arXiv:1211.2671v6 [math.ST] 13 Nov 2013

A GENERAL FRAMEWORK FOR CONSISTENCY OF PRINCIPAL COMPONENT ANALYSIS

By Dan Shen[*,†], Haipeng Shen[‡] and J. S. Marron[§]

University of North Carolina at Chapel Hill

A general asymptotic framework is developed for studying consistency properties of principal component analysis (PCA). Our framework includes several previously studied domains of asymptotics as special cases and allows one to investigate interesting connections and transitions among the various domains. More importantly, it enables us to investigate asymptotic scenarios that have not been considered before, and gain new insights into the consistency, subspace consistency and strong inconsistency regions of PCA and the boundaries among them. We also establish the corresponding convergence rate within each region. Under general spike covariance models, the dimension (or the number of variables) discourages the consistency of PCA, while the sample size and spike information (the relative size of the population eigenvalues) encourages PCA consistency. Our framework nicely illustrates the relationship among these three types of information in terms of dimension, sample size and spike size, and rigorously characterizes how their relationships affect PCA consistency.

1. Introduction. Principal Component Analysis (PCA) is an important visualization and dimension reduction tool which finds orthogonal directions reflecting maximal variation in the data. This allows the low dimensional representation of data, by projecting data onto these directions. PCA is usually obtained by an eigen decomposition of the sample variance-covariance matrix of the data. Properties of the sample eigenvalues and eigenvectors have been analyzed under several domains of asymptotics.

In this paper, we develop a *general asymptotic framework* to explore interesting transitions among the various asymptotic domains. The general framework includes the traditional asymptotic setups as special cases, which allows careful study of the connections among the various setups, and more

[*]Corresponding Author

[†]Partially supported by NSF grant DMS-0854908

[‡]Partially supported by NSF grants DMS-1106912 and CMMI-0800575, NIH Challenge Grant 1 RC1 DA029425-01, and the Xerox Foundation UAC Award

[§]Partially supported by NSF grants DMS-0606577 and DMS-0854908

AMS 2000 subject classifications: Primary 62H25; Secondary 62F12

Keywords and phrases: PCA, High Dimension, Spike Model, Consistency

importantly it investigates scenarios that have not been considered before, and offers new insights into the *consistency* (in the sense that the angle between estimated and population eigen direction tends to 0, or the inner product tends to 1) and *strong-inconsistency* (where the angle tends to $\frac{\pi}{2}$, i.e., the inner product tends to 0) properties of PCA, along with some technically challenging convergence rates.

Existing asymptotic studies of PCA roughly fall into three domains:

(a) the **classical** domain of asymptotics, under which the sample size $n \to \infty$ and the dimension d is fixed (hence the ratio $\frac{n}{d} \to \infty$). For example, see [2, 3, 11, 14, 19].
(b) the **random matrix** theory domain, where both the sample size n and the dimension d increase to infinity, with the ratio $\frac{n}{d} \to c$, a constant mostly assumed to be within $(0, \infty)$. Representative work includes [9, 13, 26, 29] from the statistical physics literature, as well as [4, 5, 8, 15, 16, 20–23] from the statistics literature.
(c) the **high dimension low sample size (HDLSS)** domain of asymptotics, which is based on the limit, as the dimension $d \to \infty$, with the sample size n being fixed (hence the ratio $\frac{n}{d} \to 0$). HDLSS asymptotics was originally studied by [10], and recently rediscovered by [12]. PCA has been studied using the HDLSS asymptotics by [1, 17].

PCA consistency and (strong) inconsistency, defined in terms of angles, are important properties that have been studied before. A common technical device is the spike covariance model, initially introduced by Johnstone [15]. This model has been used in this context by, for example, Nadler [21], Johnstone and Lu [16], and Jung and Marron [17]. An interesting, more general model has been considered by Benaych-Georges and Nadakuditi [8].

Under the spike model, the first few eigenvalues are much larger than the others. A *major point of the present paper* is that there are three critical features whose relationships drive the consistency properties of PCA, namely

(1) the *sample information*: the sample size n, which has a *positive* contribution to, i.e. *encourages*, the consistency of the sample eigenvectors.
(2) the *variable information*: the dimension d, which has a *negative* contribution to, i.e. *discourages*, the consistency of the sample eigenvectors.
(3) the *spike information*: the relative sizes of the several leading eigenvalues, which also has a positive contribution to the consistency.

Our general framework considers increasing sample size n, increasing dimension d, and increasing spike information. It clearly characterizes how their relationships determine the regions of consistency and strong-inconsistency of PCA, along with the boundary in-between. In addition, our theorems

demonstrate the transitions among the existing domains of asymptotics, and for the first time to the best of our knowledge, enable one to understand the connections among them. Note that the classical domain ((a) above) assumes increasing sample size n while fixing dimension d; the random matrix domain ((b) above) assumes increasing sample size n and increasing dimension d, while fixing the spike information; the HDLSS domain ((c) above) fixes the sample size, and increases the dimension and the spike information; thus each of these three domains is a boundary case of our framework. Finally, our theorems also contain novel results on rates of convergence.

Sections 3 and 4 formally state very general theorems for the single and multiple component spike models, respectively. For illustration purposes only, in this section we first consider Examples 1.1 and 1.2 under some strong assumptions, which provide intuitive insight regarding the much more general theory presented in Sections 3 and 4.

For these two illustrative examples, the three types of information and their relationships can be mathematically quantified by two indices, namely the *spike index* α and the *sample index* γ. Within the context of these examples, we point out the significant contributions of our results in comparison with existing results. The comparisons and connections are graphically illustrated in Figure 1 and discussed below.

EXAMPLE 1.1. *(Single-component spike model) Assume that $X_1, \ldots, X_n$ are random sample vectors from a d-dimensional normal distribution $N(0, \Sigma)$, where the sample size $n \sim d^\gamma$ ($\gamma \geq 0$ is defined as the sample index) and the covariance matrix Σ has the eigenvalues as*

$$\lambda_1 \sim d^\alpha, \lambda_2 = \cdots = \lambda_d = 1, \alpha \geq 0,$$

where the constant α is defined as the spike index.

Theorem 3.1, when applied to this example, suggests that the maximal sample eigenvector is consistent when $\alpha+\gamma > 1$ (grey region in Figure 1(A)), and strongly inconsistent when $0 \leq \alpha+\gamma < 1$ (white triangle in Figure 1(A)). These very general new results nicely connect with many existing ones:

- ***Previous Results I - the classical domain**:*
  For this example, Theorem 1 of Anderson [2] implied that for fixed dimension d and finite eigenvalues, when the sample size $n \to \infty$ (i.e. $\gamma \to \infty$, the limit on the vertical axis), the maximal sample eigenvector is consistent. This case is the upper left corner of Figure 1(A).
- ***Previous Results II - the random matrix domain**:*
  (a) The results of Johnstone and Lu [16] appear on the vertical axis in Panel (A) where the spike index $\alpha = 0$ (as they fix the spike

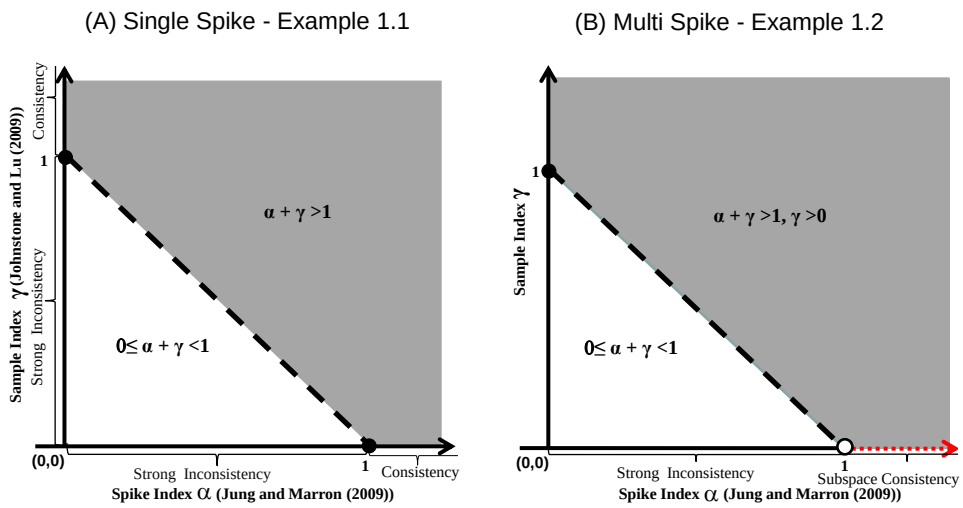

FIG 1. *General consistency and strong inconsistency regions for PCA, as a function of the spike index α and the sample index γ. Panel (A) - single spike model in Example 1.1: PCA is consistent on the grey region ($\alpha + \gamma > 1$), and strongly inconsistent on the white triangle ($0 \leq \alpha + \gamma < 1$). Panel (B) - multiple spike model in Example 1.2: the first m sample PCs are consistent on the grey region ($\alpha + \gamma > 1, \gamma > 0$), subspace consistent on the dotted line segment ($\alpha > 1, \gamma = 0$) on the horizontal axis, and strongly inconsistent on the white triangle ($0 \leq \alpha + \gamma < 1$).*

information): the first sample eigenvector is consistent when the sample index $\gamma > 1$ and strongly inconsistent when $\gamma < 1$.

(b) *Nadler [21] explored the interesting boundary case of $\alpha = 0, \gamma = 1$ (i.e. $\frac{d}{n} \to c$ for a constant c) and showed that $< \hat{u}_1, u_1 >^2 \xrightarrow{\text{a.s}} \frac{((\lambda_1-1)^2-c)_+}{(\lambda_1-1)^2+c(\lambda_1-1)}$, where $\hat{u}_1$ and u_1 are the first sample and population eigenvector. This result appears in Panel (A) as the single solid circle $\gamma = 1$ on the vertical axis.*

- ***Previous Results III - the HDLSS domain**:*

(a) *The theorems of Jung and Marron [17] are represented on the horizontal axis in Panel (A) when the sample index $\gamma = 0$ (as they fix the sample size): the maximal sample eigenvector is consistent with the first population eigenvector when the spike index $\alpha > 1$ and strongly inconsistent when $\alpha < 1$.*

(b) *Jung et al. [18] deeply explored limiting behavior at the boundary $\alpha = 1, \gamma = 0$ (i.e. $\frac{d}{\lambda_1} \to c$ for a constant c) and showed that $< \hat{u}_1, u_1 >^2 \Rightarrow \frac{\chi_n^2}{\chi_n^2+c}$, where " $\Rightarrow$ " means convergence in distribution and χ_n^2 is the chi-squared distribution with n degrees of freedom.*

This result appears in Panel (A) as the single solid circle $\alpha = 1$ on the horizontal axis.

- ***Our Results*** *hence nicely connect existing domains of asymptotics, and give a much more complete characterization for the regions of PCA consistency, subspace consistency, and strong inconsistency. We also investigate asymptotic properties of the other sample eigenvectors and all the sample eigenvalues.*

EXAMPLE 1.2. *(Multiple-component spike model) Assume that the covariance matrix Σ in Example 1.1 has the following eigenvalues*

$$\lambda_j = \begin{cases} c_j d^{\alpha} & \text{if } j \leq m, \\ 1 & \text{if } j > m, \end{cases} \quad \alpha \geq 0,$$

where m is a finite positive integer, the constants $c_j, j = 1, \cdots, m$, are positive and satisfy that $c_j > c_{j+1} > 1$, $j = 1, \cdots, m-1$.

Theorem 4.1, when applied to this example, shows that the first m sample eigenvectors are individually consistent with corresponding population eigenvectors when $\alpha + \gamma > 1, \gamma > 0$ (the grey region in Figure 1(B)), instead of being subspace consistent [17], and strongly inconsistent when $\alpha + \gamma < 1$ (the white triangle in Panel (B)). This very general new result connects with many others in the existing literature:

- ***Previous Results I - the classical domain****:*
  For this example, Theorem 1 of Anderson [2] implied that for fixed dimension d and finite eigenvalues, when the sample size $n \to \infty$ (i.e. $\gamma \to \infty$, the limit on the vertical axis), the first m sample eigenvectors are consistent, while the other sample eigenvectors are subspace consistent. This case is the upper left corner of Figure 1(B).
- ***Previous Results II - the random matrix domain****:*
  Paul [23] explored asymptotic properties of the first m eigenvectors and eigenvalues in the interesting boundary case of $\alpha = 0, \gamma = 1$, i.e., $\frac{d}{n} \to c$ with $c \in (0,1)$ and showed that $< \hat{u}_j, u_j >^2 \xrightarrow{\text{a.s}} \frac{((\lambda_j - 1)^2 - c)_+}{(\lambda_j - 1)^2 + c(\lambda_j - 1)}$ for $j = 1, \cdots, m$. This result appears in Panel (B) as the solid circle $\gamma = 1$ on the vertical axis. Paul and Johnstone [24] considered a similar framework but from a minimax risk analysis perspective. Nadler [21] and Johnstone and Lu [16] did not study multiple spike models.
- ***Previous Results III - the HDLSS domain****:*
  The theorems of Jung and Marron [17] are valid on the horizontal axis in Panel (B) where the sample index $\gamma = 0$. In particular, for this

example, their results showed that the first m sample eigenvectors are not separable when the spike index $\alpha > 1$ (the horizontal dotted red line segment), instead they are subspace consistent with their corresponding population eigenvectors, and are strongly inconsistent when the spike index $\alpha < 1$ (the horizontal solid line segment). They and Jung et al. [18] did not study the asymptotic behavior on the boundary - the single open circle $(\alpha = 1, \gamma = 0)$ on the horizontal axis.

- ***Our Results** cover the classical domain, and are stronger than what [17] obtained: the increasing sample size enables us to separate out the first few leading eigenvectors and characterize individual consistency, while only subspace consistency was obtained by [17].*

The organization of the rest of the paper is as follows. Section 2 first introduces our notations and several relevant consistency concepts. Section 3 then presents the theoretical results of single-component spike models, stating the asymptotic properties of the sample eigenvalues and eigenvectors under our general framework. Section 3.1 first considers single-component spike models with the increasing sample size n, and Section 3.2 then studies single-component spike models where the sample size n is fixed. Section 4 studies multiple-component spike models. For easy access to the main ideas, Section 4.1 first studies models with distinct eigenvalues, while Section 4.2 then considers models where the eigenvalues are grouped. Section 5 contains some discussion about the asymptotic properties of PCA when some small eigenvalues equal to zero and the challenges to obtain non-asymptotic results. Section 6 contains the technical proofs of the main theorem.

2. Notations and Concepts. We now introduce some necessary notations, and define consistency concepts relevant for our asymptotic study.

2.1. *Notation.* Let the population covariance matrix be Σ, whose eigen decomposition is

$$\Sigma = U \Lambda U^T,$$

where Λ is the diagonal matrix of population eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$, and U is the matrix of corresponding eigenvectors $U = [u_1, \ldots, u_d]$.

As in Jung and Marron [17], assume that $X_1, \ldots, X_n$ are i.i.d. d-dimensional random sample vectors and have the following representation

$$X_i = \sum_{j=1}^{d} z_{i,j} u_j, \tag{2.1}$$

where the $z_{i,j}$'s are i.i.d random variables with zero mean, unit variance and finite fourth moment. An important special case is that the $z_{i,j}$'s follow the standard normal distribution $N(0,1)$.

ASSUMPTION 2.1. *$X_1, \ldots, X_n$ are a random sample having the distribution described by* (2.1).

Denote the sample covariance matrix by $\hat{\Sigma} = n^{-1}XX^T$, where $X = [X_1, \ldots, X_n]$. Note that $\hat{\Sigma}$ can also be decomposed as

$$\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T, \tag{2.2}$$

where $\hat{\Lambda}$ is the diagonal matrix of sample eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_d$ and $\hat{U}$ is the matrix of corresponding sample eigenvectors where $\hat{U} = [\hat{u}_1, \ldots, \hat{u}_d]$.

Below we introduce asymptotic notations that will be used in our theoretical studies. Assume that $\{\xi_n : n = 1, \ldots, \infty\}$ is a sequence of random variables, and $\{a_n : n = 1, \ldots, \infty\}$ is a sequence of constant values.

- Denote $\xi_n = \mathrm{o}_{\mathrm{a.s}}(a_n)$ if $\lim_{n\to\infty} \frac{\xi_n}{a_n} = 0$ almost surely.
- Denote $\xi_n = \mathrm{O}_{\mathrm{a.s}}(a_n)$ if $\overline{\lim}_{n\to\infty} \left|\frac{\xi_n}{a_n}\right| \leq z$ almost surely, where the random variable z satisfies $P(0 < z < \infty) = 1$..
- Denote $\xi_n \overset{\mathrm{a.s}}{\sim} a_n$ if $c_2 \leq \underline{\lim}_{n\to\infty} \frac{\xi_n}{a_n} \leq \overline{\lim}_{n\to\infty} \frac{\xi_n}{a_n} \leq c_1$ almost surely, for two constants $c_1 \geq c_2 > 0$.

In addition, we introduce the following notions to help understand the assumptions on the population eigenvalues in our theorems. Assume that $\{a_k : k = 1, \ldots, \infty\}$ and $\{b_k : k = 1, \ldots, \infty\}$ are two sequence of constant values, where k can stand for either n or d.

- Denote $a_k \gg b_k$ if $\lim_{k\to\infty} \frac{b_k}{a_k} = 0$.
- Denote $a_k \sim b_k$ if $c_2 \leq \underline{\lim}_{k\to\infty} \frac{a_k}{b_k} \leq \overline{\lim}_{k\to\infty} \frac{a_k}{b_k} \leq c_1$ for two constants $c_1 \geq c_2 > 0$.

2.2. *Concepts.* Below we list three important concepts relevant for consistency and strong inconsistency, some of which are modified from the related concepts given by Jung and Marron [17] and Shen et al. [27].

Let $\hat{u}_j$ be any normalized sample estimator of u_j for $j = 1, \ldots, [n \wedge d]$.

- **Consistency with rate** a_n: The estimator $\hat{u}_j$ is consistent with its population counterpart u_j with the convergence rate a_n if $|< \hat{u}_j, u_j >| = 1 + \mathrm{O}_{\mathrm{a.s}}(a_n)$. For example, $a_n = \left(\frac{n\lambda_1}{d}\right)^{\frac{1}{2}}$.
- **Strong inconsistency with rate** a_n: $\hat{u}_j$ is strongly inconsistent with u_j with the convergence rate a_n if $| < \hat{u}_j, u_j > | = \mathrm{O}_{\mathrm{a.s}}(a_n)$.

Let H be an index set, e.g. $H = \{m+1, \cdots, d\}$. Define $S = \text{span}\{u_k, k \in H\}$ to be the linear span generated by $\{u_k, k \in H\}$.

- **Subspace consistency with rate** a_n: $\hat{u}_j$, $j \in H$, is subspace consistent with S with convergence rate a_n if

$$\text{angle}(\hat{u}_j, S) = \text{O}_{a.s}(a_n), \tag{2.3}$$

  where the angle between the estimator $\hat{u}_j$ and the subspace S is the angle between the estimator and its projection onto the subspace, see Jung and Marron [17]. For further clarification, we provide a graphical illustration of the angle in Section B of the supplement [28].

Terminology: In the following results, for simple general formulations, the term - *consistent at the rate* a_n - will mean "$\text{O}_{\text{a.s}}(a_n)$" in situations where $a_n \to 0$. Otherwise it means "$\text{o}_{\text{a.s}}(a_n)$". Similarly for *strong inconsistency* and *subspace consistency*.

3. Single component spike models. Below we state our main theorems for single-component spike models. In Section 3.1, we study the asymptotic properties of PCA with increasing sample size n. In Section 3.2 , we investigate the asymptotic properties of PCA with fixed n.

3.1. *Cases with increasing sample size* n. We first state in Theorem 3.1 one of our main theoretical results regarding PCA consistency under our general framework. We then offer several remarks in regards to the conditions of the theorem as well as the connection between our results and the earlier ones in the literature.

To fix ideas, we assume the maximal eigenvalue λ_1 dominates the other eigenvalues. WLOG, we assume that as $n \to \infty$ or $d \to \infty$,

ASSUMPTION 3.1. $\lambda_1 > \lambda_2 \to \cdots \to \lambda_d \to c_\lambda$, *where* c_λ *is a constant.*

As discussed in the Introduction, we consider the delicate balance among the positive *sample information* n, the positive *spike information* λ_1, and the negative *variable information* d, and characterize the various PCA consistency and strong-inconsistency regions.

Theorem 3.1 below suggests that the asymptotic properties of the sample eigenvalues and eigenvectors depend on the relative strength of the positive information and the negative information, as particularly measured by two ratios: $\frac{d}{n\lambda_1}$ and $\frac{d}{n}$. The value of $\frac{d}{n\lambda_1}$ determines whether the maximal sample eigenvalue is separable from the other eigenvalues, and further determines

the consistency of the maximal sample eigenvector. The value of $\frac{d}{n}$ determines the asymptotic properties of the second and higher sample eigenvalues and eigenvectors.

The following discussion and the scenarios in Theorem 3.1 are arranged according to a decreasing amount of positive information:

- Theorem 3.1(a): If the amount of positive information dominates the amount of negative information up to the maximal eigenvalue, i.e. $\frac{d}{n\lambda_1} \to 0$, then the maximal sample eigenvector is consistent, and the other sample eigenvectors are subspace consistent. In addition, the asymptotic properties of sample eigenvalues and eigenvectors whose index are greater than 1 depend on the value $\frac{d}{n}$.
- Theorem 3.1(b): On the other hand, if the amount of negative information always dominates, i.e. $\frac{d}{n\lambda_1} \to \infty$, then the sample eigenvalues are asymptotically indistinguishable, and the sample eigenvectors are strongly inconsistent.

THEOREM 3.1. *Under Assumptions 2.1 and 3.1, as $n \to \infty$, the following results hold.*

(a) If $\frac{d}{n\lambda_1} \to 0$, then $\frac{\hat{\lambda}_1}{\lambda_1} \xrightarrow{\text{a.s}} 1$, $\hat{u}_1$ is consistent with u_1, and the other $\hat{u}_j$ are subspace consistent with $S = span\{u_k, k \geq 2\}$. In addition,

i. If $\frac{d}{n} \to 0$, then $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1$, $j = 2, \cdots, [n \wedge (d-1)]$, and $\hat{\lambda}_{[n\wedge d]} = \mathrm{O}_{\text{a.s}}(1)$. The consistency rate for $\hat{u}_1$ and the subspace consistency rate for the other $\hat{u}_j$ are both $\left(\frac{1}{\lambda_1}\right)^{\frac{1}{2}}$.

ii. If $\frac{d}{n} \to \infty$, then $\frac{\hat{\lambda}_j}{\lambda_j} \overset{\text{a.s}}{\to} \frac{d}{n}$ for $j = 2, \cdots, [n \wedge d]$; $\hat{u}_1$ is consistent with rate $\left(\frac{d}{n\lambda_1}\right)^{\frac{1}{2}}$, and the other $\hat{u}_j$ are strongly inconsistent with rate $\left(\frac{n\lambda_j}{d}\right)^{\frac{1}{2}}$.

iii. If $\overline{\lim}\frac{d}{n} = c^*$ $(0 < c^* \leq \infty)$, then for $j = 2, \cdots, [n \wedge d]$, $\overline{\lim}\hat{\lambda}_j \leq c \times \overline{\lim}\frac{d}{n}$ almost surely, where c is some constant. The consistency rate for $\hat{u}_1$ and the subspace consistency rate for the other $\hat{u}_j$ are both $\left(\frac{d_n}{n\lambda_1}\right)^{\frac{1}{2}}$, where $\{d_n\}$ is a sequence converging to c^*.

(b) If $\frac{d}{n\lambda_1} \to \infty$, then $\hat{\lambda}_j \overset{\text{a.s}}{\to} c_\lambda \frac{d}{n}$ for $j = 1, \cdots, [n \wedge d]$, and the corresponding eigenvectors $\hat{u}_j$ are strongly inconsistent with rate $(\frac{n\lambda_j}{d})^{\frac{1}{2}}$.

Having stated the main results for single-component spike models, we now offer several remarks regarding the conditions assumed in Theorem 3.1 and make connections with existing results about PCA consistency.

- If Assumption 3.1 is replaced by the alternative assumption $\lambda_1 \gg \lambda_2 \sim \cdots \lambda_d \sim 1$, then except for $\frac{\hat{\lambda}_1}{\lambda_1} \xrightarrow{\text{a.s}} 1$ in Scenario (a), all other " $\xrightarrow{\text{a.s}}$ " for the sample eigenvalues should be replaced by " $\overset{\text{a.s}}{\sim}$ ". The results for the sample eigenvectors remain the same.
- An assumption of the form (3.1), i.e $\lambda_2 \to \cdots \to \lambda_d \to c_\lambda$, or else $\lambda_2 \sim \cdots \lambda_d \sim 1$ is needed to obtain general convergence results for the non-spike sample eigenvalues $\hat{\lambda}_j$, $j > 1$ under the wide range of scenarios: $\frac{d}{n} \to 0$, $\frac{d}{n} \to \infty$ or $\overline{\lim} \frac{d}{n} = c^*$ $(0 < c^* \leq \infty)$. When one focusses only on the spike eigenvalue, a weaker assumption, such as the slowly decaying non-spike eigenvalues assumed by Bai and Yao (2012) [7], is enough. Then the spike condition $\lambda_1 \gg \lambda_2$ is enough to generate the consistency properties of $\hat{\lambda}_1$ and $\hat{u}_1$ in Scenario (a). In that case, the behaviors of the other sample eigenvalues and eigenvectors are very case-wise to formulate in general.
- Nadler [21] and Johnstone and Lu (2009) [16] studied the properties of the first sample eigenvalue and eigenvector under the normality assumption for fixed λ_1. Furthermore, if d is fixed, Scenario (a) of Theorem 3.1 degenerates to the case studied by Anderson (1963) [2].
- Assuming fixed λ_1 and $\frac{d}{n} \to c$ with c being a constant, Nadler [21], Johnstone and Lu [16] and Benaych-Georges and Nadakuditi [8] obtained the results in *Previous Results II - the random matrix domain* in Example 1.1, which indicate that, as $n \to \infty$, the maximal sample eigenvector $\hat{u}_1$ is consistent when $\frac{d}{n} \to 0$, and inconsistent when $\frac{d}{n} \to \infty$. Our Theorem 3.1 includes this as a special case. In addition, Theorem 3.1 offers more than just relaxing the fixed λ_1 assumption: it characterizes how an increasing λ_1 interacts with the ratio $\frac{d}{n}$, derives the corresponding convergence rate, and also studies the asymptotic properties of the higher order sample eigenvalues and eigenvectors, all of which have not been investigated before.

3.2. *Cases with fixed n.* Theorem 3.2 summarize the results for the fixed n cases (i.e. the HDLSS domain). In comparison with Jung and Marron (2009) [17], we make more general assumptions on the population eigenvalues, and obtain the corresponding convergence rate results; furthermore, we obtain almost sure convergence, instead of convergence in probability [17].

Consider the $z_{i,j}$ in (2.1), and define

$$\widetilde{Z}_j = (z_{1,j}, \cdots, z_{n,j})^T, \quad j = 1, \cdots, d, \tag{3.1}$$

which are needed here to describe the asymptotic properties of the sample eigenvalues in HDLSS settings. In addition, define $K = \lim_{d\to\infty} \frac{\sum_{j=2}^{d} \lambda_j}{nd}$.

THEOREM 3.2. *Under Assumptions 2.1 and 3.1, for fixed n, as $d \to \infty$, the following results hold.*

(a) If $\frac{d}{\lambda_1} \to 0$, then $\frac{\hat{\lambda}_1}{\lambda_1} \xrightarrow{\text{a.s}} \frac{\widetilde{Z}_1^T \widetilde{Z}_1}{n}$, where $\widetilde{Z}_1$ is defined in (3.1), and the rest of the non-zero $\frac{\hat{\lambda}_j}{d} \xrightarrow{\text{a.s}} K$. In addition, $\hat{u}_1$ is consistent with rate $\left(\frac{d}{\lambda_1}\right)^{\frac{1}{2}}$, and the rest of the $\hat{u}_j$ are strongly inconsistent with rate $\left(\frac{\lambda_j}{d}\right)^{\frac{1}{2}}$.

(b) If $\frac{d}{\lambda_1} \to \infty$, then the non-zero $\frac{\hat{\lambda}_j}{d} \xrightarrow{\text{a.s}} K$, and the corresponding $\hat{u}_j$ are strongly inconsistent with rate $(\frac{\lambda_j}{d})^{\frac{1}{2}}$, respectively.

Some comments about the conditions and results of Theorem 3.2

- Assumption 3.1 can be replaced by $\lambda_1 \gg \lambda_2 \sim \cdots \lambda_d \sim 1$. The results remain the same.
- Even if the non-spike eigenvalues λ_j, $j > 2$, decay slowly, the condition $\lambda_1 \gg \lambda_2$ is enough to generate the same properties for $\hat{\lambda}_1$ and $\hat{u}_1$ as in Scenario (a).
- If Assumption 2.1 is strengthened to a normality assumption, then $\frac{\hat{\lambda}_1}{\lambda_1} \xrightarrow{\text{a.s}} \frac{\chi_n^2}{n}$ in Scenario (a).
- Assumption 2.1 assumes that the $z_{i,j}$'s are i.i.d rather than ρ-mixing as in [17]. Thus, convergence in probability in [17] is strengthened to almost sure convergence here.

4. Multiple component spike models. We consider multiple spike models with finite $m(\in [1, n \wedge d])$ dominating spikes. In Section 4.1, we study models where the dominating eigenvalues are distinct. In Section 4.2, we consider the cases where the eigenvalues are not all distinct, by introducing the concept of *tiered eigenvalues*.

4.1. *Multiple component spike models with distinct eigenvalues.*

4.1.1. *Cases with increasing sample size n.* WLOG, we assume that the first m population eigenvalues have different strength and dominate the rest population eigenvalues, which are asymptotically equivalent.

ASSUMPTION 4.1. *as* $n \to \infty$, $\lambda_1 > \cdots > \lambda_m > \lambda_{m+1} \to \cdots \to \lambda_d \to c_\lambda$.

A useful quantity, for distinguishing the various cases among eigenvectors in the coming theorems, is

$$a_l = \max_{1 \le k \le l} \frac{\lambda_{k+1}}{\lambda_k}, \quad l = 1, \cdots, m.$$

This lower bound on the consecutive relative gap among the first l eigenvalues provides a critical measure of the separation between the l-th sample eigenvector and the first $l-1$ sample eigenvectors.

Below we first state the main theoretical results in Theorem 4.1, and follow up with some remarks about the theorem conditions and the connections between the theorem and the existing results in the literature.

Similar to Theorem 3.1, Theorem 4.1 states the asymptotic properties of the sample eigenvalues and eigenvectors in a trichotomous manner, separated by the size of $\frac{d}{n\lambda_j}$, which again measures the relative strength of the positive information and the negative information. The three scenarios below and in Theorem 4.1 are arranged in a decreasing order of the amount of the positive information:

- Theorem 4.1(a): If the amount of positive information dominates the amount of negative information up to the mth spike, i.e. $\frac{d}{n\lambda_m} \to 0$, then each of the first m sample eigenvector is consistent, and the additional ones are subspace consistent;
- Theorem 4.1(b): Otherwise, if the amount of positive information dominates the amount of negative information only up to the hth spike ($h \in [1, m]$), i.e. $\frac{d}{n\lambda_h} \to 0$ and $\frac{d}{n\lambda_{h+1}} \to \infty$, then each of the first h sample eigenvector is consistent, and each of the remaining higher-order sample eigenvector is strongly-inconsistent;
- Theorem 4.1(c): Finally, if the amount of negative information always dominates, i.e. $\frac{d}{n\lambda_1} \to \infty$, then the sample eigenvalues are asymptotically indistinguishable, and the sample eigenvectors are strongly inconsistent.

THEOREM 4.1. *Under Assumptions 2.1 and 4.1, as* $n \to \infty$, *the following results hold.*

(a) If $\frac{d}{n\lambda_m} \to 0$, then $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1$ for $1 \le j \le m$. In addition, $\hat{u}_j$ are consistent with u_j for $1 \le j \le m$ and the other $\hat{u}_j$ are subspace consistent with $S = \text{span}\{u_k, k \ge m+1\}$.

(b) *If there exists a constant h, $1 \le h \le m$, such that $\frac{d}{n\lambda_h} \to 0$ and $\frac{d}{n\lambda_{h+1}} \to \infty$, then $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1$ for $1 \le j \le h$, and the other non-zero $\hat{\lambda}_j \overset{\text{a.s}}{\to} c_\lambda \frac{d}{n}$. In addition, $\hat{u}_j$ are consistent with rate $\left(a_j \vee \frac{d}{n\lambda_j}\right)^{\frac{1}{2}}$ for $1 \le j \le h$, and the other $\hat{u}_j$ are strongly inconsistent with rate $\left(\frac{n\lambda_j}{d}\right)^{\frac{1}{2}}$.*

(c) *If $\frac{d}{n\lambda_1} \to \infty$, then the non-zero $\hat{\lambda}_j \overset{\text{a.s}}{\to} c_\lambda \frac{d}{n}$, and the corresponding $\hat{u}_j$ are strongly inconsistent with rate $(\frac{n\lambda_j}{d})^{\frac{1}{2}}$.*

We now discuss the properties of the rest of the sample eigenvalues and the convergence rate in Scenario (a), and the conditions needed in the theorem and how the results connect with existing ones in the literature.

- The special case of $m = 1$ is Theorem 3.1 for single spike models.
- As in Theorem 3.1, Scenario (a) in Theorem 4.1 contains three different cases.
  - i. If $\frac{d}{n} \to 0$, then $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1$, $j = m+1, \cdots, [n \wedge (d-m)]$ and the rest of the non-zero $\hat{\lambda}_j = \mathrm{O}_{\text{a.s}}(1)$. In addition, the consistency rates for the $\hat{u}_j$ are $\left(a_j \vee \frac{1}{\lambda_j}\right)^{\frac{1}{2}}$ for $1 \le j \le m$, and the subspace consistency rates for the other $\hat{u}_j$ are $\left(a_m \vee \frac{1}{\lambda_m}\right)^{\frac{1}{2}}$.
  - ii. The case $\frac{d}{n} \to \infty$ is considered in Scenario (b) ($h = m$) of Theorem 4.1.
  - iii. If $\overline{\lim}\frac{d}{n} = c^*$ ($0 < c^* \le \infty$), then for $j = m+1, \cdots, [n \wedge d]$, $\overline{\lim}\hat{\lambda}_j \le c \times \overline{\lim}\frac{d}{n}$ almost surely, where c is some constant. Also the consistency rates for the $\hat{u}_j$ are $\left(a_j \vee \frac{d_n}{n\lambda_j}\right)^{\frac{1}{2}}$ for $1 \le j \le m$, and the subspace consistency rates for the rest of the $\hat{u}_j$ are $\left(\frac{d_n}{n\lambda_m}\right)^{\frac{1}{2}}$, where $\{d_n\}$ is a sequence converging to c^*.
- If Assumption 4.1 is replaced by the alternative assumption $\lambda_1 > \cdots > \lambda_m \gg \lambda_{m+1} \sim \cdots \lambda_d \sim 1$, then we still have $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1$, $1 \le j \le m$, as in Scenario (a) and $1 \le j \le h$ as in Scenario (b), but all other results of the form " $\xrightarrow{\text{a.s}}$ " for the sample eigenvalues should be replaced by " $\overset{\text{a.s}}{\sim}$ ". The results for the sample eigenvectors remain same.
- Even if the non-spike eigenvalues λ_j, $j > m$, decay slowly, the condition $\lambda_1 > \cdots > \lambda_m \gg \lambda_{m+1}$ is enough to generate the consistency

properties of $\hat{\lambda}_j$ and $\hat{u}_j$, for $1 \leq j \leq m$ in Scenario (a) and $1 \leq j \leq h$ in Scenario (b).

- In Theorem 4.1, consider the special case of fixed dimension d and $\infty > \lambda_1 > \cdots > \lambda_m > \lambda_{m+1} \to \cdots \to \lambda_d \to c_\lambda$. Then, Theorem 4.1(a) is consistent with the classical results implied by Theorem 1 of Anderson [2].
- Considering fixed $\lambda_1, \cdots, \lambda_m$ and $\frac{d}{n} \to c$, where $c \in (0,1)$, Paul [23] obtained results that are applicable to Example 1.2 to obtain *Previous Results II - the random matrix domain* in . As one can see, our Theorem 4.1 relaxes the assumptions of $\frac{d}{n} \to c \in (0,1)$ and that $\lambda_1, \cdots, \lambda_m$ are fixed. In addition, we characterize how increasing $\lambda_1, \cdots, \lambda_m$ interact with the ratio $\frac{d}{n}$ along with the corresponding convergence rates, and study the asymptotic properties of the higher order sample eigenvalues and eigenvectors, all of which have not been investigated before.

4.1.2. *Cases with fixed n.* The following Theorem 4.2 considers cases with fixed n. The multiple spike condition in Assumption 4.1 now becomes that the first m population eigenvalues are of the different order and dominate the other population eigenvalues, which are asymptotically equivalent:

ASSUMPTION 4.2. *as $d \to \infty$, $\lambda_1 \gg \cdots \gg \lambda_m \gg \lambda_{m+1} \sim \cdots \sim \lambda_d \sim 1$.*

Note that for fixed n and $d \to \infty$, assuming $\lambda_j > \lambda_{j+1}$ can not asymptotically separate the corresponding sample eigenvalues $\hat{\lambda}_j$ and $\hat{\lambda}_{j+1}$. Thus, we need to replace Assumption 4.1 with Assumption 4.2 to asymptotically separate the first m sample eigenvalues. Define $K = \lim_{d\to\infty} \frac{\sum_{j=m+1}^{d} \lambda_j}{nd}$.

THEOREM 4.2. *Under Assumptions 2.1 and 4.2, for fixed n, as $d \to \infty$, the following results hold.*

(a) If there exists a constant h, $1 \leq h \leq m$, such that $\frac{d}{\lambda_h} \to 0$ and $\frac{d}{\lambda_{h+1}} \to \infty$, then $\frac{\hat{\lambda}_j}{\lambda_j} \overset{\text{a.s}}{\to} \frac{\widetilde{Z}_j^T \widetilde{Z}_j}{n}$ for $1 \leq j \leq h$, where $\widetilde{Z}_j$ is defined in (3.1), and the other $\hat{\lambda}_j$'s satisfy $\frac{\hat{\lambda}_j}{d} \overset{\text{a.s}}{\to} K$. In addition, $\hat{u}_j$ are consistent with rate $\left(a_j \vee \frac{d}{\lambda_j}\right)^{\frac{1}{2}}$ for $1 \leq j \leq h$, and the other $\hat{u}_j$'s are strongly inconsistent with rate $\left(\frac{\lambda_j}{d}\right)^{\frac{1}{2}}$.

(b) If $\frac{d}{\lambda_1} \to \infty$, then the non-zero $\frac{\hat{\lambda}_j}{d} \overset{\text{a.s}}{\to} K$, and the corresponding $\hat{u}_j$ are strongly inconsistent with rate $(\frac{\lambda_j}{d})^{\frac{1}{2}}$.

Some comments are made for the results of Theorem 4.2

- If $m = 1$, Theorem 4.2 becomes Theorem 3.2.
- Even if the non-spike eigenvalues λ_j, $j > m$, decay slowly, the condition $\lambda_1 \gg \cdots \gg \lambda_m \gg \lambda_{m+1}$ is enough to guarantee the same properties for $\hat{\lambda}_j$ and $\hat{u}_j$, where $1 \leq j \leq h$, in Scenario (a).
- If Assumption 2.1 is strengthened to a normality assumption, then $\frac{\hat{\lambda}_j}{\lambda_j} \overset{\text{a.s}}{\to} \frac{\chi_n^2}{n}$ for $1 \leq j \leq h$ in Scenario (a).

4.2. *Multiple component spike models with tiered eigenvalues.* We now consider models where the m eigenvalues can be grouped into r tiers, where the eigenvalues within the same tier are either the same or have the same limit or are of the same order, and the eigenvalues within different tiers have either different limits or are of different orders.

4.2.1. *Cases with increasing sample size n.* To fix ideas, the first m eigenvalues are grouped into r tiers where there are $q_l(> 0)$ eigenvalues in the lth tier with $\sum_{l=1}^{r} q_l = m$. Define $q_0 = 0$, $q_{r+1} = d - \sum_{l=1}^{r} q_l$, and the index set of the eigenvalues in the lth tier as

$$H_l = \left\{ \sum_{k=0}^{l-1} q_k + 1, \sum_{k=0}^{l-1} q_k + 2, \cdots, \sum_{k=0}^{l-1} q_k + q_l \right\}, \quad l = 1, \cdots, r+1. \tag{4.1}$$

Assume the eigenvalues in the lth tier have the same limit $\delta_l(> 0)$, i.e.

Assumption 4.3. $\lim_{n\to\infty} \frac{\lambda_j}{\delta_l} = 1,\ j \in H_l, l = 1, \cdots, r.$

The above assumption suggests that it is impossible to separate the sample eigenvectors whose indexes are in the same tier, and motives us to consider subspace consistency. In addition, we assume that the population eigenvalues from different tiers are asymptotically different and dominate the other population eigenvalues that are asymptotically equivalent:

Assumption 4.4. *as $n \to \infty$, $\delta_1 > \cdots > \delta_r > \lambda_{m+1} \to \cdots \to \lambda_d \to c_r$.*

Under the above setup, we have the following Theorem 4.3 which suggests that the eigenvalues with the same limit can not be consistently estimated individually; the corresponding eigenvector estimates are either subspace consistent with the linear space spanned by the eigenvectors, or strongly inconsistent. Similar to the earlier theorems, Theorem 4.3 is arranged according to a decreasing amount of positive information:

- Theorem 4.3(a): If the amount of positive information dominates the amount of negative information up to the rth tier, i.e. $\frac{d}{n\delta_r} \to 0$, then the estimates for the eigenvectors in the first r tiers are subspace consistent, and the estimates for the rest are also subspace consistent (but) at a different rate;
- Theorem 4.3(b): Otherwise, if the amount of positive information dominates the amount of negative information only up to the hth tier ($h \in [1, r]$), i.e. $\frac{d}{n\delta_h} \to 0$ and $\frac{d}{n\delta_{h+1}} \to \infty$, then the estimates for the eigenvectors in the first h tiers are subspace consistent, and the estimates for the rest eigenvectors are strongly-inconsistent;
- Theorem 4.3(c): Finally, if the amount of negative information always dominates, i.e. $\frac{d}{n\lambda_1} \to \infty$, then the sample eigenvalues are asymptotically indistinguishable, and the sample eigenvectors are strongly inconsistent.

In this setting, one key to distinguishing the cases in the theorem is

$$a_l = \max_{1 \le k \le l} \frac{\delta_{k+1}}{\delta_k}, \quad l = 1, \cdots, r, \tag{4.2}$$

where $\delta_{r+1} = 1$, which measures the separation between the sample eigenvectors in the l-th tier and those in the first $l-1$ tiers. Define the subspace $S_l = \text{span}\{u_k, k \in H_l\}$ for $l = 1, \cdots, r+1$.

THEOREM 4.3. *Under Assumptions 2.1, 4.3 and 4.4, as $n \to \infty$, the following results hold.*

(a) If $\frac{d}{n\delta_r} \to 0$, then $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1$ for $1 \le j \le m$. In addition, $\hat{u}_j$ are subspace consistent with S_l $j \in H_l, l = 1, \cdots, r+1$.

(b) If there exists a constant h, $1 \le h \le r$, such that $\frac{d}{n\delta_h} \to 0$ and $\frac{d}{n\delta_{h+1}} \to \infty$, then $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1$ for $j \in H_l, l = 1, \cdots, h$, and the other non-zero $\hat{\lambda}_j \xrightarrow{\text{a.s}} c_\lambda \frac{d}{n}$. In addition, $\hat{u}_j$ are subspace consistent with S_l with rate $\left(a_l \vee \frac{d}{n\delta_l}\right)^{\frac{1}{2}}$ for $j \in H_l, l = 1, \cdots, h$, and the other $\hat{u}_j$ are strongly inconsistent with rate $\left(\frac{n\lambda_j}{d}\right)^{\frac{1}{2}}$.

(c) If $\frac{d}{n\delta_1} \to \infty$, then the non-zero $\hat{\lambda}_j \xrightarrow{\text{a.s}} c_\lambda \frac{d}{n}$, and the corresponding $\hat{u}_j$ are strongly inconsistent with rate $(\frac{n\lambda_j}{d})^{\frac{1}{2}}$.

The following comments can be made for the results of Theorem 4.3.

- If each tier only contains one eigenvalue, i.e. $q_1 = \cdots = q_r = 1$, then Theorem 4.3 simplifies to Theorem 4.1.
- There are additional eigenvalue properties, which are entirely parallel to those in the 2nd remark following Theorem 4.1. The corresponding convergence rates in Scenario (a) of Theorem 4.3 can be attained by replacing λ_j by δ_j in Scenario (a) of Theorem 4.1.
- Assumption 4.4 can be replaced by $\delta_1 > \cdots > \delta_r \gg \lambda_{m+1} \sim \cdots \lambda_d \sim 1$. Then, the consistency results of the first r tiers of sample eigenvalues in Scenario (a) or the first h tiers in Scenario (b) remain the same, while all other results of the form " $\xrightarrow{\text{a.s}}$ " for the sample eigenvalues should be replaced by " $\overset{\text{a.s}}{\sim}$ ". The results for the sample eigenvectors remain same.
- Even if the non-spike eigenvalues λ_j, $j > m$, decay slowly, the condition $\delta_1 > \cdots > \delta_r \gg \lambda_{m+1}$ is enough to generate the same properties for $\hat{\lambda}_j$ and $\hat{u}_j$, with $j \in H_l$, $l \leq r$ as in Scenario (a) and $j \in H_l$, $l \leq h$ as in Scenario (b).
- The cases covered by Theorem 4.3 were not studied by Paul (2007) [23], which required the eigenvalues to be individually estimable.
- In Theorem 4.3, the dimension d can be fixed. In addition, suppose $\infty > \delta_1 > \cdots > \delta_r > \lambda_{m+1} \to \cdots \to \lambda_d \to c_\lambda$ and the eigenvalues satisfying (4.3). Then, the results of Theorem 4.3(a) are consistent with the classical asymptotic subspace consistency results implied by Theorem 1 of Anderson (1963) [2].

4.2.2. *Cases with fixed n.* Similar results can be obtained for the fixed n cases (i.e. the HDLSS domain) as summarized below in Theorem 4.4. For that, we assume that as $d \to \infty$, the first m eigenvalues fall into r tiers, where the eigenvalues in the same tier are asymptotically equivalent, as stated in the following assumption:

Assumption 4.5. $\lambda_j \sim \delta_l,\ j \in H_l, l = 1, \cdots, r.$

Different from Assumption 4.3 for diverging sample size n, now with a fixed n, the eigenvalues within the same tier are assumed to be of the same order, rather than of the same limit when n increases to ∞. As we will see below in Theorem 4.4, one can not separately estimate the eigenvalues of the same order when n is fixed, which is feasible with an increasing n as long as they do not have the same limit as previously shown in Theorem 4.3.

In addition, we assume that the population eigenvalues from different tiers are of different orders and dominate the rest eigenvalues which are asymptotically equivalent:

ASSUMPTION 4.6. *as* $d \to \infty$, $\delta_1 \gg \cdots \gg \delta_r \gg \lambda_{m+1} \sim \cdots \sim \lambda_d \sim 1$.

Note that for fixed n and $d \to \infty$, the assumption $\delta_l > \delta_{l+1}$ can not guarantee asymptotic separation of the corresponding sample eigenvalues $\hat{\lambda}_j$ for $j \in H_l$ and $\hat{\lambda}_j$ for $j \in H_{l+1}$. Thus, we need to replace Assumption 4.4 with Assumption 4.6 in order to asymptotically separate the first r subgroups of sample eigenvalues. Define

$$K = \lim_{d\to\infty} \frac{\sum_{j=m+1}^{d} \lambda_j}{nd} \quad \text{and} \quad A_l^* = \frac{1}{n} \sum_{k\in H_l} \widetilde{Z}_k \widetilde{Z}_k^T, \quad l = 1, \cdots, r,$$

which are used to describe the asymptotic properties of the sample eigenvalues in HDLSS settings.

THEOREM 4.4. *Under Assumptions 2.1, 4.5 and 4.6, for fixed* n, *as* $d \to \infty$, *the following results hold.*

(a) If there exists a constant h, $1 \le h \le r$, *such that* $\frac{d}{\delta_h} \to 0$ *and* $\frac{d}{\delta_{h+1}} \to \infty$, *then for* $j \in H_l, l = 1, \cdots, h$, *we have almost surely that*

$$\lambda_{\min}(A_l^*) \times \min_{k\in H_l} \lambda_k \le \hat{\lambda}_j \le \lambda_{\max}(A_l^*) \times \max_{k\in H_l} \lambda_k, \tag{4.3}$$

and the other $\hat{\lambda}_j$*'s satisfy* $\frac{\hat{\lambda}_j}{d} \xrightarrow{\text{a.s}} K$. *In addition,* $\hat{u}_j$ *are subspace consistent with* S_l *with rate* $\left(a_l \vee \frac{d}{\delta_l}\right)^{\frac{1}{2}}$ *for* $j \in H_l, l = 1, \cdots, h$, *and the other* $\hat{u}_j$*'s are strongly inconsistent with rate* $\left(\frac{\lambda_j}{d}\right)^{\frac{1}{2}}$.

(b) If $\frac{d}{\delta_1} \to \infty$, *then the non-zero* $\frac{\hat{\lambda}_j}{d} \xrightarrow{\text{a.s}} K$, *and the corresponding* $\hat{u}_j$ *are strongly inconsistent with rate* $\left(\frac{\lambda_j}{d}\right)^{\frac{1}{2}}$.

The following comments can be made about the results of Theorem 4.4.

- If each tier only contains one eigenvalue, i.e. $q_1 = \cdots = q_r = 1$, then (4.3) becomes $\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} \frac{\widetilde{Z}_j^T \widetilde{Z}_j}{n}$ and Theorem 4.3 becomes Theorem 4.1.
- Even if the non-spike eigenvalues λ_j, $j > m$, decay slowly, the condition $\delta_1 \gg \cdots \gg \delta_r \gg \lambda_{m+1}$ can still guarantee the same properties for $\hat{\lambda}_j$ and $\hat{u}_j$, with $j \in H_l$, $l \le h$, in Scenario (a).

5. Discussion. Throughout the paper, we assume that the small eigenvalues have the same limit or the same order as 1, i.e. $\lambda_{m+1} \to \cdots \to \lambda_d \to c_\lambda$ or $\lambda_{m+1} \sim \cdots \sim \lambda_d \sim 1$. In fact, this is a convenient WLOG choice. Our results remain valid when these small eigenvalues are not of the same order, and even when some of them are 0. For example, suppose $\lambda_{d_1+1} = \cdots = \lambda_d = 0$ for $m+1 < d_1 < d$. As shown in Section C of the supplementary material [28], the asymptotic properties of PCA are independent of the basis choice for the d-dimensional space. If the population eigenvectors u_j, $j = 1, \ldots, d$, are chosen as the basis of the d-dimensional space, the population covariance matrix becomes

$$\Sigma = \Lambda = \begin{pmatrix} \Lambda_1 & 0_{d_1 \times (d-d_1)} \\ 0_{(d-d_1)\times d_1} & 0_{(d-d_1)\times(d-d_1)} \end{pmatrix}, \quad \text{where } \Lambda_1 = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{d_1} \end{pmatrix},$$

and $0_{k\times l}$ is the k-by-l zero matrix. Then, the asymptotic properties of PCA under the population covariance matrix Σ is the same as those under the covariance matrix Λ_1. Therefore, we only need to replace the dimension d by the *effective* dimension d_1, and all the earlier results can be obtained.

It would be interesting but challenging to explore the non-asymptotic results such as large deviations of the angle between the sample and population eigenvectors. The properties of sample eigenvectors heavily depend on the sample eigenvalues' properties. Since we are not aware of any non-asymptotic results for the eigenvalues of the random matrix, then it appears to be challenging to obtain non-asymptotic results for sample eigenvectors.

6. Proofs. We now provide detailed proofs for the general Theorem 4.3. To save space, proofs for Theorems 3.1, 3.2, 4.1, 4.2, and 4.4 (which are often similar, and simpler) are provided in the supplement [28]. We first provide some overview in Section 6.1 and list four lemmas in Section 6.2, and then prove the asymptotic properties of the sample eigenvalues and the sample eigenvectors in Sections 6.3 and 6.4, respectively.

In this paper, we study the consistency and strong inconsistency of PCA through the angle or the inner product between a sample eigenvector and the corresponding population eigenvector. We first note that this angle has a nice invariance property: it doesn't depend on the specific choice of the basis for the d-dimensional space, as discussed in details in the supplement [28]. Given this invariance property, for the rest of the paper, we choose to use the population eigenvectors u_j, $j = 1, \ldots, d$, as the basis of the d-dimensional space, which is equivalent to assuming that X_i, $i = 1, \ldots, n$, is a d-dimensional random vector with mean zero and a diagonal covariance

matrix as $\Sigma = \Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_d\}$. This will simplify our mathematical analysis, see for example (6.13) and (6.14).

We consider general cases where the first m eigenvalues are grouped into r tiers, and WLOG we assume that $\lambda_1 = \cdots = \lambda_{q_1} = \delta_1, \cdots, \lambda_{\sum_{l=0}^{r-1} q_l+1} = \cdots = \lambda_m = \delta_r$ where $q_0 = 0$ and q_l are positive integers for $l \geq 1$. In addition, we assume that each ratio δ_j/δ_i, where $1 \leq i < j \leq r$, converges to a constant less than 1 as $n \to \infty$. (The following arguments can be extended to cases where only the upper limits of the ratios exist as stated in the theorems, through taking a converging subsequence of the diverging sequence of n.)

6.1. *Overview.* Our proof makes use of the connection between the sample covariance matrix $\hat{\Sigma}$ and its dual matrix $\hat{\Sigma}_D$, which share the same nonzero eigenvalues. Since $\Sigma = \Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_d\}$, then it follows from (2.1) and (3.1) that the dual matrix can be expressed as

$$\hat{\Sigma}_D = n^{-1} X^T X = \frac{1}{n} \sum_{j=1}^{d} \lambda_j \widetilde{Z}_j \widetilde{Z}_j^T,$$

which can be rewritten as the sum of two matrices as follows:

$$\hat{\Sigma}_D = A + B, \quad \text{with} \quad A = \frac{1}{n} \sum_{j=1}^{m} \lambda_j \widetilde{Z}_j \widetilde{Z}_j^T, \quad B = \frac{1}{n} \sum_{j=m+1}^{d} \lambda_j \widetilde{Z}_j \widetilde{Z}_j^T. \tag{6.1}$$

The proof involves the following several steps. First, we study the asymptotic properties of the eigenvalues of A and B in Lemmas 6.1 and 6.2, respectively. Then, the *Wielandt's Inequality* (Rao [25]), now restated as Lemma 6.4, enables us to establish the asymptotic properties of the eigenvalues of the dual matrix in Section 6.3. Finally, we derive the asymptotic properties of the sample eigenvectors of $\hat{\Sigma}$ in Section 6.4. Some intuitive ideas are provided in the supplement [28] to help understanding the proof.

6.2. *Lemmas.* We list four lemmas that are used in our proof. Lemmas 6.1 and 6.2 are proven in our online supplement, the proofs of which need the following Lemma 6.3 that studies asymptotic properties of the largest and smallest non-zero eigenvalues of random matrix.

LEMMA 6.1. *As $n \to \infty$, the eigenvalues of the matrix A in (6.1) satisfy*

$$\frac{\lambda_j(A)}{\lambda_j} \xrightarrow{\text{a.s}} 1, \quad \textit{for} \quad j = 1, \cdots, m,$$

where $\lambda_j(A)$ denotes the jth largest eigenvalue of the matrix A.

LEMMA 6.2. *As $n \to \infty$, the eigenvalues of the matrix B in (6.1) satisfy that, for $j = 1, \cdots, [n \wedge (d-m)]$,*

$$\frac{\lambda_j(B)}{\lambda_{j+m}} \overset{\text{a.s}}{\to} 1, \quad \text{for} \quad \frac{d}{n} \to 0, \tag{6.2}$$

$$\frac{\lambda_j(B)}{\lambda_{j+m}} \overset{\text{a.s}}{\to} \frac{d}{n}, \quad \text{for} \quad \frac{d}{n} \to \infty, \tag{6.3}$$

and almost surely,

$$\overline{\lim} \lambda_1(B) \leq c \times \overline{\lim} \frac{d}{n}, \quad \textit{otherwise}, \tag{6.4}$$

where c is a constant.

REMARK 6.1. *If $\lambda_{m+1} \to \cdots \to \lambda_d$ is relaxed to $\lambda_{m+1} \sim \cdots \sim \lambda_d$, then "$\overset{\text{a.s}}{\to}$" is replaced by "$\overset{\text{a.s}}{\sim}$" in (6.2) and (6.3).*

LEMMA 6.3. *Suppose $B = \frac{1}{q} V V^T$ where V is an $p \times q$ random matrix composed of i.i.d. random variables with zero mean, unit variance and finite fourth moment. As $q \to \infty$ and $\frac{p}{q} \to c \in [0, \infty)$, the largest and smallest non-zero eigenvalues of B converge almost surely to $(1+\sqrt{c})^2$ and $(1-\sqrt{c})^2$, respectively.*

REMARK 6.2. *Lemma 6.3 is known as the Bai-Yin's law [6]. As in Remak 1 of [6], the smallest non-zero eigenvalue is the $p-q+1$ smallest eigenvalue of B for $c > 1$.*

LEMMA 6.4. *(Wielandt's Inequality [25]). If A, B are $p \times p$ real symmetric matrices, then for all $j = 1, \ldots, p$,*

$$\left\{ \begin{array}{ccc} \lambda_j(A) & + & \lambda_p(B) \\ \lambda_{j+1}(A) & + & \lambda_{p-1}(B) \\ & \vdots & \\ \lambda_p(A) & + & \lambda_j(B) \end{array} \right\} \leq \lambda_j(A+B) \leq \left\{ \begin{array}{ccc} \lambda_j(A) & + & \lambda_1(B) \\ \lambda_{j-1}(A) & + & \lambda_2(B) \\ & \vdots & \\ \lambda_1(A) & + & \lambda_j(B) \end{array} \right\}.$$

6.3. *Asymptotic properties of the sample eigenvalues.* We now study the asymptotic properties of the sample eigenvalues $\hat{\lambda}_j$, for $j = 1, \cdots, [n \wedge d]$, which are the same as the eigenvalues of the dual matrix $\hat{\Sigma}_D$, denoted as $\lambda_j(\hat{\Sigma}_D) = \lambda_j(A+B)$.

6.3.1. *Scenario (a) in Theorem 4.3.* Note that $\frac{d}{n\lambda_m} \to 0$ ($\delta_r = \lambda_m$) contains three different cases: $\frac{d}{n} \to 0$, ∞ or $\overline{\lim}\frac{d}{n} = c^*$ ($0 < c^* \leq \infty$). The proofs are different for each case and are provided separately below.

Consider the first one: $\frac{d}{n} \to 0$. If in addition we have $\lambda_m \to \infty$, then Lemma 6.4 suggests that

$$\frac{\lambda_j(A)}{\lambda_j} \leq \frac{\hat{\lambda}_j}{\lambda_j} \leq \frac{\lambda_j(A)}{\lambda_j} + \frac{\lambda_1(B)}{\lambda_j}, \tag{6.5}$$

which, together with $\lambda_m \to \infty$, (6.2) and Lemma 6.1, yields that

$$\frac{\hat{\lambda}_j}{\lambda_j} \overset{\text{a.s}}{\to} 1, \quad j = 1, \cdots, m. \tag{6.6}$$

Instead, if $\lambda_m < \infty$, according to Theorem 1 ($c = 0$) of [5], we still have (6.6). In addition, according to Lemma 6.4, we have that

$$\frac{\lambda_j(B)}{\lambda_j} \leq \frac{\hat{\lambda}_j}{\lambda_j} \leq \frac{\lambda_j(A)}{\lambda_j} + \frac{\lambda_1(B)}{\lambda_j}, \tag{6.7}$$

which, together with (6.2), $\lambda_j(A) = 0$ for $j \geq m+1$ and $\lambda_{m+1} \to \lambda_d \to c_\lambda$ yields that

$$\begin{aligned} &\frac{\hat{\lambda}_j}{\lambda_j} \overset{\text{a.s}}{\to} 1, \quad j = m+1, \cdots, [n \wedge (d-m)], \\ &\hat{\lambda}_j = \mathrm{O}_{\text{a.s}}(1), \quad j = [n \wedge (d-m)] + 1, \cdots, [n \wedge d]. \end{aligned} \tag{6.8}$$

Now, consider the second case: $\frac{d}{n} \to \infty$. Since $\frac{d}{n\lambda_m} \to 0$, then $\lambda_m \to \infty$, which, together with (6.3), (6.5) and Lemma 6.1, yields (6.6). In addition, it follows from (6.3), (6.7), $\lambda_j(A) = 0$ for $j \geq m+1$ and $\lambda_{m+1} \to \lambda_d$ that

$$\frac{\hat{\lambda}_j}{\lambda_j} \overset{\text{a.s}}{\to} \frac{d}{n}, \quad j = m+1, \cdots, [n \wedge d]. \tag{6.9}$$

Finally, consider the third case: $\overline{\lim}\frac{d}{n} = c^*$ ($0 < c^* \leq \infty$). Similarly, it follows from $\frac{d}{n\lambda_m} \to 0$ that $\lambda_m \to \infty$, which, jointly with (6.4), (6.5) and Lemma 6.1, yields (6.6). In addition, note that (6.4), (6.7), $\lambda_j(A) = 0$ for $j \geq m+1$ and $\lambda_{m+1} \to \lambda_d \to c_\lambda$, and then almost surely we have

$$\overline{\lim}\hat{\lambda}_j \leq c \times \overline{\lim}\frac{d}{n}, \quad j = m+1, \cdots, [n \wedge d], \tag{6.10}$$

where c is a constant.

All together, we have proved the consistency of the first m sample eigenvalues in (6.6) and the asymptotic properties of the rest of the non-zero sample eigenvalues in (6.8), (6.9) and (6.10) for Scenario (a).

6.3.2. *Scenario (b) in Theorem 4.3.* Since $\frac{d}{n\delta_{h+1}} \to \infty$, then $\frac{d}{n} \to \infty$. According to (6.3), (6.5) and $\frac{d}{n\delta_h} \to 0$, we have

$$\frac{\hat{\lambda}_j}{\lambda_j} \overset{\text{a.s}}{\to} 1, \quad j \in H_l, l = 1, \cdots, h. \tag{6.11}$$

In addition, note that

$$\frac{\lambda_j(B)n}{d} \leq \frac{\hat{\lambda}_j n}{d} \leq \frac{\lambda_j(A)n}{d} + \frac{\lambda_1(B)n}{d}, \tag{6.12}$$

which, together with $\frac{d}{n\delta_{h+1}} \to \infty$, (6.3) and Lemma 6.1, yields that

$$\hat{\lambda}_j \overset{\text{a.s}}{\to} c_\lambda \frac{d}{n}, \quad j = \sum_{k=1}^{h} q_k + 1, \cdots, [n \wedge d].$$

6.3.3. *Scenario (c) in Theorem 4.3.* Since $\frac{d}{n\delta_1} \to \infty$, then $\frac{d}{n} \to \infty$. Then it follows from (6.3), (6.12) and $\frac{d}{n\delta_1} \to \infty$ that

$$\hat{\lambda}_j \overset{\text{a.s}}{\to} c_\lambda \frac{d}{n}, \quad j = 1, \cdots, [n \wedge d].$$

6.4. *Asymptotic properties of the sample eigenvectors.* We first state two results that simplify the proof. As aforementioned, in light of the invariance property of the angle, we choose the population eigenvectors u_j, $j = 1, \ldots, d$, as the basis of the d-dimensional space. It then follows that $u_j = e_j$ where the jth component of e_j equals to 1 and all the other components equal to zero. This suggests that

$$|< \hat{u}_j, u_j >|^2 = |< \hat{u}_j, e_j >|^2 = \hat{u}_{j,j}^2, \tag{6.13}$$

and for any index set H,

$$\cos\left[\text{angle}\left(\hat{u}_j, \text{span}\{u_k, k \in H\}\right)\right] = \sum_{k \in H} \hat{u}_{k,j}^2. \tag{6.14}$$

As a reminder, the population eigenvalues are grouped into $r+1$ tiers and the index set of the eigenvalues in the lth tier H_l is defined in (4.1). Define

$$\hat{U}_{k,l} = (\hat{u}_{i,j})_{i \in H_k, j \in H_l}, \quad 1 \leq k, l \leq r+1. \tag{6.15}$$

Then, the sample eigenvector matrix $\hat{U}$ can be rewritten as the following:

$$\hat{U} = [\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_d] = \begin{pmatrix} \hat{U}_{1,1} & \hat{U}_{1,2} & \cdots & \hat{U}_{1,r+1} \\ \hat{U}_{2,1} & \hat{U}_{2,2} & \cdots & \hat{U}_{2,r+1} \\ \vdots & \vdots & & \vdots \\ \hat{U}_{r+1,1} & \hat{U}_{r+1,2} & \cdots & \hat{U}_{r+1,r+1} \end{pmatrix}.$$

To derive the asymptotic properties of the sample eigenvectors $\hat{u}_j$, we consider the three scenarios of Theorem 4.3 separately.

6.4.1. *Scenario (b) in Theorem 4.3.* Under this scenario, there exists a constant $h \in [1, r]$, such that $\frac{d}{n\delta_h} \to 0$ and $\frac{d}{n\delta_{h+1}} \to \infty$. From (6.14), to show the subspace consistency with S_l and rate $\left(a_l \vee \frac{d}{n\delta_l}\right)^{\frac{1}{2}}$, we only need to show that

$$\sum_{k \in H_l} \hat{u}_{k,j}^2 = 1 + \mathrm{o}_{a.s}(a_l) \vee \mathrm{O}_{a.s}(\frac{d}{n\delta_l}), \quad j \in H_l, l = 1, \cdots, h, \tag{6.16}$$

where, as defined in (4.2) in Section 4.2, $a_l = \max_{1 \le k \le l} \frac{\delta_{k+1}}{\delta_k}$, $l = 1, \cdots, r$. Below we provide the proof for $l = 1$. The process is similar for $l = 2, \cdots, h$, which is omitted to save space.

Note that for $l = 1$, the left hand side of (6.16) becomes the sum of squares of the column elements in the matrix $\hat{U}_{1,1}$ (defined in (6.15)). Thus, to prove (6.16), we first show that this sum of squares converges to 1, and then establish the convergence rate $a_1 \vee \frac{d}{n\delta_1}$.

For the first step, let $Z = (Z_1, \cdots, Z_n)$, where $Z_i = (z_{i,1}, \cdots, z_{i,d})^T = \Lambda^{-\frac{1}{2}} X_i$ from (2.1). Denote $S = \Lambda^{-\frac{1}{2}} \hat{U} \hat{\Lambda}^{\frac{1}{2}}$ where $\hat{U}$ is the sample eigenvector matrix and $\hat{\Lambda}$ is the sample eigenvalue matrix defined in (2.2). We can show that $SS^T = \frac{1}{n} ZZ^T$. Considering the k-th diagonal entry of the matrices on the two sides and noting that $s_{k,j} = \lambda_k^{-\frac{1}{2}} \hat{\lambda}_j^{\frac{1}{2}} \hat{u}_{k,j}$, we have the following

$$\frac{1}{n} \sum_{i=1}^{n} z_{i,k}^2 = \sum_{j=1}^{d} s_{k,j}^2 = \lambda_k^{-1} \sum_{j=1}^{d} \hat{\lambda}_j \hat{u}_{k,j}^2, \quad k = 1, \cdots, d. \tag{6.17}$$

Select the first $[n \wedge d]$ rows of Z and denote the resulting random matrix as Z^*. Then, we have $\max_{1 \le k \le [n \wedge d]} \frac{1}{n} \sum_{i=1}^{n} z_{i,k}^2 \le \lambda_{\max}(\frac{1}{n} Z^* Z^{*T})$. Note that $\frac{d}{n} \to \infty$ here, so $[n \wedge d] = n$. According to Lemma 6.3, we have $\lambda_{\max}(\frac{1}{n} Z^* Z^{*T}) \overset{\text{a.s}}{\to} 4$, which suggests that almost surely $\hat{u}_{k,j}^2 \le 4 \frac{\lambda_k}{\hat{\lambda}_j}$ for

$j = 1, \cdots, [n \wedge d]$, as $n \to \infty$. Then, given the asymptotic properties of $\hat{\lambda}_j$ in Scenario (b) of Theorem 4.3 (Section 6.3), it follows that

$$\hat{u}_{k,j}^2 = \begin{cases} \mathrm{O}_{\mathrm{a.s}}(\frac{\lambda_k}{\lambda_j}) & j \in H_l, l = 1, \cdots, h, \\ \mathrm{O}_{\mathrm{a.s}}(\frac{n\lambda_k}{d}) & j = \sum_{l=1}^{h} q_l + 1, \cdots, [n \wedge d]. \end{cases} \tag{6.18}$$

In addition, the kth diagonal entry of $S^T S$ is less than or equal to its largest eigenvalue, i.e. the largest eigenvalue of $\frac{1}{n} Z^T Z$. Hence, we have

$$\hat{\lambda}_j \sum_{k=1}^{d} \lambda_k^{-1} \hat{u}_{k,j}^2 = \sum_{k=1}^{d} s_{k,j}^2 \leq \lambda_{\max}(\frac{1}{n} Z^T Z), \quad j \in H_l, l = 1, \cdots, h. \tag{6.19}$$

According to Lemma 6.3 and $\frac{d}{n} \to \infty$, we have that

$$\lambda_{\max}(\frac{1}{n} Z Z^T) \overset{\text{a.s}}{\to} (\frac{d}{n}). \tag{6.20}$$

From (6.11), (6.19), (6.20) and $\lambda_{m+1} \to \lambda_d \to c_\lambda$, we have

$$\sum_{k=m+1}^{d} \hat{u}_{k,j}^2 = \mathrm{O}_{\mathrm{a.s}}(\frac{d}{n\lambda_j}), \quad j \in H_l, l = 1, \cdots, h. \tag{6.21}$$

Note that $\lambda_j \ll \frac{d}{n}$, for $j = \sum_{l=1}^{h} q_l + 1, \cdots, m$, which together with (6.18) and (6.21), yields that

$$\sum_{k=\sum_{t=1}^{h} q_t + 1}^{d} \hat{u}_{k,j}^2 = \mathrm{O}_{\mathrm{a.s}}\left(\frac{d}{n\lambda_j}\right), \quad j \in H_l, l = 1, \cdots, h. \tag{6.22}$$

According to (6.17) and $\lambda_k = \delta_1$, $k \in H_1$, we obtain that for $k \in H_1$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} z_{i,k}^2 = \lambda_k^{-1} \sum_{j=1}^{d} \hat{\lambda}_j \hat{u}_{k,j}^2 \quad &\leq \quad \delta_1^{-1} \sum_{j \in H_1} \hat{\lambda}_1 \hat{u}_{k,j}^2 + \delta_1^{-1} \sum_{j \notin H_1} \hat{\lambda}_{q_1+1} \hat{u}_{k,j}^2 \\ &= \quad \delta_1^{-1} (\hat{\lambda}_1 - \hat{\lambda}_{q_1+1}) \sum_{j \in H_1} \hat{u}_{k,j}^2 + \delta_1^{-1} \hat{\lambda}_{q_1+1}. \end{aligned} \tag{6.23}$$

In addition, it follows from (6.11) that $\delta_1^{-1}(\hat{\lambda}_1 - \hat{\lambda}_{q_1+1}) \xrightarrow{\text{a.s}} \delta_1^{-1}(\delta_1 - \delta_2) = (1 - c)$, and $\delta_1^{-1} \hat{\lambda}_{q_1+1} \xrightarrow{\text{a.s}} c$, where $c = \lim_{n \to \infty} \frac{\delta_2}{\delta_1} < 1$.

Note that $\frac{1}{n}\sum_{i=1}^{n} z_{i,k}^2 = 1 + \mathrm{o}_{a.s}(1)$, which together with (6.23), yields that

$$\begin{aligned} 1 + \mathrm{o}_{a.s}(1) &\leq (1-c)\underline{\lim}_{n\to\infty} \sum_{j\in H_1} \hat{u}_{k,j}^2 + c \\ &\leq (1-c)\overline{\lim}_{n\to\infty} \sum_{j\in H_1} \hat{u}_{k,j}^2 + c \leq 1, \end{aligned}$$

which yields $\sum_{j\in H_1} \hat{u}_{k,j}^2 \xrightarrow{\text{a.s}} 1$, $k \in H_1$. The above means that the sum of squares of the row elements of $\hat{U}_{1,1}$ converges to 1. Given that the sample eigenvectors all have norm 1, the sum of squares of the row or the column elements of $\hat{U}_{1,1}$ is less than or equal to 1. It then follows that the sum of squares of the column elements of $\hat{U}_{1,1}$ converges to 1, which finishes the first step of the proof.

For the second step of the proof, we need to establish the convergence rate $a_1 \vee \frac{d}{n\delta_1}$ of the above sum of squares. Having shown that the sum of squares of the row elements of $\hat{U}_{1,1}$ converges to 1, it follows that the sum of squares of the row elements of $\hat{U}_{1,2}$ converges to 0. Furthermore, the sum of the squares of the column elements of $\hat{U}_{1,2}$ converges to 0, as follows:

$$\sum_{k\in H_1} \hat{u}_{k,j}^2 = \mathrm{o}_{a.s}(1), \quad j \in H_2. \tag{6.24}$$

WLOG, we assume that $\frac{\delta_3}{\delta_2} \to 0$. (If the limit is greater than 0, we can combine the index sets H_2 and H_3 together to check whether $\frac{\delta_4}{\delta_2} \to 0$ converges to 0. If not, we keep combining the index sets together until the big jump appears.) Given that $\frac{\delta_3}{\delta_2} \to 0$, (6.18) and (6.22), it follows that

$$\sum_{k\in H_3\cup\cdots\cup H_{r+1}} \hat{u}_{k,j}^2 = \mathrm{o}_{a.s}(1), \quad j \in H_2. \tag{6.25}$$

From (6.24) and (6.25), we have that

$$\sum_{k\in H_2} \hat{u}_{k,j}^2 = 1 + \mathrm{o}_{a.s}(1), \quad j \in H_2,$$

which means that the sum of squares of the column elements of $\hat{U}_{2,2}$ also converges to 1. Again, since the sum of squares of the row or column elements of $\hat{U}_{2,2}$ is less than or equal to 1, it follows that the sum of squares of the row elements of $\hat{U}_{2,2}$ must converge to 1:

$$\sum_{j\in H_2} \hat{u}_{k,j}^2 = 1 + \mathrm{o}_{a.s}(1), \quad k \in H_2. \tag{6.26}$$

Given that $\hat{\lambda}_j \xrightarrow{a.s} \lambda_j = \delta_2$, $j \in H_2$, and (6.26), it follows that, for $k \in H_2$,

$$\begin{aligned} 1 + \mathrm{o}_{a.s}(1) &= \frac{1}{n}\sum_{i=1}^{n} z_{i,k}^2 = \lambda_k^{-1}\sum_{j=1}^{d}\hat{\lambda}_j\hat{u}_{k,j}^2 \\ &\geq \delta_2^{-1}\sum_{j\in H_1}\hat{\lambda}_j\hat{u}_{k,j}^2 + \delta_2^{-1}\sum_{j\in H_2}\hat{\lambda}_j\hat{u}_{k,j}^2 = \delta_2^{-1}\sum_{j\in H_1}\hat{\lambda}_j\hat{u}_{k,j}^2 + 1 + \mathrm{o}_{a.s}(1), \end{aligned}$$

which yields $\delta_2^{-1}\sum_{j\in H_1}\hat{\lambda}_j\hat{u}_{k,j}^2 = \mathrm{o}_{a.s}(1)$, $k \in H_2$.

For $j \in H_1$, we have that $\hat{\lambda}_j \xrightarrow{a.s} \lambda_j = \delta_1$; hence, it follows that

$$\sum_{j\in H_1}\hat{u}_{k,j}^2 = \mathrm{o}_{a.s}(\frac{\delta_2}{\delta_1}), \quad k \in H_2,$$

which yields that

$$\sum_{k\in H_2}\hat{u}_{k,j}^2 = \mathrm{o}_{a.s}(\frac{\delta_2}{\delta_1}), \quad j \in H_1. \tag{6.27}$$

In addition, from (6.18) and (6.22), we have

$$\sum_{k\in H_3\cup\cdots\cup H_{r+1}}\hat{u}_{k,j}^2 = \mathrm{o}_{a.s}(\frac{\delta_2}{\delta_1}), \quad j \in H_1. \tag{6.28}$$

From (6.27), (6.28) and $\frac{\delta_2}{\delta_1} \gg \frac{d}{n\delta_1}$, we have

$$\sum_{k\in H_1}\hat{u}_{k,j}^2 = 1 + \mathrm{o}_{a.s}(\frac{\delta_2}{\delta_1}) = 1 + \mathrm{o}_{a.s}(a_1) \vee \mathrm{O}_{a.s}(\frac{d}{n\delta_1}), \quad j \in H_1,$$

which suggests that the sum of squares of the column elements of $\hat{U}_{1,1}$ converges to 1 with the convergence rate $a_1 \vee \frac{d}{n\delta_1}$, as stated in (6.16) for $l = 1$. The proof of (6.16) is similar for $l = 2, \cdots, h$. Thus, we have shown the subspace consistency portion of the results in Scenario (b).

Finally, the strong inconsistency in Scenario (b) follows directly from (6.18) by setting $k = j$:

$$|< \hat{u}_j, u_j >|^2 = \hat{u}_{j,j}^2 = \mathrm{O}_{\mathrm{a.s}}\left(\frac{n\lambda_j}{d}\right), \quad j = \sum_{l=1}^{h} q_l + 1, \cdots, [n \wedge d]. \tag{6.29}$$

Hence, we have finished the proof of Scenario (b) in Theorem 4.3.

6.4.2. *Scenario (a) in Theorem 4.3.* As in Section 6.3.1, $\frac{d}{n\lambda_m} \to 0$ ($\delta_r = \lambda_m$) contains three different cases: $\frac{d}{n} \to 0$, ∞, or $\overline{\lim}\frac{d}{n} = c^*$ ($0 < c^* \leq \infty$), which we shall prove separately.

Consider that $\frac{d}{n} \to 0$, then (6.16) in Section 6.4.1 becomes

$$\sum_{k \in H_l} \hat{u}_{k,j}^2 = 1 + \mathrm{o}_{a.s}(a_l) \vee \mathrm{O}_{a.s}(\frac{1}{\delta_l}), \quad j \in H_l, l = 1, \cdots, r, \tag{6.30}$$

which, together with similar arguments as those in proving Scenario (b), leads to

$$\sum_{k \in H_{r+1}} \hat{u}_{k,j}^2 = 1 + \mathrm{o}_{a.s}(a_r) \vee \mathrm{O}_{a.s}(\frac{1}{\delta_r}), \quad m+1 \leq j \leq [n \wedge d]. \tag{6.31}$$

Note that if $\delta_l < \infty$, then "$\mathrm{O}_{a.s}(\frac{1}{\delta_l})$" in (6.30) and (6.31) becomes "$\mathrm{o}_{a.s}(1)$".

For the second case: $\frac{d}{n} \to \infty$, we have that "$\mathrm{O}_{a.s}(\frac{1}{\delta_l})$" in (6.30) and (6.31) are replaced by "$\mathrm{O}_{a.s}(\frac{d}{n\delta_l})$".

Finally, consider the third case: $\overline{\lim}\frac{d}{n} = c^*$ ($0 < c^* \leq \infty$). If $c^* < \infty$, then the convergence rates are the same as those in (6.30) and (6.31). Otherwise, "$\mathrm{O}_{a.s}(\frac{1}{\delta_l})$" in (6.30) and (6.31) are replaced by "$\mathrm{O}_{a.s}(\frac{d_n}{n\delta_l})$", where the sequence $\{d_n\}$ is defined in Section E.2 of the Supplement [28].

6.4.3. *Scenario (c) in Theorem 4.3.* Finally, for Scenario (c) where $\frac{d}{n\delta_1} \to 0$, the strong inconsistency in Theorem 4.3 follows from (6.18) by setting $k = j$.

SUPPLEMENTARY MATERIAL

Additional Proofs
(http://www.unc.edu/ dshen/PCA/PCASupplment.pdf). Detailed proofs are provided for Theorems 3.1, 3.2, 3.3, 4.1, 4.2, 4.4, and the necessary lemmas.

References.

[1] Ahn, J., Marron, J., Muller, K. and Chi, Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94** 760–766.

[2] Anderson, T. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics* **34** 122–148.

[3] Anderson, T. (1984). *An introduction to multivariate statistical analysis.* John Willey & Sons, New York.

[4] Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability* **33** 1643–1697.

[5] Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97** 1382–1408.

[6] Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability* 1275–1294.

[7] Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis* **106** 167–177.

[8] Benaych-Georges, F. and Nadakuditi, R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics* **227** 494–521.

[9] Biehl, M. and Mietzner, A. (1994). Statistical mechanics of unsupervised structure recognition. *Journal of Physics A: Mathematical and General* **27** 1885–1897.

[10] Casella, G. and Hwang, J. (1982). Limit expressions for the risk of James-Stein estimators. *Canadian Journal of Statistics* **10** 305–309.

[11] Girshick, M. (1939). On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics* **10** 203–224.

[12] Hall, P., Marron, J. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B* **67** 427–444.

[13] Hoyle, D. and Rattray, M. (2003). PCA learning for sparse high-dimensional data. *Europhysics Letters* **62** 117–123.

[14] Jackson, J. (1991). *A user's guide to principal components.* John Willey & Sons, New York.

[15] Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29** 295–327.

[16] Johnstone, I. and Lu, A. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.

[17] Jung, S. and Marron, J. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics* **37** 4104–4130.

[18] Jung, S., Sen, A. and Marron, J. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis* **109** 190–203.

[19] Lawley, D. (1956). Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika* **43** 128–136.

[20] Lee, S., Zou, F. and Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *The Annals of Statistics* **38** 3605–3629.

[21] Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics* **36** 2791–2817.

[22] Onatski, A. (2006). Asymptotic distribution of the principal components estimator of large factor models when factors are relatively weak. *Manuscript, Columbia University.*

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

[24] Paul, D. and Johnstone, I. (2007). Augmented Sparse Principal Component Analysis for High Dimensional Data. *Technical Report, UC Davis.*

[25] Rao, C. (2002). *Linear statistical inference and its applications.* John Willey & Sons, New York.

[26] Reimann, P., Broeck, C. and Bex, G. (1996). A Gaussian scenario for unsupervised learning. *Journal of Physics A: Mathematical and General* **29** 3521–3535.

[27] Shen, D., Shen, H. and Marron, J. (2012). Consistency of sparse PCA in high dimension and low sample size contexts. *Journal of Multivariate Analysis, forthcoming.*

[28] Shen, D., Shen, H. and Marron, J. (2012). A General Framework for Consistency of Principal Component Analysis: Supplement Materials. *Available online at http://www.unc.edu/ dshen/PCA/PCASupplement.pdf.*

[29] Watkin, T. and Nadal, J. (1994). Optimal unsupervised learning. *Journal of Physics A: Mathematical and General* **27** 1899–1915.

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
E-mail: dshen@email.unc.edu
haipeng@email.unc.edu
marron@email.unc.edu