\]
For $D_2$, let $\tilde{Z} = [\tilde{Z}_1, \ldots, \tilde{Z}_p]$, with $\tilde{Z}_\nu = Z_\nu/\|Z_\nu\|_2$. Thus, $D_2$ satisfies
\[
\|D_2\| \leq \frac{1}{n} \max_{\nu \in H} \|Z_\nu\|^2 \cdot \|\tilde{Z}_M\|_{B.}.
\]
Note that $\|Z_\nu\|^2 \sim \text{i.i.d. } \chi_n^2$. Lemma $B.2$ ensures that with probability at least $1 - C_0 p_n^{-2}$, $\frac{1}{n} \max_{\nu \in H} \|Z_\nu\|^2 = 1 + o(1)$. On the other hand, since $B^o$ only depends on $\{\|Z_\nu\|_2, \nu \in H\}$, which are independent of $\{\tilde{Z}_\nu, \nu \in H\}$, we obtain that $\tilde{Z}_H$ is independent of $B^o$. So, $\tilde{Z}_M$ is independent of $\tilde{Z}_{B^o}$. Let $U = \text{diag}(u_1, \ldots, u_{\text{card}(B^o)})$ with $u_j$ i.i.d. $\chi_n^2$ random variables. Then, $\tilde{Z}_M'\tilde{Z}_{B^o}U$ has i.i.d. $N(0,1)$ entries, and so Lemma $B.3$ asserts that with probability at least $1 - C_0 p_n^{-2}$,
\[
\|\tilde{Z}_M'\tilde{Z}_{B^o}U\| \leq C\left(\sqrt{\text{card}(M)} + \sqrt{\text{card}(B^o)} + \sqrt{\log(p_n)}\right) \leq C\left(\sqrt{\text{card}(H)} + \sqrt{\log(p_n)}\right).
\]
In addition, Lemma $B.2$ shows that with probability at least $1 - C_0 p_n^{-2}$, $\min_j u_j = n(1 - o(1))$. Thus, we obtain that, with probability at least $1 - C_0 p_n^{-2}$,
\[
\|\tilde{Z}_M'\tilde{Z}_{B^o}\| \leq \frac{\|\tilde{Z}_M'\tilde{Z}_{B^o}\|_{\min_j u_j}}{\lambda_1} \leq C\left(\sqrt{\text{card}(H)} + \sqrt{\log(p_n)}\right) = o(\lambda_1^2 e_n).
\]
Here, the last inequality comes from Lemma $A.4$. Thus, we obtain $\|D_2\| = o(e_n)$. For $D_3$ and $D_4$, on the event $\{A.11\}$, Lemmas $A.2$ and $A.4$ lead to
\[
\|D_3\| \leq \frac{1}{n} \|\mathbf{A}\|\|V'Z_{B^o}\|\|\tilde{Q}_M.\|_F \leq \frac{1}{n} \lambda_1 \|V'Z_{H}\|\|\tilde{Q}_{B^c.}\|_F \leq \lambda_1 \xi(n, \tilde{m}, \text{card}(H); 2)\|\tilde{Q}_{B^c.}\|_F = o(\lambda_1^2 e_n),
\]
\[
\|D_4\| \leq \frac{1}{n} \|\mathbf{A}\|\|Z'_M V\|\|\tilde{Q}_{B^o}\| \leq \frac{1}{n} \lambda_1 \|Z'_H V\| \leq \lambda_1 \xi(n, \tilde{m}, \text{card}(H); 2) = o(\lambda_1^2 e_n).
\]
Combining the four parts leads to $\|S_{MB^o}\| \leq C\lambda_1^2 e_n$.

Switch to $S_{MM} - I_{MM}$. We decompose it as $S_{MM} - I_{MM} = \sum_{k=1}^2 E_{MM,k}$, where
\[
E_{MM,1} = \tilde{Q}_M A\left(\frac{1}{n} V'V - I\right) \Lambda \tilde{Q}_M, \quad E_{MM,2} = \frac{1}{n} Z'_M Z - I_{MM},
\]
and
\[
E_{MM,3} = \frac{1}{n} \left(\tilde{Q}_M A V'Z_{M} + Z'_M V A\tilde{Q}_M\right).
\]
Consider the intersection of the events in $\{A.9\}$, $\{A.10\}$ and $\{A.11\}$. Lemmas $A.2$ and $A.4$ lead to
\[
\|E_{MM,1}\| \leq \lambda_1^2 \frac{1}{n} V'V\|\|\tilde{Q}_M.\|_F^2 \leq C\lambda_1^2 \|\tilde{Q}_{B^c.}\|_F^2 = o(\lambda_1^2 e_n),
\]
\[
\|E_{MM,2}\| \leq \frac{1}{n} \lambda_1 \|Z'_H Z_{H} - I_{HH}\| \leq \xi(n, \text{card}(H)) = o(\lambda_1^2 e_n),
\]
\[
\|E_{MM,3}\| \leq \frac{2}{n} \lambda_1 \|V'Z_{H}\|\|\tilde{Q}_{M.}\|_F \leq C\lambda_1 \xi(n, \tilde{m}, \text{card}(H); 2)\|\tilde{Q}_{B^c.}\|_F = o(\lambda_1^2 e_n).
\]
Thus, we combine the three parts to obtain $\|S_{MM} - I_{MM}\| = o(\lambda_1^2 e_n)$.

Condition SP implies that $e_n = o(1)$. Thus, $\|S_{MB^o}\|$, together with the bounds on $\|S_{MB^o}\|$ and $\|S_{MM} - I_{MM}\|$, leads to $|\ell_j(S_{B^o}) - \tilde{c}_j| = o(\lambda_1^2)$, for $j = 1, \ldots, \tilde{m}$. In addition, condition AD($m, \kappa$) implies that $\lambda_1^2 \geq \lambda_2^2/\kappa$. Thus, the last display also implies that $\ell_j(S_{B^o}) = \ell_j(S_{B^o}) > 1$ for $j = 1, \ldots, m$. This completes the proof claim (2).
Proof of claim (3) Turn to claim (3). Since \( \ell_j(S_0^o) = \ell_j(S_{B^o B^o}) > 1 \) for \( j = 1, \ldots, m \), \( \hat{T}_{m}^{(0),o} \) is the leading principal subspace of \( S_0^o \). Thus, we apply Theorem 13.1 with \( A = S^o \), \( A + E = S_0^o \) and \( \delta = \ell_m(S_0^o) - \ell_{m+1}(S^o) \). Claim (1) and Lemma 6.2 imply that \( \delta = (\lambda_m^2 - \lambda_{m+1}^2)(1 + o(1)). \) Therefore, Theorem 13.1 leads to

\[
L(\hat{T}_{m}^{(0),o}, \hat{T}_{m}^{(0),o}) \leq \frac{\|S^o - S_0^o\|^2}{(\lambda_m^2 - \lambda_{m+1}^2)^2(1 + o(1))} \leq \frac{C\lambda_m^4 e_2}{(\lambda_m^2 - \lambda_{m+1}^2)^2} \leq C\epsilon_n^2 \leq \frac{1}{5}(1 - \rho)^2.
\]

The last inequality comes from conditions SP, AD\((m, \kappa)\) and the definition of \( \rho \), and holds when \( n \) is sufficiently large.

A.4 Proof of Proposition 6.1

Here, we restrict the proof on the event such that the conclusions of Lemmas 6.1, 6.3 hold. Thus, all the arguments are deterministic. Recall that \( T^{(k),o} \) denotes the matrix obtained after the multiplication step of the \( k \)-th iteration in oracle Algorithm 1. Denote \( \text{ran}(T^{(k),o}) \) by \( \mathcal{T}^{(k),o} \), and let \( \phi^{(k)} \) be the largest canonical angle between \( \mathcal{T}^{(k),o} \) and \( \hat{T}_{m}^{(k-1),o} \). So, \( \sin^2 \phi^{(k)} = L(\mathcal{T}^{(k),o}, \hat{T}_{m}^{(k-1),o}). \)

In what follows, we first prove both claims for the first iteration, and then extend them to subsequent iterations.

The First Iteration Claim (1) relies on the following lemma, which describes the effects of the multiplication and the thresholding steps separately.

Lemma A.5. In the first iteration of oracle Algorithm 1,

1. After the multiplication step, \( T^{(1),o} \) has full column rank and \( \sin \phi^{(1)} \leq \rho \tan \theta(0); \)
2. After the thresholding step, \( \hat{T}^{(1),o} \) has full column rank, and \( L(T^{(1),o}, \hat{T}_{m}^{(1),o}) \leq \omega^2 \sec^2 \theta(0). \)

Proof. Claim (1) is essentially one-step analysis of orthogonal iteration, which is obtained by directly applying Theorem 8.2.2 in [9] to a single iteration.

The proof of claim (2) relies on Wedin’s sin \( \theta \) theorem for singular subspaces (Theorem 13.2). In the theorem, let \( A = T^{(0),o} \) and \( B = \hat{T}^{(1),o}. \) Then, the theorem gives that

\[
L(T^{(1),o}, \hat{T}_{m}^{(1),o}) \leq \frac{\|T^{(1),o} - T^{(1),o}\|^2}{\sigma_m^2(T^{(1),o})}. \quad (A.21)
\]

In what follows, we study the numerator and the denominator of the right side separately.

First, we derive a lower bound for \( \sigma_m(T^{(1),o}) \). To this end, for any unit vector \( x \in \mathbb{R}^m \), let \( y = \hat{Q}^{(0),o} x \), then \( y \) is a unit vector in \( \mathbb{R}^p \). Decompose \( y \) as \( y = \tilde{y} + \tilde{y}_c \), with \( \tilde{y} \in \text{ran}(\hat{Q}^{o}) \) and \( \tilde{y}_c \in \text{ran}(\hat{Q}_c^{o}) \). Then, it follows that \( \|T^{(1),o} x\|^2 = \|S^o y\|^2 = \|S^o \tilde{y}\|^2 + \|S^o \tilde{y}_c\|^2 \geq \|S^o \tilde{y}\|^2 \geq (\hat{\rho}_m)^2 \|\tilde{y}\|^2 \). Since \( \|\tilde{y}\|^2 \geq \cos^2 \theta(0) \), we obtain that

\[
\sigma_m^2(T^{(1),o}) \geq \inf_{\|x\|_2 = 1} \|T^{(1),o} x\|^2 \geq (\hat{\rho}_m)^2 \cos^2 \theta(0). \quad (A.22)
\]

To bound the numerator, we define matrix \( \Delta T \in \mathbb{R}^{p \times m} \), whose \((\nu, j)\)-th entry is given by

\[
(\Delta T)_{\nu j} = \gamma_{nj} 1_{(\nu \in H)}.
\]

32
Then, we obtain
\[
\|\hat{T}^{(1),0} - T^{(1),0}\| \leq \|\hat{T}^{(1),0} - T^{(1),0}\|_F \leq \|\Delta T\|_F = \hat{c}_m \omega. \tag{A.23}
\]
Here, the second inequality comes from the constraint on the thresholding function \(\eta\).

Plug the bounds \(A.22\) and \(A.23\) into \(A.21\), we obtain the bound for \(L(T^{(1),0}, \hat{T}_m^{(1),0})\) in the lemma. Finally, we are to show that \(\hat{\eta}\) is equivalent to the constraint on the thresholding function \(\eta\). Here, the second inequality comes from Lemma \(A.5\). This completes the proof of claim (1).

Note that Lemmas 6.1–6.3 lead to \(Q\). Here, \(\eta\) tends to 0 as \(n\) tends to \(\infty\). Therefore, for sufficiently large \(n\), which completes the proof.

By the discussion following the loss function \(L\), we have the triangle inequality
\[
\sin \theta^{(1)} \leq \sin \phi^{(1)} + \frac{L^{1/2}(T^{(1),0}, \hat{T}_m^{(1),0})}{\cos \theta^{(0)} \leq \rho \tan \theta^{(0)} + \omega \sec \theta^{(0)}.}
\]
Here, the second inequality comes from Lemma \(A.3\). This completes the proof of claim (1).

Turn to claim (2). We first show that, for any \(a \leq 1/2\),
\[
\sin^2 \theta^{(0)} > \frac{1.01 \omega^2}{(1 - a)^2(1 - \rho)^2} \tag{A.24}
\]
implies \(\sin^2 \theta^{(1)}/\sin^2 \theta^{(0)} \leq [1 - a(1 - \rho)]^2\). By claim (1), it suffices to show
\[
\frac{\rho}{\cos \theta^{(0)}} + \frac{\omega}{\sin \theta^{(0)} \cos \theta^{(0)}} \leq 1 - a(1 - \rho).
\]
Let \(x = \sin \theta^{(0)}\). Multiply both sides of the last display with \(x\sqrt{1 - x^2}\) and collect terms. Then it is equivalent to \(\omega \leq x \{[1 - a(1 - \rho)] \sqrt{1 - x^2} - \rho\}\). The right side of last inequality is non-negative, because \(x^2 \leq (1 - \rho)^2/5\) (by Lemma \(6.3\)) and \(a \leq 1/2\). So, we can square both sides of it, and some simple algebra shows that the inequality holds if
\[
Q(x^2) = [1 - a(1 - \rho)]^2x^4 - (1 - a)^2(1 - \rho)^2x^2 + \omega^2 \leq 0. \tag{A.25}
\]
Here, \(Q(x^2)\) is a quadratic form of \(x^2\), with discriminant \(\Delta = (1 - a)^4(1 - \rho)^4 - 4\omega^2[1 - a(1 - \rho)]^2\).

Note that Lemmas \(6.1, 6.3\) lead to
\[
\frac{\omega^2}{(1 - \rho)^4} \geq \text{card}(H)\gamma^2 \sum_{j=1}^{m}(\lambda_j^2 + 1) \log(p_n) \frac{\log(\lambda_m + 1)}{n^2} \left(1 - \frac{\lambda_1 + 1}{\lambda_m + 1}\right) (1 + o(1)) \tag{A.26}
\]
Here, \(C = C(\gamma, r, m, \kappa)\). Moreover, conditions GR and SP imply that the rightmost side of \(A.26\) tends to 0 as \(n \to \infty\). Therefore, for sufficiently large \(n \geq n_0(\gamma, r, m, \kappa)\), we have
\[
\frac{\omega^2}{(1 - \rho)^4} \leq \frac{3}{16} \leq \frac{(1 - a)^4}{4[1 - a(1 - \rho)]^2},
\]
and so \(\Delta > 0\). Let the two roots of \(Q\) be \(x_{\pm}^2\). We solve \(Q\) to obtain that \(A.25\), and hence the conclusion, holds when
\[
x^2 \geq x_{+}^2 = \frac{\omega^2(1 + o(1))}{(1 - a)^2(1 - \rho)^2}. \tag{A.27}
\]
Note that, by (A.26), \( x^2 \leq (1 - \rho)^2 / 5 \leq x^2 \) always holds. Therefore, the case is completed because (A.24) implies (A.27).

Next, we show that if \( \sin^2 \theta(0) \leq 1.01 \omega^2 / [(1 - a)(1 - \rho)]^2 \), then so is \( \sin^2 \theta(1) \). By claim (1), it suffices to show that \( \rho \tan \theta(0) + \omega \sec \theta(0) \leq \sqrt{1.01 \omega^2 / [(1 - a)(1 - \rho)]} \). Multiply both sides by \( \cos \theta(k\omega)(1 - a)(1 - \rho) \). We obtain that the last inequality holds when

\[
\rho + \frac{(1 - a)(1 - \rho)}{\sqrt{1.01}} \leq \sqrt{1 - \frac{1.01 \omega^2}{(1 - a)^2(1 - \rho)^2}},
\]

because the right side is no greater than \( \cos \theta(0) \) by the condition on \( \sin \theta(0) \). Since \( \sqrt{x} \geq x \) for \( x \in [0, 1] \), it further suffices to have the last inequality hold but with the square root of the right side removed, which is equivalent to have

\[
\frac{1.01 \omega^2}{(1 - a)^2(1 - \rho)^2} \leq (1 - \rho)[1 - (1 - a) / \sqrt{1.01}].
\]

By (A.26), the last desired inequality holds when \( n \geq n_0(\gamma, r, m, \kappa) \). This completes the case.

**Subsequent Iterations** Now we have \( \hat{Q}^{(1)} \) is orthonormal with \( \sin^2 \theta(1) \leq (1 - \rho)^2 / 5 \). Therefore, we can essentially repeat the whole argument to prove both claims for \( k = 2 \), and so on by induction. This is valid because the minimum sample size needed, \( n_0 = n_0(\gamma, r, m, \kappa) \), such that both claims hold does not depend on the iteration index \( k \).

### A.5 Proof of Proposition 6.2

we divide the proof into two parts, depending on whether the following condition holds or not:

\[
1.01 \omega^2 / (1 - \rho)^2 \geq \epsilon_{nm}^2 / 4. \tag{A.28}
\]

Without loss of generality, we assume that \( n = 2^l \) for some \( l \geq 1 \). So, \( l = \log n / \log 2 \).

First, consider the case where (A.28) holds. Let \( k_1 \) be the number of iterations needed to achieve

\[
L(\hat{P}_{m, o}, \hat{D}^{(k_1)}) \leq 1.01 \omega^2 / [(1 - \frac{1}{2})^2(1 - \rho)].
\]

Before the last inequality is satisfied, the decay rate of the approximation error satisfies the bound in claim (2) of Proposition 6.1 with \( a = 1/2 \). Thus, it suffices to have \( k_1 \) satisfy \( [1 - (1 - \rho)/2]^{2k_1} < 1.01 \omega^2 / [(1 - \frac{1}{2})(1 - \rho)]^2 \), i.e., \( 2k_1 \log(1 - \frac{1}{2} - (1 - \rho)) \geq \log([(1 - \frac{1}{2})^2(1 - \rho)] / 1.01 \omega^2) \). Condition (A.28) implies \( (1 - 1/2)^2(1 - \rho)^2 / 1.01 \omega^2 \leq \epsilon_{nm}^2 / \leq nh(\lambda_m^2) \). In addition, \( \log(1 - x) \geq x \), for all \( x \in (0, 1) \). Thus, it suffices to set

\[
k_1 = \frac{\log n + \log h(\lambda_m^2)}{1 - \rho} = \frac{\lambda_m^2 + 1}{\lambda_m^2 - \lambda_{m+1}^2} [\log n + \log h(\lambda_m^2)] (1 + o(1)).
\]

Now, let \( k_2 - k_1 \) be the number of additional iterations needed to achieve \( L(\hat{P}_{m, o}, \hat{D}^{(k_2)}) \leq 1.01 \omega^2 / [(1 - \frac{1}{4})(1 - \rho)]^2 \). Then, before the inequality is satisfied, the decay rate satisfies claim (2)
of Proposition 6.1 with \( a = 1/4 \). Thus, it suffices to have \( [1 - \frac{1}{4}(1 - \rho)]^{2(k_2 - k_1)} \leq (1 - \frac{1}{4})^2 \).
This is guaranteed if we set
\[
k_2 - k_1 \geq \frac{4}{1 - \rho} \left[ \log \left( 1 - \frac{1}{4} \right) - \log \left( 1 - \frac{1}{2} \right) \right].
\]

Recursively, we define \( k_i \) for \( i = 3, \ldots, l \), such that \( L(\widehat{P}_m^{(k_i)}, \widehat{P}_m^{(k)}) \leq 1.01 \omega^2/[(1 - 2^{-i})(1 - \rho)]^2 \).
Repeating the above argument shows it suffices to have
\[
k_i - k_{i-1} \geq \frac{2^i}{1 - \rho} \left[ \log \left( 1 - \frac{1}{2^i} \right) - \log \left( 1 - \frac{1}{2^{i-1}} \right) \right], \quad \text{for } i = 3, \ldots, l.
\]
Therefore, if we set
\[
k_1 - k_1 = \frac{l + 1/2}{1 - \rho} = \frac{\lambda^2 + 1}{\lambda^2_n - \lambda^2_{n+1}} \log 2 (1 + o(1)) \geq \sum_{i=2}^{l} \frac{2^i}{1 - \rho} \left[ \log \left( 1 - \frac{1}{2^i} \right) - \log \left( 1 - \frac{1}{2^{i-1}} \right) \right],
\]
then \( L(\widehat{P}_m^{(k)}, \widehat{P}_m^{(k)}) \leq 1.01 \omega^2/[(1 - n^{-1})(1 - \rho)]^2 \) for all \( k \geq k_1 \). We complete the proof for this case
by noting that \( K \geq k_1 \) for large enough \( n \) and that
\[
\frac{\omega^2}{(1 - \rho)^2} \leq C \gamma^2 \text{card}(H)h(\lambda^2_n) \leq C M_n \epsilon^2_{nm}.
\]

Turn to the case where (A.28) does not hold. Then, the decay rate in claim (2) of Proposition 6.1 holds with \( a = 1/2 \) for all \( k \) such that \( L(\widehat{P}_m^{(k)}, \widehat{P}_m^{(k)}) > \epsilon_{nm}^2 \). Since the approximation error is
monotone decreasing, it suffices to verify that \( [1 - (1 - \rho)/2]^{2K} \leq \epsilon_{nm}^2 \), i.e.,
\[
2K \left| \log \left( 1 - \frac{1 - \rho}{2} \right) \right| \geq \log \frac{n}{\log(p_n)} + \log \frac{(\lambda^2_n - \lambda^2_{n+1})^2}{(\lambda^2 + 1)(\lambda^2_{n+1} + 1)}.
\]
Note that the second term on the right side is bounded above by \( \log h(\lambda^2_n) \). Since \( \log(p_n) = o(n) \)
and \( |\log(1 - x)| \geq x \) for \( x \in (0, 1) \), it suffices to verify
\[
K \geq \frac{1}{1 - \rho} \left[ \log n + \log h(\lambda^2_n) \right] = \frac{\lambda^2_n + 1}{\lambda^2_n - \lambda^2_{n+1}} \log n (1 + o(1)),
\]
which is indeed satisfied by \( K \) for large enough \( n \). Thus, for all \( k \geq K \), \( L(\widehat{P}_m^{(k)}, \widehat{P}_m^{(k)}) \leq \epsilon_{nm}^2 \). This
completes the proof.

A.6 Proof of Lemma 6.4

First of all, we restrict our attention to the event on which all the conclusions of Lemmas 6.1, 6.3 hold. Lemma [A.3] shows that \( B^0 = B \), which leads to \( \widehat{Q}^{(0)} = \widehat{Q}^{(0),0} \). In what follows, we start with
the equivalence between \( \widehat{Q}^{(1)} \) and \( \widehat{Q}^{(1),0} \), and then extend the argument to subsequent iterations.
The First Iteration  Fix $\nu \in L$ and $1 \leq j \leq m$. Let $\hat{Q}^{(0),o} = [\hat{q}^{(0),o}_1, \ldots, \hat{q}^{(0),o}_m]$. Since $\hat{Q}^{(0)} = \tilde{Q}^{(0),o}$, the $(\nu,j)$-th element of $T^{(1)}$ is then $t^{(1)}_{\nu j} = S_{\nu} \hat{q}^{(0),o}_{jL} = S_{\nu H} \hat{q}^{(0),o}_{jH}$. Here, the second equality holds because $\hat{q}^{(0),o}_{jL} = 0$. Our goal is to show that for sufficiently large numeric constant $\gamma$, the size of $t^{(1)}_{\nu j}$ is bounded by the threshold $\gamma_{nj}$ in Theorem 3.1 with high probability. Following the discussion after Lemma 3.3 it suffices to show that $t^{(1)}_{\nu j}$ is bounded by a constant multiple of $\sqrt{(\lambda_j^2 + 1) \log (p_n) / n}$, since $\ell_j(S_{BB}) / (\lambda_j^2 + 1) \rightarrow 1$.

By (A.44), we decompose $t^{(1)}_{\nu j}$ as

$$t^{(1)}_{\nu j} = S_{\nu H} \left[ Q_{H}^{o}(Q_{H}^{o})' + \Lambda V' V \Lambda Q_{H}^{o} \right] \hat{q}^{(0),o}_{jH} = t^{(1)}_{1} + t^{(1)}_{2}. \quad (A.29)$$

Moreover, $S_{\nu H}$ admits the decomposition $S_{\nu H} = \sum_{i=1}^{4} S_{\nu H,i}$, where

$$S_{\nu H,1} = \frac{1}{n} \bar{Q}_{\nu} \Lambda V' V \Lambda Q_{H}^{o}, \quad S_{\nu H,2} = \frac{1}{n} Z' \nu Z_{H}$$

$$S_{\nu H,3} = \frac{1}{n} \bar{Q}_{\nu} \Lambda V' Z_{H}, \quad S_{\nu H,4} = \frac{1}{n} Z' \nu \Lambda Q_{H}^{o}. \quad (A.30)$$

In what follows, we bound $t_{1}^{(1)}$, $i = 1, 2$, respectively. The argument is lengthy but elementary. The key conclusions are (A.31) and (A.36), which then lead to (A.43).

1°. Bound for $t_{1}^{(1)}$. We show below

$$|t_{1}^{(1)}| \leq (1 + o(1)) \|Q_{H}^{o}(Q_{H}^{o})' \|_{2} (\beta \sqrt{\bar{m}} + 2 \sqrt{3} \gamma_{nj}) \frac{\gamma_{nj}}{\gamma}. \quad (A.31)$$

To this end, using (A.30), we decompose $t_{1}^{(1)} = \sum_{i=1}^{4} S_{\nu H,i} Q_{H}^{o}(Q_{H}^{o})' \hat{q}^{(0),o}_{jH} = \sum_{i=1}^{4} t_{1i}^{(1)}$.

For $t_{1}^{(1)}$, we further decompose it as

$$t_{11}^{(1)} = \bar{Q}_{\nu} \Lambda \bar{Q}_{H}^{o}(Q_{H}^{o})' \hat{q}^{(0),o}_{jH} + \bar{Q}_{\nu} \Lambda \left( \frac{1}{n} V' V - I \right) \Lambda \bar{Q}_{H}^{o}(Q_{H}^{o})' \hat{q}^{(0),o}_{jH}. \quad (A.32)$$

Recall (A.4). We have $\|\bar{Q}_{\nu} \Lambda \bar{Q}_{H}^{o}(Q_{H}^{o})' \| \leq \|Q_{H}^{o}(Q_{H}^{o})' \| + \|Q_{H}^{o} \Lambda_{V}^{2} \| \|Q_{H}^{o}(Q_{H}^{o})' \|. \quad (A.33)$

Let $\rho_{lj} = (\lambda_{l}^2 + 1) / (\lambda_{j}^2 + 1), \; 1 \leq l, j \leq \bar{m}$. Since $\nu \in L$, we have $|q_{lj}| \leq \beta \gamma_{nj}$, and so $\|Q_{H}^{o} \Lambda_{V}^{2} \|^2 \leq \sum_{l=1}^{m} \beta^{2} \rho_{lj}^{2} \lambda_{l}^2 = \sum_{l=1}^{m} \rho_{lj}^{1/2} (\beta / \gamma) \gamma_{nj}$, with $\gamma$ and $\gamma_{nj}$ specified as in Theorem 3.1. Similarly, $\|Q_{1} \nu \Lambda_{V}^{2} \|^2 \leq \sum_{l=1}^{m} \rho_{lj}^{1/2} (\beta / \gamma) \gamma_{nj}$. Moreover, we have $\|Q_{H}^{o}(Q_{H}^{o})' \| \leq 1$ and $\|Q_{1}^{H}(Q_{H}^{o})' \| \leq \|Q_{1} \nu \Lambda_{V}^{2} \| \leq L^{1/2}(P_{m}, \bar{P}_{m}^{\rho}) = o(1)$. Therefore,

$$\|\bar{Q}_{\nu} \Lambda \bar{Q}_{H}^{o}(Q_{H}^{o})' \| \leq (1 + o(1)) \left( \sum_{l=1}^{m} \rho_{lj}^{1/2} \beta \gamma_{nj} \right).$$

On the other hand, on the event (A.43), we have

$$\|\bar{Q}_{\nu} \Lambda \left( \frac{1}{n} V' V - I \right) \Lambda \bar{Q}_{H}^{o} \| \leq \|\bar{Q}_{\nu} \Lambda \|_{2} \left( \sum_{l=1}^{m} \rho_{lj}^{1/2} \beta \gamma_{nj} \right) \leq \lambda_{1} \zeta(n, \bar{m}) \left( \sum_{l=1}^{m} \rho_{lj}^{1/2} \beta \gamma_{nj} \right) = o(\gamma_{nj}).$$

36
Here, the last equality comes from conditions GR and SP. Finally, the triangle inequality leads to

\[ |t_{11}^{(1)}| \leq (1 + o(1)) \left( \sum_{i=1}^{m} \rho_{ij} \right)^{1/2} \| (Q_{H}^{o})^{'0,0,0} \|_{2} \frac{\beta \gamma_{nj}}{\gamma}. \]

For \( t_{12}^{(1)} \), define event

\[ \left\{ \| Z_{\nu}^{'0} Z_{H} Q_{H}^{o} \|_{2} \leq n \xi(n, 1, m; \sqrt{6}), \forall \nu \in L \right\}. \quad (A.33) \]

Note that \( \xi(n, 1, m; \sqrt{6}) = (1 + o(1))[6/(\lambda_{j}^2 + 1)]^{1/2}(\beta/\gamma)\gamma_{nj} \). On this event, we have

\[ |t_{12}^{(1)}| \leq (1 + o(1)) \left( \frac{6}{\lambda_{j}^2 + 1} \right)^{1/2} \| (Q_{H}^{o})^{'0,0,0} \|_{2} \frac{\gamma_{nj}}{\gamma}. \]

For \( t_{13}^{(1)} \), on the event \( (A.34) \), we have

\[ |t_{13}^{(1)}| \leq \frac{1}{n} \| \bar{Q}_{\nu} \Lambda \|_{2} \| V_{\nu}^{'0} Z_{H} Q_{H}^{o} \| \| (Q_{H}^{o})^{'0,0,0} \|_{2} \]

\[ \leq (1 + o(1)) 2 \sqrt{\frac{\log(p_{n})}{n}} \left( \sum_{i=1}^{m} \frac{\rho_{ij}}{\lambda_{i}^2} \right)^{1/2} \| (Q_{H}^{o})^{'0,0,0} \|_{2} \frac{\beta \gamma_{nj}}{\gamma} = o(\gamma_{nj}). \]

Switch to \( t_{14}^{(1)} \). We note that \( \| \Lambda \bar{Q}_{H} Q_{H}^{o} \| \leq \lambda_{1}(1 + o(1)). \) So, on the event

\[ \left\{ \| Z_{\nu}^{'0} V \|_{2} \leq n \xi(n, 1, \tilde{m}; \sqrt{6}), \forall \nu \in L \right\}, \quad (A.34) \]

we have

\[ |t_{14}^{(1)}| \leq (1 + o(1)) \left( \frac{6 \lambda_{1}^2}{\lambda_{j}^2 + 1} \right)^{1/2} \| (Q_{H}^{o})^{'0,0,0} \|_{2} \frac{\gamma_{nj}}{\gamma}. \]

Finally, we apply the triangle inequality to obtain

\[ |t_{1}^{(1)}| \leq \sum_{i=1}^{4} |t_{1i}^{(1)}| \leq (1 + o(1)) \| (Q_{H}^{o})^{'0,0,0} \|_{2} \left[ \beta \left( \sum_{i=1}^{m} \rho_{ij} \right)^{1/2} + 2 \sqrt{3} \rho_{ij}^{1/2} \right] \frac{\gamma_{nj}}{\gamma}. \quad (A.35) \]

Without loss of generality, we could assume that \( \lambda_{1}/\lambda_{m} \to 1 \) as \( n \to \infty \). When this is the case, \( (A.35) \) reduces to \( (A.31) \). If this is not the case, let \( 1 < m_{0} \leq m \) be the smallest index such that \( \lambda_{m_{0}}/\lambda_{m} \to 1 \). We further divide the set \( \{1, \ldots, m\} \) into \( \{1, \ldots, m_{0} - 1\} \cup \{m_{0}, \ldots, m\} \), and similar but lengthier argument would also lead to \( (A.31) \).

2\textsuperscript{a}. Bound for \( t_{2}^{(1)} \). Here, we show

\[ |t_{2}^{(1)}| \leq (1 + o(1)) \| (P_{H}^{o})^{'0,0,0} \|_{2} \left[ \beta \left( \sum_{i=m_{0}+1}^{m} \rho_{ij} \right)^{1/2} + \sqrt{7} \rho_{ij}^{1/2} \right] \frac{\gamma_{nj}}{\gamma}. \quad (A.36) \]

To this end, observe that \( (A.30) \) leads to

\[ t_{2}^{(1)} = \sum_{i=1}^{4} S_{\nu H, i} P_{H}^{o} (P_{H}^{o})^{'0,0,0} = \sum_{i=1}^{4} t_{2i}^{(1)}. \]
For $t^{(1)}_{21}$, by argument similar to that for $t^{(1)}_{11}$, we have on event \( (A.9) \),

\[
|t^{(1)}_{21}| \leq (1 + o(1)) \left( \sum_{t=m+1}^{\hat{m}} \rho_{tj} \right)^{1/2} \| (P^o_H)_{tjH}^{(0),o} \|_2 \| 2 \beta \gamma n_j \gamma .
\]

For $t^{(1)}_{22}$, define

\[
y^{(1)}_{\nu j} = n t^{(1)}_{22} / \| Z_H P^o_H (P^o_H)_{\nu jH}^{(0),o} \|_2.
\]

By our construction of the initial value, $q^{(0),o}_{\nu jH}$ is a function of $S_{HH}$ and hence is independent of $Z_{\nu}$. In addition, $Z_H$ is independent of $Z_{\nu}$. Thus, we have $y^{(1)}_{\nu j} \sim N(0,1)$. Define events

\[
\left\{ |y^{(1)}_{\nu j}| \leq \sqrt{7 \log(p_n)}, 1 \leq j \leq m, \forall \nu \in L \right\}, \quad (A.38)
\]

\[
\left\{ \| Z_H \| \leq \sqrt{n} + \sqrt{\text{card}(H)} + 2 \sqrt{\log(p_n)} \right\} . \quad (A.39)
\]

On the intersection of the above events, we have

\[
|t^{(1)}_{22}| \leq \frac{1}{n} \| Z_H \| \| (P^o_H)_{tjH}^{(0),o} \|_2 \leq (1 + o(1)) \| (P^o_H)_{\nu jH}^{(0),o} \|_2 \left( \frac{7}{\lambda_j^2 + 1} \right)^{1/2} \frac{\beta \gamma n_j \gamma}{\gamma}.
\]

Switch to $t^{(1)}_{23}$. On the event \( (A.11) \), we have

\[
|t^{(1)}_{23}| \leq \frac{1}{n} \| \tilde{Q}_\nu \Lambda \|_2 \| V' Z_H \| \| (P^o_H)_{\nu jH}^{(0),o} \|_2 \leq \frac{C}{\lambda_m} \sqrt{\frac{\text{card}(H)}{n}} \| (P^o_H)_{\nu jH}^{(0),o} \|_2 \frac{\beta \gamma n_j \gamma}{\gamma}.
\]

Condition (GR) implies

\[
\frac{1}{\lambda_m^2} \frac{\text{card}(H)}{n} \leq \frac{\beta - r}{n \lambda_m^2} \sum_{j=1}^{m} s_j \tau_{n_j} = \frac{\beta - r}{n^{1/2 - r/4} \log(p_n)^{1/2 + r/4}} \sum_{j=1}^{m} s_j \left[ \frac{\log(p_n)}{n \lambda_j^2} \right]^{1/2 - r/4} \frac{\lambda_j^2}{\lambda_m^2} \left( \frac{\lambda_j^2}{\lambda_m^2 + 1} \right)^{r/2}
\]

\[
\leq \frac{\beta - r}{n^{1/2 - r/4} \log(p_n)^{1/2 + r/4}} \sum_{j=1}^{m} s_j \left[ \frac{\log(p_n)}{n \lambda_j^2} \right]^{1/2 - r/4} = o(1).
\]

Thus, we obtain $|t^{(1)}_{23}| = o(\gamma n_j)$. For $t^{(1)}_{24}$, we could bound it using the same strategy as for $t^{(1)}_{22}$. In particular, let

\[
w^{(1)}_{\nu j} = n t^{(1)}_{24} / \| V' \tilde{Q}_\nu P^o_H (P^o_H)_{tjH}^{(0),o} \|_2 \quad (A.40)
\]

For any $\nu \in L$, $Z_{\nu}$ is independent of both $V$ and $q^{(0),o}_{\nu jH}$, and so $w^{(1)}_{\nu j} \sim N(0,1)$. Define events

\[
\left\{ |w^{(1)}_{\nu j}| \leq \sqrt{7 \log(p_n)}, 1 \leq j \leq m, \forall \nu \in L \right\}, \quad (A.41)
\]

\[
\left\{ \| V \| \leq \sqrt{n} + \sqrt{m} + 2 \sqrt{\log(p_n)} \right\} . \quad (A.42)
\]
Then on the intersection of them, we have
\[
|t_{24}^{(1)}| \leq \frac{1}{n}|w_{\nu_j}^{(1)}||V||\Lambda|| (P_H^o \hat{q}_{jH}^{(0),o})_2 \leq (1 + o(1)) \left( \frac{7\lambda^2}{\lambda^2 + 1} \right)^{1/2} \|(P_H^o)\hat{q}_{jH}^{(0),o}\|_2 \frac{\gamma_{nj}}{\gamma}.
\]

Finally, the triangle inequality leads to (A.36).

3°. Summary. By Lemmas 6.1–6.3 we have
\[
\|(Q_H^o)\hat{q}_{jH}^{(0),o}\|_2 = 1 + o(1), \quad \|(P_H^o)\hat{q}_{jH}^{(0),o}\|_2 = o(1).
\]

This, together with (A.29), (A.31) and (A.36), implies that, if we choose \( \beta = c/\sqrt{m} \), then for large enough \( \gamma \geq \gamma_0(c) \) (e.g., \( \gamma \geq 1.01 \cdot (2\sqrt{3} + c) \)), when \( n \) is sufficiently large,
\[
|t_{\nu_j}^{(1)}| \leq (1 + o(1))(\beta\sqrt{m} + 2\sqrt{3}) \frac{\gamma_{nj}}{\gamma} \leq \gamma_{nj}.
\]

Since the above inequality holds for all \( \nu \in L \) and \( 1 \leq j \leq m \), we obtain \( \hat{T}^{(1)}_L = 0 \), and so, \( \hat{T}^{(1)}_L = \hat{T}^{(1),o} \). Therefore, \( \hat{Q}^{(1)} = \hat{Q}^{(1),o} \). Last but not least, Propositions B.1, B.2 and Lemmas B.1–B.3 ensure that the intersection of all the above defined events has probability at least \( 1 - C_0p_n^{-2} \).

Subsequent Iterations Fix the choice of \( \gamma_{nj} \) such that (A.43) holds. Here, we extend the above calculation up to the \( 2K \)-th iteration. The goal is to show that with high probability,
\[
\hat{Q}^{(k)} = \hat{Q}^{(k),o}, \quad k = 1, \ldots, 2K.
\]

To this end, we first recall the events defined in (A.38) and (A.41). Note that for any \( k > 1 \), \( y_{\nu_j}^{(k)} \) and \( w_{\nu_j}^{(k)} \) can be defined analogously by replacing \( \hat{q}_{jH}^{(0),o} \) with \( \hat{q}_{jH}^{(k),o} \). For any \( 1 \leq j \leq m \) and \( k \geq 0 \), \( \hat{q}_{jH}^{(k),o} \) is a function of \( V \) and \( Z_H \), which is independent of \( Z_L \), and so both \( y_{\nu_j}^{(k)} \) and \( w_{\nu_j}^{(k)} \) follow the \( N(0, 1) \) distribution. Thus, we define events
\[
\begin{align*}
\left\{ |y_{\nu_j}^{(k)}| \leq \sqrt{7 \log(p_n)}, 1 \leq j \leq m, \forall \nu \in L, 1 \leq k \leq 2K \right\}, & \quad (A.45) \\
\left\{ |w_{\nu_j}^{(k)}| \leq \sqrt{7 \log(p_n)}, 1 \leq j \leq m, \forall \nu \in L, 1 \leq k \leq 2K \right\}. & \quad (A.46)
\end{align*}
\]

Under our assumption, for any given \( \nu \), \( j \), and \( k \), the tail bound of standard normal distribution gives
\[
P\left\{ |y_{\nu_j}^{(k)}| > \sqrt{7 \log(p_n)} \right\} \leq \frac{C}{\sqrt{\log(p_n)}} p_n^{-7/2}.
\]

Since \( m \) is fixed, \( \text{card}(L) \leq p_n \), and \( K = O(\sqrt{n \log(p_n)}) \), we obtain that
\[
P\left\{ |y_{\nu_j}^{(k)}| \leq \sqrt{7 \log(p_n)}, 1 \leq j \leq m, \forall \nu \in L, 1 \leq k \leq 2K \right\}
\geq 1 - \sum_{j=1}^{m} \sum_{\nu \in L} \sum_{k=1}^{2K} P\left\{ |y_{\nu_j}^{(k)}| > \sqrt{7 \log(p_n)} \right\} \geq 1 - C_0p_n^{-2}.
\]

The same argument extends to the event in (A.46). Finally, on the intersection of (A.45), (A.46), and the other events involved in the calculation of the first iteration, we can essentially repeat the arguments for the first iteration \( 2K - 1 \) times to obtain (A.44). This complete the proof of Lemma 6.4.
B Auxiliary Results

We collect here some auxiliary results used in the proofs.

First, we present two \(\sin \theta\) theorems. One is for principal subspaces, which comes from [3] and is used extensively in the proofs of Lemmas 6.1, 6.2, and 6.3. The other is for singular subspaces, which comes from [35] and is used in the proof of Proposition 6.1.

**Theorem B.1** (\(\sin \theta\) theorem for principal subspaces). Let \(A\) and \(A+E\) be symmetric matrices satisfying

\[
A = \begin{bmatrix} G_0 & G_1 \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} G'_0 \\ G'_1 \end{bmatrix} \quad \text{and} \quad A+E = \begin{bmatrix} F_0 & F_1 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} F'_0 \\ F'_1 \end{bmatrix},
\]

with \(G'_0 G_1 = 0\), \(G_1\) orthonormal, and \([F_0 F_1]\) orthogonal. If the eigenvalues of \(A_0 G'_0 G_0\) are contained in an interval \((a, b)\), and the eigenvalues of \(A_1\) are excluded from the interval \((a-\delta, b+\delta)\) for some \(\delta > 0\), then

\[
L(\text{ran}(G_0), \text{ran}(F_0)) = \|F'_0 G_0\|^2 \leq \frac{\|F'_1 E G_0\|^2}{\delta^2 \sigma_{\text{min}}(G_0)^2}.
\]

**Theorem B.2** (\(\sin \theta\) theorem for singular subspaces). Suppose that \(p \geq k\). Let \(p \times k\) matrices \(A\) and \(B = A+E\) both have full column rank, and let \(w_{k+1}, \ldots, w_p\) be orthonormal vectors spanning \(\text{ran}(B)^\perp\), the orthogonal complement of \(\text{ran}(B)\). Take \(W = [w_{k+1}, \ldots, w_p]\), and define \(R = A' W\). If \(\sigma_{\text{min}}(A) \geq \delta > 0\), then

\[
L(\text{ran}(A), \text{ran}(B)) \leq \|R\|^2 / \delta^2 \leq \|E\|^2 / \delta^2.
\]

The last inequality holds as \(R = A' W = -E' W\).

Next, we present two probabilistic bounds on matrix norms.

**Proposition B.1** ([22] and [3]). Let \(Y\) be an \(n \times p\) matrix with i.i.d. \(N(0, 1)\) entries. For \(t_n = 6 \sqrt{\log n / n}\) and any fixed \(c > 0\), there exist \(n_0(c) > 0\), such that for any \(n \geq n_0(c)\),

\[
\mathbb{P}\left\{ \|\frac{1}{n} Y' Y - I\| \geq 2 \left( \sqrt{\frac{p}{n}} + \frac{p}{n} + c t_n \right) \right\} \leq 2 n^{-c^2}.
\]

**Proposition B.2.** Let \(Y \in \mathbb{R}^{n \times l}\) and \(Z \in \mathbb{R}^{n \times m}\) be two independent matrices with i.i.d. \(N(0, 1)\) entries. Then for any \(0 < a < \frac{1}{7} \sqrt{n}\) and \(b > 0\),

\[
\mathbb{P}\left\{ \|Y' Z\| \geq n \sqrt{1 + \frac{a}{n}} \left( \sqrt{\frac{l}{n}} + \sqrt{\frac{m}{n}} + \frac{b}{\sqrt{n}} \right) \right\} \leq (l \land m) e^{-3a^2/16} + e^{-\delta^2/2}.
\]

**Proof.** Without loss of generality, suppose \(l \leq m\). Define \(Y = [\tilde{y}_1, \ldots, \tilde{y}_l]\), where \(\tilde{y}_i = y_i / \|y_i\|_2\) with \(y_i\) the \(i\)-th column of \(Y\). Then, we obtain \(\|Y' Z\| \leq (\max_{1 \leq i \leq l} \|y_i\|_2) \|Y' Z\|\). Note that \(\|y_i\|_2^2\) are i.i.d. \(\chi_n^2\) random variables. We apply Lemma B.2 to obtain that

\[
\mathbb{P}\{\|y_i\|_2 > \sqrt{n + a}, i = 1, \ldots, l\} \leq \sum_{i=1}^l \mathbb{P}\{\|y_i\|_2 > \sqrt{n + a}\} \leq l e^{-3a^2/16}.
\]
On the other hand, $\bar{Y}'Z$ is an $l \times m$ matrix with i.i.d. $N(0,1)$ entries. So, Lemma B.3 leads to $\mathbb{P}\{\|\bar{Y}'Z\| > \sqrt{l} + \sqrt{m} + b\} \leq e^{-b^2/2}$. Therefore, we obtain

$$\mathbb{P}\{\|Y'Z\| \geq n\sqrt{1 + a/n} \left(\sqrt{\frac{l}{n}} + \sqrt{\frac{m}{n}} + \frac{b}{\sqrt{n}}\right)\} \leq \mathbb{P}\{\|y_i\| > \sqrt{n + a}, i = 1, \ldots, l\}$$

$$+ \mathbb{P}\{\|\bar{Y}'Z\| > \sqrt{l} + \sqrt{m} + b\} \leq le^{-3a^2/16} + e^{-b^2/2}.$$

This completes the proof.

Finally, we list three probabilistic bounds which have been used multiple times.

**Lemma B.1.** Let $x \sim N(0,1)$. For any $t \geq 0$, $\mathbb{P}\{|x| > t\} \leq \sqrt{\frac{2}{\pi}} t e^{-t^2/2}$.

**Lemma B.2 ([13]).** Let $\chi_n^2$ denote a Chi-square random variable with $n$ degrees of freedom. Then

$$\mathbb{P}\{\chi_n^2 > n(1 - \epsilon)\} \leq e^{-n\epsilon^2/4}, \text{ when } 0 < \epsilon < 1,$$

$$\mathbb{P}\{\chi_n^2 > n(1 + \epsilon)\} \leq e^{-3n\epsilon^2/16}, \text{ when } 0 < \epsilon < \frac{1}{2},$$

$$\mathbb{P}\{\chi_n^2 > n(1 + \epsilon)\} \leq \frac{\sqrt{2}}{\epsilon\sqrt{n}} e^{-n\epsilon^2/4}, \text{ when } 0 < \epsilon < n^{1/16}, n \geq 16.$$

**Lemma B.3 ([4]).** Let $Y$ be an $n \times p$ matrix with i.i.d. $N(0,1)$ entries. Then, for any $t > 0$,

$$\mathbb{P}\{\|Y\| > \sqrt{n} + \sqrt{p} + t\} \leq e^{-t^2/2}.$$

**References**

[1] A.A. Amini and M.J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Stat.*, 37(5B):2877–2921, 2009.

[2] T.W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Stat.*, 34(1):122–148, 1963.

[3] A. d’Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, 49:434–448, 2007.

[4] K.R. Davidson and S. Szarek. *Handbook on the Geometry of Banach Spaces*, volume 1, chapter Local operator theory, random matrices and Banach spaces, pages 317–366. Elsevier Science, 2001.

[5] C. Davis and W.M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.

[6] D. L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.*, 1:100–115, 1993.

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96:1348–1360, 2001.

[8] O. N. Feldheim and S. Sodin. A universality result for the smallest eigenvalues of certain sample covariance matrices. *Geom. Funct. Anal.*, 20:88–123, 2010.
[9] G.H. Golub and C.F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore and London, 3rd edition, 1996.

[10] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24:417–441, 498–520, 1933.

[11] D.C. Hoyle and M. Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Phys. Rev. E*, 69(2):026124, 2004.

[12] I.M. Johnstone. On the distribution of the largest eigenvalue in principal component analysis. *Ann. Stat.*, 29:295–327, 2001.

[13] I.M. Johnstone. Thresholding for weighted chi-squared. *Stat. Sinica*, 11:691–704, 2001.

[14] I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.*, 104(486):682–693, 2009.

[15] I. Jolliffe, N. Trendafilov, and M. Uddin. A modified principal component technique based on the LASSO. *J. Comput. Graph. Stat.*, 12:531–547, 2003.

[16] S. Jung and J. S. Marron. PCA consistency in high dimension, low sample size context. *Ann. Stat.*, 37(6B):4104–4130, 2009.

[17] A.Y. Lu. *Sparse principal component analysis for functional data*. PhD thesis, Stanford University, Stanford, CA, 2002.

[18] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 2009.

[19] B. Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Stat.*, 36(6):2791–2817, 2008.

[20] B. Nadler. Discussion of “On consistency and sparsity for principal components analysis in high dimensions” by Johnstone and Lu. *J. Am. Stat. Assoc.*, 104(486):694–697, 2009.

[21] A. Onatski. Asymptotics of the principal components estimator of large factor models with weak factors. Available at http://www.columbia.edu/~ao2027/onatski_weak_factors_revision1.pdf, 2009.

[22] D. Paul. *Nonparametric estimation of principal components*. PhD thesis, Stanford University, 2005.

[23] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sinica*, 17(4):1617–1642, 2007.

[24] D. Paul and I.M. Johnstone. Augmented sparse principal component analysis for high dimensional data. Available at http://anson.ucdavis.edu/~debashis/techrep/augmented-spca.pdf, 2007.

[25] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag. Ser. 6*, 2(11):559–572, 1901.

[26] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, NY, 2nd edition, 2005.

[27] P. Reimann, C. Van den Broeck, and G.J. Bex. A Gaussian scenario for unsupervised learning. *J. Phys. A*, 29:3521–3535, 1996.

[28] D. Shen, H. Shen, and J. S. Marron. Consistency of sparse PCA in high dimension, low sample size contexts. 2011.
[29] H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.*, 99:1015–1034, 2008.

[30] G.W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990.

[31] R.S. Tsay. *Analysis of Financial Time Series*. Wiley-Interscience, 2005.

[32] M.O. Ulfarsson and V. Solo. Sparse variable PCA using geodesic steepest descent. *IEEE T. Signal Process.*, 56(12):5823–5832, 2008.

[33] K. Varmuza and P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, 2009.

[34] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE T. Acoust. Speech*, 33(2):387–392, 1985.

[35] P.-A. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12:99–111, 1972.

[36] D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[37] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J. Comput. Graph. Stat.*, 15:265–286, 2006.
This figure "testsigorig.jpg" is available in "jpg" format from:

http://arxiv.org/ps/1112.2432v1
This figure "testsigwave.jpg" is available in "jpg" format from:

http://arxiv.org/ps/1112.2432v1