2016

The Statistics and Mathematics of High Dimension Low Sample Size Asymptotics

Dan Shen
University of South Florida, danshen@usf.edu

Haipeng Shen
University of Hong Kong

Hongtu Zhu
University of North Carolina

J S Marron
University of North Carolina

Statistica Sinica **26** (2016), 1747-1770
doi:http://dx.doi.org/10.5705/ss.202015.0088

THE STATISTICS AND MATHEMATICS OF HIGH DIMENSION LOW SAMPLE SIZE ASYMPTOTICS

Dan Shen[1], Haipeng Shen[2], Hongtu Zhu[3] and J. S. Marron[3]

[1] *University of South Florida,* [2] *University of Hong Kong and* [3] *University of North Carolina at Chapel Hill*

Abstract: The aim of this paper is to establish several theoretical properties of principal component analysis for multiple-component spike covariance models. Our results reveal an asymptotic conical structure in critical sample eigendirections under the spike models with distinguishable (or indistinguishable) eigenvalues, when the sample size and/or the number of variables (or dimension) tend to infinity. The consistency of the sample eigenvectors relative to their population counterparts is determined by the ratio between the dimension and the product of the sample size with the spike size. When this ratio converges to a nonzero constant, the sample eigenvector converges to a cone, with a certain angle to its corresponding population eigenvector. In the High Dimension, Low Sample Size case, the angle between the sample eigenvector and its population counterpart converges to a limiting distribution. Several generalizations of the multi-spike covariance models are explored, and additional theoretical results are presented.

Key words and phrases: Big data, conical behavior, high dimension low sample size, PCA.

1. Introduction

As the field of statistics continues to evolve, there is ongoing discussion about the role that should be played by mathematics. There are those who wish to focus on data as much as possible, and thus base their work on experiential insights, and those who never actuaIly work with data, but instead develop mathematical ideas about how data should be analyzed. Among the many mathematical methods that have been used to gain statistical insights, asymptotic techniques stand out as having provided a large number of insights over the years.

What is the value of asymptotics? Some might say one should not consider asymptotics on the grounds that one never has an infinite sample size, while others view asymptotics as "understanding what happens in situations where the sample size grows". Some take the latter notion to an extreme by insisting that all asymptotics should "follow some type of sampling process". We bring in a different view of asymptotics, in which the focus is moved beyond mere sampling to any type of limiting operation that gives statistical insights. It will

be seen that the role of asymptotic insights can go beyond academic indulgence to understanding even the most basic statistical concepts in challenging modern data analytic settings.

A currently fashionable statistical topic is *Big Data*. This notion goes well beyond a large sample size and includes high dimension, but, while the size of modern data sets indeed presents serious statistical challenges, *complexity* of modern data sets is an even more serious and challenging aspect. Useful terminology for approaching a complex data set is *object oriented data analysis*, introduced in Wang and Marron (2007) and more recently discussed in Marron and Alonso (2014).

There has been a realization by many that asymptotics should include calculation of limits as the dimension grows. For example, in the biological field of gene expression, the technology of microarrays (see Murillo et al. (2008) for an overview) enabled the measurement of expressions of tens of thousands of genes at once, and one could say this leads to vectors effectively understood by a limiting process of growing dimension. Others might contend that such a limit is inappropriate. This line of discussion could be continued in the direction of RNAseq techonology (see Denoeud et al. (2008) for an introduction), where instead of a single number for each gene, expression estimates at the level of resolution of genetic base pairs are available, thus resulting in a factor of around ten thousand more dimensions. Here we argue that *the goal of asymptotics is to find insightful simple structures that underly complex statistical contexts*.

There are a number of ways that growing dimension asymptotics have been studied. Pioneering work by Portnoy (1984) and Portnoy et al. (1988) studied cases where the dimension grew relatively slowly, resulting in an asymptotic domain that was not far from classical fixed-dimensional analysis. Another asymptotic domain, random matrix theory, arises when the dimension and sample size grow at the same rate. Work in this area has been done mostly outside the statistical community, with landmark results including Marcenko and Pastur (1967) on the distribution of eigenvalues of the sample covariance matrix, and Tracy and Widom (1996) on the distribution of the largest eigenvalue. A good overview of the literature in this area can be found in Bai and Silverstein (2009). Statistical implications have been developed in a series of papers by Johnstone and co-authors, see e.g., Johnstone and Lu (2009), and a number of others since.

An asymptotic domain whose importance has only recently begun to be recognized has the dimension growing more rapidly than the sample size. Hall, Marron, and Neeman (2005) coined the terminology *high dimension, low sample size* (HDLSS) for the case where the sample size is fixed while the dimension grows.

2. HDLSS Backround

As noted in Hall, Marron, and Neeman (2005), the world of HDLSS asymptotics is full of concepts and ideas that can run counter to intuition, many of which are discussed in the following. For example, in the limit as the dimension $d \to \infty$ with a fixed sample size n, a standard Gaussian sample will lie near the surface of a growing sphere of radius $d^{1/2}$ and the angle between each pair of points with vertex at the origin will approach 90°. Thus the increasing randomness inherent in growing dimension tends towards random rotation and, modulo that random rotation and scaling by a factor of $d^{-1/2}$, the data tend to lie near vertices of the unit simplex. This phenomenon, *geometric representation*, leads to a number of interesting statistical insights.

Beran (1996) and Beran et al. (2010) have it that some of these ideas lie at the heart of the famous paper of Stein (1956) on inadmissability of the sample mean. Early use of HDLSS asymptotics in the Stein estimation context can be found in Casella and Hwang (1982).

A closely related asymptotic domain is the *ultra high dimension* of Fan and Lv (2008), who consider the case of $d = exp(cn^a)$, for constants c and a, in the classical limit as $n \to \infty$. The ultra high case can also be equivalently formulated as $n = (\log(d)/c)^{1/a}$ in the limit as $d \to \infty$. The later formulation reveals that this case is quite near to the HDLSS context.

The HDLSS geometric representation has been established under a range of conditions. Hall, Marron, and Neeman (2005) went beyond the independent Gaussian case by assuming the data vectors satisfy a mixing condition. Ahn et al. (2007) proposed a more palatable eigenvalue condition for geometric representation. In unpublished correspondence John Kent pointed out, using a Gaussian scale mixture example, that more than univariate moment conditions are needed. Conditions that are especially appealing, because they are based only on the covariance matrix (with no assumption of Gaussianity), have been developed in a series of papers Yata and Aoshima (2009, 2010b); Aoshima and Yata (2011); Yata and Aoshima (2012a); Aoshima and Yata (2015). A non-Gaussian condition that makes intuitive sense based on the types of data found in genomics can be found in Jung and Marron (2009).

3. PCA and HDLSS Asymptotics

PCA is a well-proven workhorse method for many tasks involving high-dimensional data, see Jolliffe (2005) for excellent introduction and overview. A strong indicator of its utility comes from the fact that it has been rediscovered and renamed a number of times. For example, it is called *empirical orthogonal*

functions in the earth / climate sciences, *proper orthogonal decomposition* in applied mathematics, the *Karhunen Loeve expansion* in electrical engineering and probability, and *factor analysis* in a number of non-statistical areas.

One can motivate PCA from a Gaussian likelihood viewpoint, but it is more generally viewed as a fully nonparameteric method, with a number of uses. While PCA is a common dimension reduction method, it is arguably even more useful for data visualization in high-dimensional contexts, for example in Functional Data Analysis, see Ramsay and Silverman (2002), Ramsay (2005), and Ferraty and Vieu (2004).

An example of PCA data visualization is shown in Figure 1. The curves in the left panel are read depth curves from RNAseq measurements Wilhelm and Landry (2009) of $n = 180$ lung cancer patients. Each curve is a detailed measurement of biological expression of the gene CDKN2A for one tissue sample, with the horizontal axis indicating $d = 1{,}709$ base pair locations of the reference genome, and the height of the curve a $\log_{10}$ count that indicates level of gene expression. Insight into how the curves relate to each other comes from the PCA scatterplot in the right panel of Figure 1. The axes of the scatterplot are the first two principal component scores. There is some apparent interesting structure, in the form of three distinct clusters. To explore the relevance of these clusters, they have been *brushed*, i.e., colored with grey levels. These same colors have been used on the curves in the left panel. The lightest grey curves tend to be much lower than the others, representing cases where the CDKN2A gene is essentially not expressed (typical in some cases). The genome location of the gene includes several disconnected loci called exons, which here have been connected giving the block-like structure apparent in the other curves. The third exon shows very low levels of expression for all cases. The fifth exon is interesting because it is fully expressed for only the black cases, while all others show a very low level of expression. This phenomenon is called *alternate splicing*, and is very important in cancer research because it can become the target of drug treatments. This example motivated a search for alternate splice events, based on screening for clusters in read depth curves, over all genes. An important challenge to the implemention of this was the assessment of the statistical significance of clusters in very high dimensions, done using the SigClust method of Liu et al. (2008). This screening method was named SigFuge by Kimes and Cabanski (2013), who reported on the results of a full genome screen that found previously unknown splicing events, one of which was then confirmed by a biological experiment. The point is that PCA found important structure in this HDLSS data set through visualizing the PC scores.

The asymptotic underpinnings of PCA in HDLSS contexts were first considered by Ahn et al. (2007), and more deeply by Jung and Marron (2009). Insight

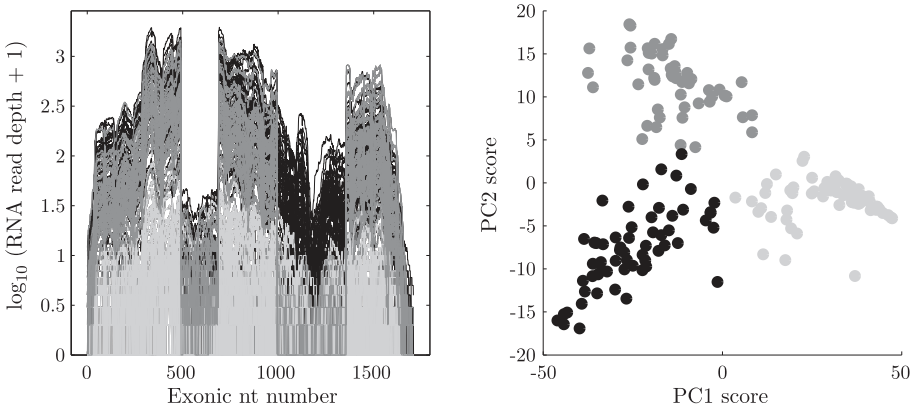

Figure 1. PCA of RNAseq log read depth maps, for the union of exons in the gene CDKN2A. Left panel shows data curves. Right panel is the PC1 vs. Pc2 scores scatterplot, showing three distinct clusters (brushed with gray levels). Use of these same colors in the left panel shows essentially unexpressed cases (lightest grey), and a clear alternate splicing event (other grey shades). This is an HDLSS example where PCA clearly reveals important biological structure.

into the behavior of PCA comes from study of the spike covariance model, made popular in statistics by Paul (2007) and Johnstone and Lu (2009). In the simplest form of the spike covariance model, there is a single large eigenvalue with the rest much smaller and constant, where the largest eigenvalue has size of growing order d^{α}. The key result is that, in the limit as $d \to \infty$ with n fixed, letting u_1 and $\hat{u}_1$ denote the first population and sample eigenvectors respectively,

$$\text{angle} < \hat{u}_1, u_1 > \to \begin{cases} 90^{\circ} \text{ for } & \alpha < 1, \\ 0 \text{ for } & \alpha > 1. \end{cases} \tag{3.1}$$

This defines a notion of *consistency* when the spike is large, $\alpha > 1$, and a notion of *strong inconsistency* when the spike is small, $\alpha < 1$. One or the other holds in almost all such settings, with the exception of $\alpha = 1$. The boundary at $\alpha = 1$ can be understood from the geometric representation: in the HDLSS limit, standard Gaussian data tends to lie on the surface of the sphere at the origin, with radius $d^{1/2}$. In the spike model, when $\alpha > 1$, the distribution reaches outside this sphere (eigenvalues are on the scale of variance, so the largest standard deviation is much greater than $d^{1/2}$), which results in consistency of the first eigendirection. When $\alpha < 1$, the distribution is essentially contained within the sphere, so the first eigendirection is random, and random directions are asymptotically orthogonal to any given direction.

Related results are available in Yata and Aoshima (2009, 2010a, 2012a, 2013). New insights about sparse PCA were dicovered by Shen, Shen, and Marron

(2012a); See Shen, Shen, and Marron (2012b) for broader results in this spirit, including all combinations of n and d tending to infinity. Limiting behavior when $\alpha = 1$ was first explored by Jung, Sen, and Marron (2012) in the HDLSS case. Yata and Aoshima (2012a) proposed a noise reduction estimator that relaxes the boundary of the eigenvalue estimator to $\alpha = 1/2$ in the HDLSS case.

While under a sufficiently strong signal in the covariance structure, PCA can find the right direction vectors, estimation of the eigenvalues is more challenging. Indeed, letting λ_1 and $\hat{\lambda}_1$ denote the population (and sample resp.) first eignevalues, there are a number of results in the spirit of

$$\frac{\hat{\lambda}_1}{\lambda_1} \xrightarrow{\text{L}} \frac{\chi^2_n}{n}, \tag{3.2}$$

under various HDLSS conditions. Thus sample eigenvalues are generally *inconsistent* when the sample size is fixed but, as noted by Yata and Aoshima (2009, 2010b), sample eigenvalues are consistent if it is assumed that, in addition to $d \to \infty$, $n \to \infty$ as well. Where n grows more slowly than d, one speaks of High Dimension Moderate Sample Size, Borysov, Hannig, and Marron (2014). Yata and Aoshima (2010a, 2012a, 2013) have given modified versions of PCA that provide consistent eigenvalue estimates.

4. Understanding Variation in Scores

In HDLSS situations PCA scores, which form the basis of informative scatterplots, such as Figure 1, are generally inconsistent in the HDLSS limit. In particular, the ratio of the sample and population PC scores converge to a non-degenerate random variable, as formulated here.

In Section 4.2, we show that, for a given component, the ratios for each data point indeed converge to a random variable, but it is the *same realization* of the random variable for each data point. Thus, while all the scores are off by a random factor, it is the *same* factor for each data point; in scatterplots the axis labels are off by a random factor, but the relationship between points is still correct.

This issue was presaged in Yata and Aoshima (2009) and is quite similar to the eignevalue inconsistency at (3.2). The improved variation of PCA proposed in Yata and Aoshima (2012a, 2013) gives asymptotically correct scalings. Under the random matrix framework, Lee, Zou, and Wright (2010) showed that the ratios of the sample and population PC scores converge to a constant. Hellton and Thoresen (2014) have used the ideas of *pervasive signal* and *visual content* to explore this inconsistent phenomenon.

4.1. Assumptions and notation

If $\{(\lambda_k, u_k) : k = 1, \cdots, d\}$ are the eigenvalue-eigenvector pairs of the covariance matrix Σ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$, we can write

$$\Sigma = U \Lambda U^T, \tag{4.1}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ and $U = [u_1, \ldots, u_d]$.

Assumption 1. *$X_1, \ldots, X_n$ are i.i.d. d-dimensional random sample vectors with the representation*

$$X_i = \sum_{j=1}^{d} \lambda_j^{1/2} z_{i,j} u_j, \tag{4.2}$$

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 has the X_is' $N(0, \Sigma)$. Consider that the X_i are i.i.d. $N(\xi, \Sigma)$ with $\xi \neq 0$. As in Paul and Johnstone (2007), it is well known that

$$\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T \quad \text{has the same distribution as} \quad \sum_{i=1}^{n-1} Y_i Y_i^T,$$

where $\overline{X}$ is the sample mean and Y_i are i.i.d. $N(0, \Sigma)$. Then the asymptotic properties of PCA can be studied through Y_i. Since the sample covariance matrix is location invariant, we can assume without loss of generality that X_i has zero mean at least for the normal case. In general, one has to consider the theoretical properties of $\overline{X} - \mu$; these have been widely investigated in the literature Rollin (2013); Chernozhukov, Chetverikov, and Kato (2014) even when d is much larger than n.

Denote the jth normalized population PC score vector by

$$S_j = (S_{1,j}, \ldots, S_{n,j})^T = \lambda_j^{-1/2} (u_j^T X_1, \ldots, u_j^T X_n)^T, \quad j = 1, \ldots, d. \tag{4.3}$$

Denote the data matrix by $X = [X_1, \ldots, X_n]$ and the sample covariance matrix by $\hat{\Sigma} = (1/n) X X^T$. The sample covariance matrix can be decomposed as

$$\hat{\Sigma} = \hat{U} \hat{\Lambda} \hat{U}^T, \tag{4.4}$$

where, similarly, $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$ and $\hat{U} = [\hat{u}_1, \ldots, \hat{u}_d]$. Since

$$n^{-1/2} X = \sum_{j=1}^{d} \hat{\lambda}_j^{1/2} \hat{u}_j \hat{v}_j^T, \quad \text{where} \quad \hat{v}_j = (\hat{v}_{1,j}, \ldots, \hat{v}_{n,j})^T, \quad j = 1, \cdots, d,$$

the jth normalized sample PC score vector is

$$\hat{S}_j = (\hat{S}_{1,j}, \ldots, \hat{S}_{n,j})^T = n^{1/2}(\hat{v}_{1,j}, \ldots, \hat{v}_{n,j})^T, \quad j = 1, \ldots, d. \tag{4.5}$$

Let $\{a_k : k = 1, \ldots, \infty\}$ and $\{b_k : k = 1, \ldots, \infty\}$ be two sequence of constants, where k can stand for either n or d. We write $a_k \gg b_k$ if $\lim_{k\to\infty} b_k/a_k = 0$, and $a_k \sim b_k$ if $c_2 \le \underline{\lim}_{k\to\infty} a_k/b_k \le \overline{\lim}_{k\to\infty} a_k/b_k \le c_1$ for constants $c_1 \ge c_2 > 0$. If $\{\xi_k : k = 1, \ldots, \infty\}$ is a sequence of random variables and $\{e_k : k = 1, \ldots, \infty\}$ is a sequence of constants, we write $\xi_k = \mathrm{O}_{\mathrm{a.s}}(e_k)$ if $\overline{\lim}_{k\to\infty} |\xi_k/e_k| \le \zeta$ almost surely with $P(0 < \zeta < \infty) = 1$.

4.2. HDLSS inconsistency

In this subsection, we show the asymptotic properties of PC scores in HDLSS when the sample size n is fixed and the dimension $d \to \infty$. We consider multiple spike models Jung and Marron (2009) under which, as $d \to \infty$,

$$\lambda_1 \gg \cdots \gg \lambda_m \gg \lambda_{m+1} \sim \cdots \sim \lambda_d \sim 1, \tag{4.6}$$

where $m \in [1, n \wedge d]$ is finite. Under these spike models, Jung and Marron (2009) showed that when n is fixed, if $d/\lambda_m \to 0$, the angle between each of the first m sample eigenvectors $\hat{u}_j$ and its corresponding population eigenvector u_j goes to 0 with probability 1, the *consistency* of the sample eigenvector.

However, under the same assumptions, the sample PC scores are not consistent. We show that, for a particular principal component, the proportion between the sample PC scores and the corresponding population scores converges to a random variable, the realization of which remains the same for all data points. Since we study $\hat{S}_{i,j}/S_{i,j}$, we need to assume

$$P(z_{i,j} \neq 0) = 1, \quad i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{4.7}$$

to ensure that $P(S_{i,j} \neq 0) = 1$ ($S_{i,j} = z_{i,j}$ from (4.2) and (4.3)). Let

$$\widetilde{Z}_j = (z_{1,j}, \ldots, z_{n,j})^T \quad \text{and} \quad R_j = \frac{\widetilde{Z}_j^T \widetilde{Z}_j}{n}, \quad j = 1, \ldots, d, \tag{4.8}$$

where $z_{i,j}$ is defined in (4.2).

Theorem 1. *Under Assumption* 1, (4.6), *and* (4.7), *for fixed* n *as* $d \to \infty$, *if* $d/\lambda_m \to 0$, *then*

$$\left| \frac{\hat{S}_{i,j}}{S_{i,j}} \right| \xrightarrow{a.s} R_j^{-1/2}, \quad i = 1, \ldots n,\ j = 1, \ldots, m, \tag{4.9}$$

where $\xrightarrow{a.s}$ *stands for almost sure convergence. In addition, if Assumption* 1 *holds with normal* $z_{i,j}$*'s, then* nR_j *is the Chi-square with* n *degrees of freedom.*

Remark 1. The results in Theorems 1 differ from those in Lee, Zou, and Wright (2010). Under the random matrix framework with $n \sim d \to \infty$ and $\lambda_j < \infty$, Lee, Zou, and Wright (2010) showed that the ratios between the sample and population eigenvalues converge to a constant. Lee, Zou, and Wright (2010) did not consider properties of PC scores under the framework of our Theorem 2.

Remark 2. As the ratio R_j only depends on j, scores scatter plots, such as right panel of Figure 1, have incorrectly labeled axes but asymptotically correct relative positions of points.

Remark 3. For non-normalized PC scores with $S^o_{i,j} = u_j^T X_i = \lambda_j^{1/2} S_{i,j}$ and $\hat{S}^o_{i,j} = \hat{\lambda}_j^{1/2} \hat{S}_{i,j}$, $\left| \hat{S}^o_{i,j} \,/\, S^o_{i,j} \right| \xrightarrow{a.s} 1$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$, under the assumptions of Theorem 1.

4.3. Growing sample size analysis

In this subsection, we consider growing sample size contexts, and study the asymptotic properties of the PC scores. Here both the sample eigenvectors and the sample principal component scores can be consistent.

Consider the spike models, as $n, d \to \infty$,

$$\lambda_1 > \cdots > \lambda_m \gg \lambda_{m+1} \sim \cdots \sim \lambda_d \sim 1. \tag{4.10}$$

Theorem 2. *Under Assumption* 1, (4.7), *and* (4.10), *for* $n, d \to \infty$, *if* $d/(n^{1/2}\lambda_m) \to 0$, *then*

$$\left| \frac{\hat{S}_{i,j}}{S_{i,j}} \right| \xrightarrow{a.s} 1, \quad i = 1, \ldots n, \; j = 1, \ldots, m. \tag{4.11}$$

Remark 4. In the current context, the consistency of the sample PC scores fits, as expected, with the fact that the sample eigenvectors are consistent under the assumptions of Theorem 2. In particular, Shen, Shen, and Marron (2012b) shows that, under the same assumptions, the angle between the sample eigenvector $\hat{u}_j$ and the corresponding population eigenvector u_j, $j = 1, \ldots, m$, converges almost surely to 0.

5. Deeper Conical Behavior

This section reveals an asymptotic conical structure in critical sample eigendirections under the spike models when the sample size and/or the number of variables (or dimension) tend to infinity. The consistency of the sample eigenvectors relative to their population counterparts is determined by the ratio between the

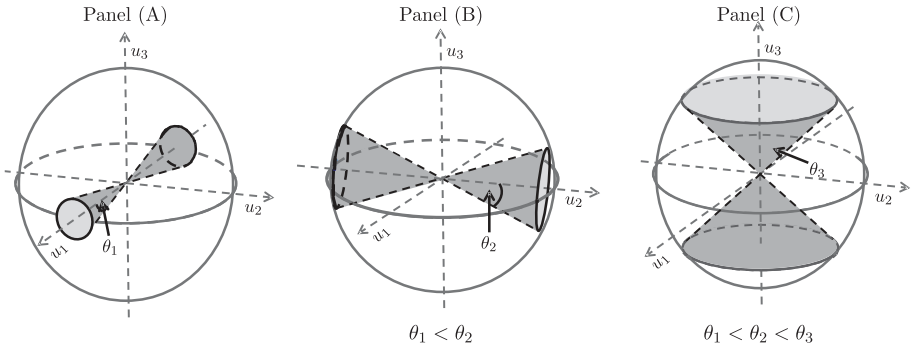

Figure 2. Geometric representation of PC directions in Example 1. The sphere represents the space of possible sample eigenvectors. Panel (A) shows that the first sample eigenvector tends to lie in the dark gray cone, with the θ_1 angle. Similarly, Panels (B) and (C) show that the second and the third sample eigenvectors respectively tend to lie in the dark gray cones, whose angles are θ_2 and θ_3. Note that θ_1 is less than θ_2, which is again less than θ_3.

dimension and the product of the sample size with the spike size. When this ratio converges to a nonzero constant, the sample eigenvector converges to a cone, with a certain angle to its corresponding population eigenvector. In the HDLSS case, the angle between the sample eigenvector and its population counterpart converges to a limiting distribution. Several generalizations of the multi-spike covariance models are also explored, and additional theoretical results are presented.

We first introduce two examples to help understand the asymptotic results of conical structure for sample eigendirections.

Example 1 (Multiple-component spike models with distinguishable eigenvalues)**.** Let $X_1, \ldots, X_n$ be random sample vectors according to (4.2), where the population eigenvalues satisfy, as $n, d \to \infty$,

$$\begin{cases} \lambda_1 > \lambda_2 > \lambda_3 \gg \lambda_4 = \cdots = \lambda_d = 1, \\ \dfrac{d}{n\lambda_j} \to c_j, \quad j = 1, 2, 3, \text{ with } 0 \le c_1 < c_2 < c_3 \le \infty. \end{cases}$$

In Figure 2, the sphere represents the space of all possible sample eigendirections, with the first three population eigenvectors as the coordinate axes. From Theorem 3, as $n, d \to \infty$, the sample eigenvector $\hat{u}_1$ lies in the dark gray cone, shown in Panel (A) of Figure 2, with the angle of the cone $\theta_1 = \arccos(1/\sqrt{1+c_1})$. Similarly, as $n, d \to \infty$, the sample eigenvectors $\hat{u}_2$ and $\hat{u}_3$,

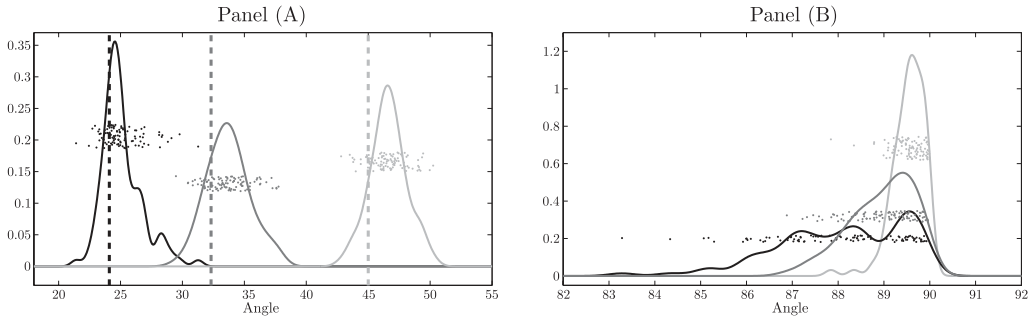

Figure 3. Example 1: Simulated angles between sample and population eigenvectors. Panel (A) shows realizations of angles between sample and population eigenvectors as black gray dots (black is first, dark gray is second, light gray is third). Distributions are studied using kernel density estimates, and compared with the theoretical values θ_j for $j = 1, 2, 3$, shown as dashed lines. Panel (B) studies randomness of eigen-directions within the cones shown in Figure 2, by showing the distribution of pairwise angles between realizations of the sample eigenvectors. All 3 colors are overlaid here, and all angles are close to 90 degrees, which is consistent with the randomness of the respective sample eigenvectors within the cones.

respectively, lie in the dark gray cones, shown in Panels (B) and (C) of Figure 2, with angles $\theta_2 = \arccos(1/\sqrt{1+c_2})$ and $\theta_3 = \arccos(1/\sqrt{1+c_3})$. For $c_1 < c_2 < c_3$, we have $\theta_1 < \theta_2 < \theta_3$, as shown in Figure 2.

Our Proposition S4.1 in the supplementary material includes the two boundary cases studied by Shen, Shen, and Marron (2012b) as special cases. When $c_1 = c_2 = c_3 = 0$, it follows that $\theta_1 = \theta_2 = \theta_3 = 0$, in the domain of consistency (Shen, Shen, and Marron (2012b)). When $c_1 = c_2 = c_3 = \infty$, we have $\theta_1 = \theta_2 = \theta_3 = 90$ degrees and strong inconsistency Shen, Shen, and Marron (2012b).

We investigated convergence, using simulations, over a range of settings, with $n = 50$, 100, 200, 500, 1,000, 2,000, where $d/n = 50$, and $c_1 = 0.2$, $c_2 = 0.4$, $c_3 = 1$. The full sequence, illustrating this convergence, is shown in Figure 3 A of the supplementary material. Figure 3 shows the intermediate case of $n = 200$. For one data set with this distribution, we computed angles between the sample and population eigenvectors. Repeating this procedure over 100 replications, we got 100 angles for each of the first three eigenvectors, shown as black, dark gray, and light gray points in Panel (A). The black, dark gray, light gray curves are the corresponding kernel density estimates. Panel (A) shows that the simulated angles close to the corresponding theoretical angles θ_j, $j = 1, 2, 3$, shown as dashed vertical lines.

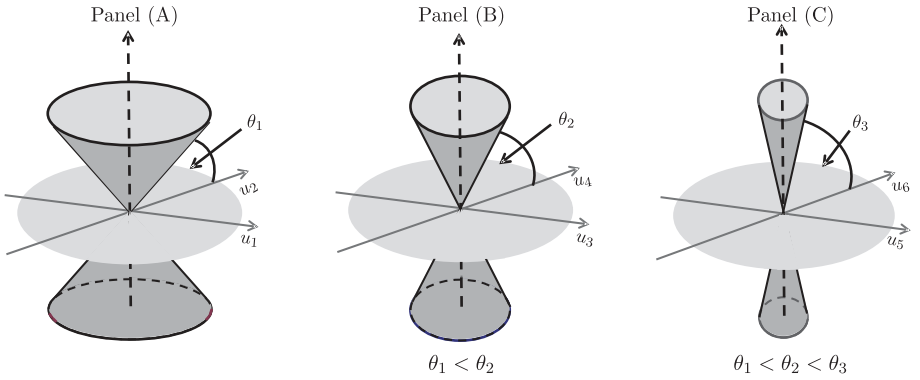

Figure 4. Example 2: Geometric representation of PC directions. Panel (A) shows the cone to which the first group of sample eigenvectors converge in the dark gray. This cone has angle θ_1 with the light gray subspace, generated by the first group of population eigenvectors. Similarly, Panel (B) (Panel (C)) shows the cone to which the second (third) group of sample eigenvectors converges shown as a dark gray cone, which has angle θ_2 (θ_3) with the subspace, generated by the second (third) group of population eigenvectors.

Panel (B) in Figure 3 studies randomness of eigen-directions within the cones shown in Figure 2. We calculated pairwise angles between realizations of the sample eigenvectors for the three cones, showing angles and kernel density estimates using colors as in Panel (A) of Figure 3. All angles ware close to 90 degrees, consistent with randomness in high dimensions, see Hall, Marron, and Neeman (2005); Yata and Aoshima (2012a); Jung and Marron (2009); Jung, Sen, and Marron (2012); Cai, Fan, and Jiang (2013).

Example 2. (Multiple-component spike models with indistinguishable eigenvalues) Here we take the six leading population eigenvalues to satisfy, as $n, d \to \infty$

$$\begin{cases} \lambda_1 = \lambda_2 > \lambda_3 = \lambda_4 > \lambda_5 = \lambda_6 \gg \lambda_7 = \cdots = \lambda_d = 1, \\ \dfrac{d}{n\lambda_{2j-1}} \to c_j, \quad j = 1, 2, 3, \text{ with } 0 \le c_1 < c_2 < c_3 \le \infty. \end{cases}$$

From Theorem 4, Panel (A) in Figure 4 shows, as a dark gray cone, the region where the first group of sample eigenvectors $\hat{u}_1$ and $\hat{u}_2$ lie in the limit as $n, d \to \infty$. This has the angle $\theta_1 = \arccos(1/\sqrt{1+c_1})$ with the light gray subspace generated by the first group of population eigenvectors, u_1 and u_2. Similarly, Panel (B) (Panel (C)) presents, as a dark gray cone, the region where the second (third) group of sample eigenvectors $\hat{u}_3$ and $\hat{u}_4$ ($\hat{u}_5$ and $\hat{u}_6$) lie in the limit as $n, d \to \infty$. This has the angle $\theta_2 = \arccos(1/\sqrt{1+c_2})$ ($\theta_3 = \arccos(1/\sqrt{1+c_3})$) with the subspace generated by the second (third) group of population eigenvectors, u_3 and u_4 (u_5 and u_6). For $c_1 < c_2 < c_3$, we have $\theta_1 < \theta_2 < \theta_3$, as shown in Figure 4.

Our Proposition S2.1 in the supplementary material includes boundary cases studied by Shen, Shen, and Marron (2012b) as special cases. For $c_1 = c_2 = c_3 = 0$, it follows that $\theta_1 = \theta_2 = \theta_3 = 0$, in the domain of subspace consistency; see Theorem 4.3 of Shen, Shen, and Marron (2012b). When $c_1 = c_2 = c_3 = \infty$, we have $\theta_1 = \theta_2 = \theta_3 = 90$ degrees and strong inconsistency; see Theorem 4.3 of Shen, Shen, and Marron (2012b).

5.1. Growing sample size asymptotics

We now study asymptotic properties of PCA as $n \to \infty$. We consider multiple component spike models with distinguishable population eigenvalues in Section 5.1.1, and with indistinguishable eigenvalues in Section 5.1.2. We vary d from $d \ll n$, through the random matrix version with $d \sim n$, to the high dimension medium sample size (HDMSS) asymptotics of Cabanski et al. (2010), Yata and Aoshima (2012b), and Aoshima and Yata (2015), with $d \gg n \to \infty$. Aoshima and Yata (2015) improves the results of Yata and Aoshima (2012b) under mild conditions.

5.1.1. Multiple component spike models with distinguishable eigenvalues

We consider multiple component spike models where the population eigenvalues satisfy the following.

$\mathcal{A}1$. As $n, d \to \infty$, $\lambda_1 > \cdots > \lambda_m \gg \lambda_{m+1} \to \cdots \to \lambda_d = 1$.

$\mathcal{A}2$. As $n, d \to \infty$, $\quad \frac{d}{n\lambda_j} \to c_j, \quad$ where $\quad 0 < c_1 < \cdots < c_m < \infty$.

Here, that $\lambda_1 > \cdots > \lambda_m$ makes it possible to separately consider the first m principle component signals and their corresponding asymptotic properties. That $\lambda_m \gg \lambda_{m+1} \to \cdots \to \lambda_d = 1$ enables clear separation of the signal in the first m components from the noise in the higher order components.

In $\mathcal{A}2$, the *positive* and *negative* information are of the same order: increasing n and the spike positively impacts the consistency of PCA, whereas increasing d has a negative impact.

In our context, we take $H = \{m + 1, \ldots, d\}$ as the noise index set and the space spanned by the noise eigenvectors as

$$\mathbb{S} = \text{span}\{u_j, j \in H\}. \tag{5.1}$$

For each sample eigenvector $\hat{u}_j$, $j \in H$, we study the angle between $\hat{u}_j$ and the space $\mathbb{S}$, as defined in Jung and Marron (2009) and Shen, Shen, and Marron (2012b), and illustrated in Figure B of the supplementary material.

Theorem 3. *Under Assumptions* 1, $\mathcal{A}1$, *and* $\mathcal{A}2$, *as* $n, d \to \infty$, *the sample eigenvalues satisfy*

$$\begin{cases} \dfrac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} 1 + c_j, \ 1 \le j \le m, \\ \dfrac{n\hat{\lambda}_j}{d\lambda_j} \xrightarrow{\text{a.s}} 1, \ m+1 \le j \le [n \wedge d], \end{cases} \tag{5.2}$$

and the sample eigenvectors satisfy

$$\begin{cases} |< \hat{u}_j, u_j >| \xrightarrow{a.s} (1 + c_j)^{-1/2}, \quad 1 \le j \le m, \\ |< \hat{u}_j, u_j >| = \mathrm{O}_{\text{a.s}} \left\{ (\frac{n}{d})^{1/2} \right\}, \quad m+1 \le j \le [n \wedge d], \\ angle < \hat{u}_j, \mathbb{S} > \xrightarrow{a.s} 0, \quad m+1 \le j \le [n \wedge d]. \end{cases} \tag{5.3}$$

Remark 5. The $d/n\lambda_j \to c_j$ contains three scenarios: n, d, and $\lambda_j \to \infty$; $d, \lambda_j \to \infty$ and $n < \infty$ (HDLSS); $n, d \to \infty$ and $\lambda_j < \infty$. We study the first two. Theorem 3 studies the first scenario and Paul (2007) studied the third. The results in Paul (2007) are based on the normal assumption, unnecessary here, and do not pertain to indistinguishable eigenvalues.

Remark 6. The results of (5.2) and (5.3) suggest that, as the eigenvalue index increases, the proportional bias between the sample and population eigenvalue increases, so the angle between the sample and corresponding population eigenvectors increases. This is because larger eigenvalues (i.e. with small indices) contain more positive information, which makes the corresponding sample eigenvalues/eigenvectors less biased. These results are graphically illustrated in Figure 1 and empirically verified in Figure 2, for the specific model in Example 1. More empirical support is provided in the supplementary material.

Remark 7. Theorem 3 can be extended to include the random matrix and HDMSS cases; This is shown in Section S4.1 of the supplementary material.

5.1.2. Multiple component spike models with indistinguishable eigenvalues

We consider spike models with the m leading eigenvalues grouped into $r(\ge 1)$ *tiers*, each of which contains all given eigenvalues that are either the same or have the same limit. Specifically, the first m eigenvalues are grouped into r tiers, in which there are q_k eigenvalues in the kth tier such that $\sum_{l=1}^{r} q_l = m$. Let $q_0 = 0$, $q_{r+1} = d - \sum_{l=1}^{r} q_l$, and the index set of the eigenvalues in the kth tier be

$$H_k = \left\{ \sum_{l=0}^{k-1} q_l + 1, \sum_{l=0}^{k-1} q_l + 2, \ldots, \sum_{l=0}^{k-1} q_l + q_k \right\}, \quad k = 1, \ldots, r+1. \tag{5.4}$$

We make these formal assumptions.

$\mathcal{B}1$. The eigenvalues in the kth tier have the limit $\delta_k(>0)$:

$$\lim_{n,d\to\infty}\frac{\lambda_j}{\delta_k}=1,\quad j\in H_k,\ k=1,\dots,r.$$

$\mathcal{B}2$. The eigenvalues in different tiers have different limits:

$$\text{as } n,d\to\infty,\quad \delta_1>\cdots>\delta_r\gg\lambda_{m+1}\to\cdots\to\lambda_d=1.$$

$\mathcal{B}3$. The ratio between the dimension and the product of the sample size with eigenvalues in the same tier converges to a constant:

$$\text{as } n,d\to\infty,\quad \frac{d}{n\delta_k}\to c_k,\quad \text{with}\quad 0<c_1<\cdots<c_r<\infty.$$

Since the sample eigenvalues within the same tier can not be asymptotically identified, the corresponding sample eigenvectors are indistinguishable. For $j\in H_k$, in order to study the asymptotic properties of the sample eigenvector $\hat{u}_j$, we consider the angle between $\hat{u}_j$ and the subspace spanned by the population eigenvectors u_j in the same tier,

$$\mathbb{S}_k=\text{span}\{u_j,j\in H_k\}. \tag{5.5}$$

Theorem 4. *Under Assumptions* 1, $\mathcal{B}1$, $\mathcal{B}2$, *and* $\mathcal{B}3$, *as* $n,d\to\infty$, *the sample eigenvalues satisfy*

$$\begin{cases}\dfrac{\hat{\lambda}_j}{\lambda_j}\xrightarrow{\text{a.s}}1+c_k,\quad j\in H_k,\ k=1,\dots,r,\\[2ex] \dfrac{n\hat{\lambda}_j}{d\lambda_j}\xrightarrow{\text{a.s}}1,\quad m+1\le j\le[n\wedge d],\end{cases} \tag{5.6}$$

and the sample eigenvectors satisfy

$$\begin{cases}\text{angle}<\hat{u}_j,\mathbb{S}_k>\xrightarrow{a.s}\arccos\left\{(1+c_k)^{-1/2}\right\},\ j\in H_k,\ k=1,\dots,r,\\ |<\hat{u}_j,u_j>|=\mathrm{O}_{\text{a.s}}\left\{(\frac{n}{d})^{1/2}\right\},\quad m+1\le j\le[n\wedge d],\\ \text{angle}<\hat{u}_j,\mathbb{S}_{r+1}>\xrightarrow{a.s}1,\quad m+1\le j\le[n\wedge d].\end{cases} \tag{5.7}$$

Theorem 4 extends Theorem 3. For higher-order eigenvalues, the sample eigenvalues are more biased, while the angles between the sample eigenvectors and the subspaces spanned by their population counterparts in the same tiers are larger. See Figure 3 for an illustration of the specific model considered in Example 2. Theorem 4 can be extended to cover the random matrix and HDMSS cases, see Section S4.2 of the supplementary material.

5.2. HDLSS asymptotics

We study the asymptotic properties of PCA in the HDLSS context. Here, the ratios between the sample eigenvalues and their population counterparts converge to non-degenerate random variables, as do the angles between the sample eigenvectors and the space spanned by the corresponding population eigenvectors.

Since the sample size is fixed, we can't distinguish the two types of spike models considered in Sections 5.1.1 and 5.1.2. Hence, we merge the model assumptions there as follows.

$\mathcal{C}1$. For fixed n, as $d \to \infty$, $\lambda_1 \geq \cdots \geq \lambda_m \gg \lambda_{m+1} \to \cdots \to \lambda_d = 1$.

$\mathcal{C}2$. For fixed n, as $d \to \infty$,

$$\frac{d}{n\lambda_j} \to c_j, \qquad \text{with} \quad 0 < c_1 \leq \cdots \leq c_m < \infty.$$

Now the sample eigenvalues and eigenvectors converge to non-degenerate random variables rather than constants. Consider the $m \times d$ matrix $\mathbb{M} = [\mathbb{C}, 0_{m\times(d-m)}]_{m\times d}$, where $\mathbb{C} = \operatorname{diag}\{c_1^{-1/2}, \ldots, c_m^{-1/2}\}$ is an $m \times m$ diagonal matrix and $0_{m\times(d-m)}$ is the $m \times (d-m)$ zero matrix. Take

$$Z = (z_{i,j})_{n\times d} \quad \text{and} \quad \mathcal{W} = \mathbb{M}Z^T Z\mathbb{M}^T, \tag{5.8}$$

where $z_{i,j}$ is defined in (4.2).

Given a fixed sample size, the sample eigenvalues can't be asymptotically distinguished, nor can the corresponding sample eigenvectors. To study the asymptotic behavior of the sample eigenvectors, we need to consider the space $\mathbb{S}_k$ spanned by the corresponding population eigenvectors, as defined in (5.5), with the index sets $H_1 = \{1, \ldots, m\}$ and $H_2 = \{m+1, \ldots, d\}$.

Theorem 5. *Under Assumptions* 1, $\mathcal{C}1$ *and* $\mathcal{C}2$, *for fixed* n, *as* $d \to \infty$, *the sample eigenvalues satisfy*

$$\begin{cases} \dfrac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s}} \frac{c_j}{n}\lambda_j(\mathcal{W}) + c_j, & 1 \leq j \leq m, \\ \dfrac{n\hat{\lambda}_j}{d\lambda_j} \xrightarrow{\text{a.s}} 1, & m+1 \leq j \leq n, \end{cases} \tag{5.9}$$

where $\mathcal{W}$ *is defined in* (5.8), *and the sample eigenvectors satisfy*

$$\begin{cases} \text{angle} < \hat{u}_j, \mathbb{S}_1 > \xrightarrow{a.s} \arccos\left\{\left(1 + \frac{n}{\lambda_j(\mathcal{W})}\right)^{-1/2}\right\}, & 1 \leq j \leq m, \\ |< \hat{u}_j, u_j >| = \mathrm{O}_{\text{a.s}}(d^{-1/2}), & m+1 \leq j \leq n, \\ \text{angle} < \hat{u}_j, \mathbb{S}_2 > \xrightarrow{a.s} 1, & m+1 \leq j \leq n. \end{cases} \tag{5.10}$$

Remark 8. Under the normal distribution, then Theorem 5 reduces to studies in Jung, Sen, and Marron (2012). Theorem 5 shows that the results in Jung, Sen, and Marron (2012) can be strengthened to almost sure convergence.

Remark 9. For $1 \leq j \leq m$, as the relative size of the eigenvalue decreases, the angle between $\hat{u}_j$ and $\mathbb{S}_1$ increases. However, this phenomenon is not as strong as in the growing sample size settings studied in Section 5.1, where the sample eigenvectors can be separately studied, and the corresponding angles have a non-random increasing order.

Remark 10. Assumption $\mathcal{C}2$ can be relaxed to include boundary cases, in which there exists an integer $m_0 \in [1, m]$ such that $c_{m_0} = 0$, positive information dominates in the leading m_0 spikes, or $c_{m_0+1} = \infty$, negative information dominates in the remaining high-order spikes. These results are presented in Section S5 of the supplementary material.

6. Proofs

We only present a detailed proof for Theorems 4. Theorem 3 is a special case of Theorem 4. The proof of properties of sample eigenvectors is in Section 6.1, while the properties of sample eigenvalues' properties are shown in Section S6.1 of the supplementary material.

The proofs of Theorems 1, 2, and 5 are in Sections S6 and S7 of the supplementary material, which also contains proofs of extensions of Theorems 3, 4, and 5.

6.1. The proof of sample eigenvectors' properties

This subsection gives the proof for the properties of the sample eigenvectors in Theorem 4.

The population eigenvalues are grouped into $r + 1$ tiers and H_k at (5.4) is the index set of the eigenvalues in the kth tier. Let

$$\hat{u}_j = (\hat{u}_{1,j}, \ldots, \hat{u}_{d,j})^T,\ j = 1, \ldots, d \text{ and } \hat{U}_{k,l} = (\hat{u}_{i,j})_{i \in H_k, j \in H_l},\ 1 \leq k, l \leq r + 1.$$

Then, the sample eigenvector matrix $\hat{U}$ can be expressed as

$$\hat{U} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_d] = \begin{pmatrix} \hat{U}_{1,1} & \hat{U}_{1,2} & \cdots & \hat{U}_{1,r+1} \\ \hat{U}_{2,1} & \hat{U}_{2,2} & \cdots & \hat{U}_{2,r+1} \\ \vdots & \vdots & & \vdots \\ \hat{U}_{r+1,1} & \hat{U}_{r+1,2} & \cdots & \hat{U}_{r+1,r+1} \end{pmatrix}. \tag{6.1}$$

Since $u_j = e_j$, $j = 1, \ldots, d$, the inner product between the sample and population eigenvectors satisfies

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

and the angle between the sample eigenvector and the corresponding population subspace $\mathbb{S}_k$ in (5.5) satisfies

$$\{\cos[\text{angle}(\hat{u}_j, \mathbb{S}_k)]\}^2 = \sum_{l \in H_k} \hat{u}_{l,j}^2, \quad k = 1, \ldots, r+1. \tag{6.2}$$

We first state Bai-Yin's law Bai and Yin (1993).

Lemma 1. *Suppose $B = (1/s)Z_{s\times m}^T Z_{s\times m}$, where $Z_{s\times m}$ is an $s \times m$ random matrix whose elements are i.i.d. and have zero mean, unit variance, and finite fourth moment. As $s \to \infty$ and $m/s \to c \in [0, \infty)$, the largest and smallest non-zero eigenvalues of B converge almost surely to $(1+\sqrt{c})^2$ and $(1-\sqrt{c})^2$, respectively.*

6.1.1. Asymptotic properties of the sample eigenvectors $\hat{u}_j$ with $j > m$

We derive the asymptotic properties as follows. First, we show that as $n, d \to \infty$, the angle between $\hat{u}_j$ and u_j converges to 90 degrees:

$$|< \hat{u}_j, u_j >|^2 = \hat{u}_{j,j}^2 = \mathrm{O}_{\mathrm{a.s}}\left(\frac{n}{d}\right), \quad j = m+1, \ldots, [n \wedge d]. \tag{6.3}$$

We then show that, as $n, d \to \infty$, the angle between $\hat{u}_j$ and the corresponding subspace $\mathbb{S}_{r+1}$ converges to 0, where $\mathbb{S}_{r+1}$ is defined as in (5.5):

$$\text{angle} < \hat{u}_j, \mathbb{S}_{r+1} > \xrightarrow{a.s} 0, \quad j = m+1, \ldots, [n \wedge d]. \tag{6.4}$$

To account for the first step, let $W = \Lambda^{-1/2} U^T \hat{U} \hat{\Lambda}^{1/2}$, where U and V are defined in (4.1) and $\hat{U}$ and $\hat{\Lambda}$ are defined in (4.4). It follows from (4.1), (4.2), and (4.4) that $WW^T = (1/n)Z^T Z$, where Z is defined in (5.8). Considering the kth diagonal entry of the equivalent matrices WW^T and $(1/n)Z^T Z$, and noting that $w_{k,j} = \lambda_k^{-1/2} \hat{\lambda}_j^{1/2} \hat{u}_{k,j}$ $(U = I_d)$, it follows that

$$\lambda_k^{-1} \sum_{j=1}^{d} \hat{\lambda}_j \hat{u}_{k,j}^2 = \sum_{j=1}^{d} w_{k,j}^2 = \frac{1}{n} \sum_{i=1}^{n} z_{i,k}^2. \quad k = 1, \ldots, d. \tag{6.5}$$

According to (6.5), we have

$$\hat{u}_{j,j}^2 \le \frac{\lambda_{m+1}}{\hat{\lambda}_{[n\wedge d]}} \left(\frac{1}{n} \sum_{i=1}^{n} z_{i,j}^2 \right), \quad j = m+1, \ldots, [n \wedge d]. \tag{6.6}$$

Select the $m+1$th to nth columns of Z in (5.8) to form the $n \times [n \wedge d]$ random matrix $\bar{Z}$. Note that $\sum_{i=1}^{n} z_{i,j}^2$, $j = m+1, \ldots, [n \wedge d]$, are the diagonal elements of $\bar{Z}^T \bar{Z}$ and less than or equal to the largest eigenvalue of $\bar{Z}^T \bar{Z}$. Then it follows from (6.6) that

$$\max_{m+1\le j\le[n\wedge d]}\hat{u}_{j,j}^2\le\frac{\lambda_{m+1}}{\hat{\lambda}_{[n\wedge d]}}\lambda_{\max}(\frac{1}{n}\bar{Z}^T\bar{Z}) \tag{6.7}$$

which, together with the asymptotic properties of the sample eigenvalues (5.6) and Lemma 1, yields (6.3).

For the second step, according to (6.2) we need to show that

$$\sum_{k=m+1}^{d}\hat{u}_{k,j}^2\xrightarrow{a.s}1,\quad j=m+1,\dots,[n\wedge d]. \tag{6.8}$$

The non-zero kth diagonal entry of W^TW is between its smallest and largest eigenvalues. Since W^TW shares the same non-zero eigenvalues as $(1/n)ZZ^T$, it follows that, for $j=1,\dots,[n\wedge d]$,

$$\lambda_{\min}(\frac{1}{n}ZZ^T)\le\hat{\lambda}_j\sum_{k=1}^{d}\lambda_k^{-1}\hat{u}_{k,j}^2=\sum_{k=1}^{d}w_{k,j}^2\le\lambda_{\max}(\frac{1}{n}ZZ^T), \tag{6.9}$$

which yields that, for $j=m+1,\dots,[n\wedge d]$,

$$\frac{\lambda_j}{\hat{\lambda}_j}\lambda_{\min}(\frac{1}{n}ZZ^T)\le\sum_{k=1}^{d}\lambda_j\lambda_k^{-1}\hat{u}_{k,j}^2\le\frac{\lambda_j}{\hat{\lambda}_j}\lambda_{\max}(\frac{1}{n}ZZ^T). \tag{6.10}$$

According to Lemma 1 and the asymptotic properties of the sample eigenvalues (5.6), we have that, for $j=m+1,\dots,[n\wedge d]$,

$$\frac{\lambda_j}{\hat{\lambda}_j}\lambda_{\min}\left(\frac{1}{n}ZZ^T\right)\quad\text{and}\quad\frac{\lambda_j}{\hat{\lambda}_j}\lambda_{\max}\left(\frac{1}{n}ZZ^T\right)\xrightarrow{a.s}1. \tag{6.11}$$

In addition, it follows from Assumption $\mathcal{B}2$ that, for $j=m+1,\dots,[n\wedge d]$,

$$\begin{cases}\lambda_j\lambda_k^{-1}\to0, & k=1,\dots,m,\\ \lambda_j\lambda_k^{-1}\to1, & k=m+1,\dots d.\end{cases} \tag{6.12}$$

Combining (6.10), (6.11), and (6.12), we have (6.8), which further leads to (6.4).

6.1.2. Asymptotic properties of the sample eigenvectors $\hat{u}_j$ with $j\in[1,m]$

We need to prove that, for $j=1,\dots,m$, the angle between the sample eigenvector $\hat{u}_j$ and the corresponding population subspace $\mathbb{S}_l$, $j\in H_l$, converges to $\arccos(1/\sqrt{1+c_l})$, $l=1,\dots,r$. According to (6.2), we only need to show that

$$\sum_{k\in H_l}\hat{u}_{k,j}^2\xrightarrow{a.s}\frac{1}{1+c_l},\quad j\in H_l,\ l=1,\dots,r. \tag{6.13}$$

We provide the detailed proof of (6.13) for $l=1$, and briefly illustrate how repeating the same procedure can lead to (6.13) for $l>2$.

In order to show (6.13) for $l=1$, we need a lemma about the asymptotic properties of the eigenvector matrix $\hat{U}$ in (6.1):

Lemma 2. *Under Assumptions in Theorem 4 and as $n, d \to \infty$, the rows of the eigenvector matrix $\hat{U}$ satisfy*

$$\sum_{l=1}^{r}(1+c_l)c_h c_l^{-1}\sum_{j\in H_l}\hat{u}_{k,j}^2 \xrightarrow{a.s} 1, \quad k \in H_h,\ h=1,\ldots,r, \tag{6.14}$$

and the columns of the eigenvector matrix $\hat{U}$ satisfy

$$\sum_{h=1}^{r}\sum_{k\in H_h}\hat{u}_{k,j}^2 \xrightarrow{a.s} \frac{1}{1+c_l}, \quad j \in H_l,\ l=1,\ldots,r. \tag{6.15}$$

In addition, we have

$$\sum_{l=1}^{r}(1+c_l)\sum_{j\in H_l}\hat{u}_{k,j}^2 \xrightarrow{a.s} 1, \quad k \in H_1. \tag{6.16}$$

Lemma 2 is proved in Section S6.4.3 of the supplementary material. We now show how to use Lemma 2 to prove (6.13) for $l=1$. If $h=1$ in (6.14), we have that

$$\sum_{l=1}^{r}(1+c_l)c_1 c_l^{-1}\sum_{j\in H_l}\hat{u}_{k,j}^2 \xrightarrow{a.s} 1, \quad k \in H_1. \tag{6.17}$$

Note that $c_1 c_l^{-1} < 1$ for $l > 1$, and comparing (6.16) with (6.17), we get that

$$\sum_{l=2}^{r}\sum_{j\in H_l}\hat{u}_{k,j}^2 \xrightarrow{a.s} 0, \quad \sum_{j\in H_1}\hat{u}_{k,j}^2 \xrightarrow{a.s} \frac{1}{1+c_1}, \quad k \in H_1, \tag{6.18}$$

which then yields that

$$\sum_{k\in H_1}\sum_{j\in H_1}\hat{u}_{k,j}^2 \xrightarrow{a.s} \frac{q_1}{1+c_1}, \tag{6.19}$$

where q_1 is the number of eigenvalues in H_1 (5.4). Summing over $j \in H_1$ in (6.15), we have that

$$\sum_{h=1}^{r}\sum_{k\in H_h}\sum_{j\in H_1}\hat{u}_{k,j}^2 \xrightarrow{a.s} \frac{q_1}{1+c_1}. \tag{6.20}$$

It follows from (6.19) and (6.20) that

$$\sum_{h=2}^{r}\sum_{k\in H_h}\sum_{j\in H_1}\hat{u}_{k,j}^2 \xrightarrow{a.s} 0, \tag{6.21}$$

which, together with (6.15) for $l=1$, yields

$$\sum_{k\in H_1}\hat{u}_{k,j}^2 \xrightarrow{a.s} \frac{1}{1+c_1}, \quad j \in H_1.$$

This is (6.13) for $l=1$.

We now prove (6.13) for $l = 2, \ldots, r$. Note that it follows from (6.21) that (6.14) becomes

$$\sum_{l=2}^{r}(1+c_l)c_h c_l^{-1}\sum_{j\in H_l}\hat{u}_{k,j}^2 \xrightarrow{a.s} 1, \quad k \in H_h,\ h = 2, \ldots, r. \tag{6.22}$$

It follows from (6.18) that (6.15) becomes

$$\sum_{h=2}^{r}\sum_{k\in H_h}\hat{u}_{k,j}^2 \xrightarrow{a.s} \frac{1}{1+c_l}, \quad j \in H_l,\ l = 2, \ldots, r. \tag{6.23}$$

Similar to (6.16), we have

$$\sum_{l=2}^{r}(1+c_l)\sum_{j\in H_l}\hat{u}_{k,j}^2 \xrightarrow{a.s} 1, \quad k \in H_2. \tag{6.24}$$

Combining (6.22), (6.23), and (6.24), we can prove (6.13) for $l = 2$. We can repeat the same procedure for $l = 3, \ldots, r$.

Supplementary Materials

Additional results, simulations, and proofs can be found in the online supplementary materials.

Acknowledgements

This material was based upon work partially supported by the NSF grant DMS-1127914 to the Statistical and Applied Mathematical Science Institute. This work was partially supported by NIH grants MH086633, 1UL1TR001111, and MH092335, and NSF grants SES-1357666 and DMS-1407655. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

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

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequent. Anal.* **30**, 356-399.

Aoshima, M. and Yata, K. (2015). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Method. Comput. Appl. Probab.* **17**, 419-439.

Bai, Z. and Silverstein, J. W. (2009). *Spectral Analysis of Large Dimensional Random Matrices.* Springer.

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.* **21**, 1275-1294.

Beran, R. (1996). Stein estimation in high dimensions: a retrospective. *Research Developments in Probability and Statistics: Madan L. Puri Festschrift*, 91-110.

Beran, R. et al. (2010). The unbearable transparency of stein estimation. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jureckova*, 25-34. Institute of Mathematical Statistics.

Borysov, P., Hannig, J. and Marron, J. S. (2014). Asymptotics of hierarchical clustering for growing dimension. *J. Multivariate Anal.* **124**, 465-479.

Cabanski, C., Qi, Y., Yin, X., Bair, E., Hayward, M., Fan, C., Li, J., Wilkerson, M., Marron, J. S., Perou, C. and Hayes, D. (2010). Swiss made: standardized within class sum of squares to evaluate methodologies and dataset elements. *PloS One* **5**, e9905.

Cai, T., Fan, J. and Jiang, T. (2013). Distributions of angles in random packing on spheres. *J. Mach. Learn. Res.* **14**, 1837-1864.

Casella, G. and Hwang, J. T. (1982). Limit expressions for the risk of james-stein estimators. *Canad. J. Statist.* **10**, 305-309.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Central limit theorems and bootstrap in high dimensions. arXiv:1412.3661.

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C. et al. (2008). Annotating genomes with massive-scale rna sequencing. *Genome Biol* **9**, R175.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Nonparametr. Statist.* **16**, 111-125.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. Roy. Statist. Soc. Ser. B* **67**, 427-444.

Hellton, K. and Thoresen, M. (2014). Asymptotic distribution of principal component scores for pervasive, high-dimensional eigenvectors. arXiv preprint arXiv:1401.2781.

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

Jolliffe, I. (2005). *Principal Component Analysis*. Wiley Online Library.

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

Jung, S., Sen, A. and Marron, J. S. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. *J. Multivariate Anal.* **109**, 190-203.

Kimes, P. K. and Cabanski, C. R. (2013). Sigfuge (tutorial).

Lee, S., Zou, F. and Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist.* **38**, 3605-3629.

Liu, Y., Hayes, D. N., Nobel, A. and Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low sample size data. *J. Amer. Statist. Assoc.* **103**, 1281-1293.

Marcenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics* **1**, 457-483.

Marron, J. S. and Alonso, A. M. (2014). An overview of object oriented data analysis. *Biometr. J.* **56**, 732-753.

Murillo, F. et al. (2008). The incredible shrinking world of DNA microarrays. *Molecular BioSystems* **4**, 726-732.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17**, 1617-1642.

Paul, D. and Johnstone, I. (2007). Augmented sparse principal component analysis for high-dimensional data. Technical Report, UC Davis.

Portnoy, S. (1984). Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. i. consistency. *Ann. Statist.* **12**, 1298-1309.

Portnoy, S. et al. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356-366.

Ramsay, J. O. (2005). *Functional Data Analysis*. Wiley Online Library.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer New York.

Rollin, A. (2013). Stein's method in high dimensions with applications. arXiv:1101.4454.

Shen, D., Shen, H. and Marron, J. S. (2012a). Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.* **115**, 317-333.

Shen, D., Shen, H. and Marron, J. S. (2012b). A general framework for consistency of principal component analysis. *arXiv:1211.2671*.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 197-206.

Tracy, C. A. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics* **177**, 727-754.

Wang, H. and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* **35**, 1849-1873.

Wilhelm, B. T. and Landry, J.-R. (2009). RNA-seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249-257.

Yata, K. and Aoshima, M. (2009). PCA consistency for non-gaussian data in high dimension, low sample size context. *Comm. Statist. Theory Methods* **38**, 2634-2652.

Yata, K. and Aoshima, M. (2010a). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *J. Multivariate Anal.* **101**, 2060-2077.

Yata, K. and Aoshima, M. (2010b). Intrinsic dimensionality estimation of high-dimension, low sample size data with d-asymptotics. *Comm. Statist. Theory Methods* **39**, 1511-1521.

Yata, K. and Aoshima, M. (2012a). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* **105**, 193-215.

Yata, K. and Aoshima, M. (2012b). Inference on high-dimensional mean vectors with fewer observations than the dimension. *Methodology and Computing in Applied Probability* **14**, 459-476.

Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high- dimensional settings. *J. Multivariate Anal.* **122**, 334-354.

Interdisciplinary Data Sciences Consortium, Department of Mathematics and Statistics, University of South Florida, Tampa, FL, 33620, USA.

E-mail: danshen@usf.edu

School of Business, University of Hong Kong, Pokfulam, Hong Kong.

E-mail: haipeng@hku.hk

Department of Biostatistics and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: htzhu@email.unc.edu

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: marron@unc.edu

(Received March 2015; accepted December 2015)