Boundary behavior in High Dimension, Low Sample Size asymptotics of PCA

Sungkyu Jung[a,*], Arusharka Sen[b], J. S. Marron[c]

[a] *Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.*
[b] *Department of Mathematics and Statistics, Concordia University, Montreal, Quebec H3G 1M8, Canada*
[c] *Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*

Abstract

In High Dimension, Low Sample Size (HDLSS) data situations, where the dimension d is much larger than the sample size n, principal component analysis (PCA) plays an important role in statistical analysis. Under which conditions does the sample PCA well reflect the population covariance structure? We answer this question in a relevant asymptotic context where d grows and n is fixed, under a generalized spiked covariance model. Specifically, we assume the largest population eigenvalues to be of the order d^α, where $\alpha <, =,$ or > 1. Earlier results show the conditions for consistency and strong inconsistency of eigenvectors of the sample covariance matrix. In the boundary case, $\alpha = 1$, where the sample PC directions are neither consistent nor strongly inconsistent, we show that eigenvalues and eigenvectors do not degenerate but have limiting distributions. The result smoothly bridges the phase transition represented by the other two cases, and thus gives a spectrum of limits for the sample PCA in the HDLSS asymptotics. While the results hold under a general situation, the limiting distributions under Gaussian assumption are illustrated in greater detail. In addition, the geometric representation of HDLSS data is extended to give three different representations, that depend on the magnitude of variances in the first few principal components.

*Corresponding author

Email addresses: sungkyu@pitt.edu (Sungkyu Jung), asen@mathstat.concordia.ca (Arusharka Sen), marron@email.unc.edu (J. S. Marron)

Keywords: Principal Component Analysis, high dimension low sample size, geometric representation, ρ-mixing, consistency and strong inconsistency, spiked covariance model

1. Introduction

The study of the covariance matrix and its usual estimator, the sample covariance matrix, is an important issue in multivariate statistics. In particular, the sample covariance matrix provides the conventional estimator of principal component analysis (PCA) through the eigenvalue-eigenvector decomposition. PCA plays an important role in dimension reduction and visualization of important data structure. The High Dimension, Low Sample Size (HDLSS) data situation, where the dimension d of the sample space is much larger than the sample size n, occurs in many areas of modern science, and thus the dimension reduction through PCA is becoming more important for analysis of such data. The sample PCA (through the eigen-decomposition of the sample covariance matrix) is still well-defined when $d > n$, and thus frequently used in practice. Even when the dimension is much higher than the sample size, the PCA has shown to be successful such as in microarray studies [1]. What is the underlying mechanism which leads to the success of PCA in the HDLSS situation? This is the question we answer in this paper.

A central question is whether the sample principal components reflect true underlying distributional structure in the HDLSS context. This has been investigated by comparing the sample eigenvalues and eigenvectors with their population counterparts, in a relevant asymptotic context where the dimension d grows while the sample size n is fixed ([2], [3], [4], [5]). The asymptotic direction of d growing and n fixed is also studied in different contexts; see for instance, [6], [7] and Chapter 4.5 of [8]. While we focus on this asymptotic context in this paper, we also point out that there has been a different approach for the problem where the limits are taken along the direction where d and n grow at the same rate, i.e. $d/n \to c \in (0, \infty)$ as $d \to \infty$. For the result of this type, we refer to [9], [10], [11], [12], [13] and references therein.

In both investigations, the majority of results are well described in a spiked covariance model, proposed by [9]. An exception we point out is a work by [14], which proposes to estimate the spectral distribution of eigenvalues without assuming a spike model.

A spiked covariance model assumes that the first few eigenvalues are distinctively larger than the others. We use a generalized version of the spike model, as described in Section 3, which is different from that of [9] and [11]. Let $\Sigma_{(d)}$ denote the population covariance matrix and $S_{(d)}$ denote the sample covariance matrix. The eigen-decomposition of $\Sigma_{(d)}$ is $\Sigma_{(d)} = U_d \Lambda_d U_d'$, where Λ_d is a diagonal matrix of eigenvalues $\lambda_{1,d} \geq \lambda_{2,d} \geq \cdots \geq \lambda_{d,d}$ in non-increasing order, U_d is a matrix of corresponding eigenvectors so that $U_d = [u_{1,d}, \ldots, u_{d,d}]$, and $'$ denotes the transpose of the preceding matrix. The eigen-decomposition of $S_{(d)}$ is similarly defined as $S_{(d)} = \hat{U}_d \hat{\Lambda}_d \hat{U}_d'$. As a simple example of the spiked model, consider $\lambda_{1,d} = \sigma^2 d^\alpha$, $\lambda_{2,d} = \cdots = \lambda_{d,d} = \tau^2$, for $\alpha, \sigma^2, \tau^2 > 0$ fixed. The first eigenvector of $S_{(d)}$ corresponding to the largest eigenvalue is of interest, as it contains the most important variation of the data. The first sample eigenvector $\hat{u}_{1,d}$ is assessed with the angle formed by itself and its population counterpart $u_{1,d}$. The direction $\hat{u}_{1,d}$ is said to be *consistent* with $u_{1,d}$ if $\text{Angle}(\hat{u}_{1,d}, u_{1,d}) \to 0$ as $d \to \infty$. However in the HDLSS context, a perhaps counter-intuitive phenomenon frequently occurs, where the two directions tend to be as far away as possible. We say the direction $\hat{u}_{1,d}$ is *strongly inconsistent* with $u_{1,d}$ if $\text{Angle}(\hat{u}_{1,d}, u_{1,d}) \to \frac{\pi}{2}$ as $d \to \infty$. In the one spike model above, the order of magnitude α of the first eigenvalue is the key condition for these two limiting phenomena. [3] have shown that

$$\text{Angle}(\hat{u}_{1,d}, u_{1,d}) \to \begin{cases} 0, & \alpha > 1; \\ \frac{\pi}{2}, & \alpha < 1, \end{cases} \tag{1}$$

in probability under some conditions. Although the gap between consistency and strong inconsistency is relatively thin, the case $\alpha = 1$ has not been investigated, and is a main focus of this paper.

It is natural to conjecture from (1) that when $\alpha = 1$, the angle does not degenerate but converges to a random quantity in $(0, \pi/2)$. This claim is established in the simple one spike model in the next section, where we describe a range of limits for the eigenvalues and eigenvectors, depending on the order of magnitude α of $\lambda_{1,d}$. In Section 3, the claim is generalized for multiple spike cases, and is proved in a much more general distributional setting.

The parameter α gives a sharp mathematical boundary for the set of HDLSS situations where the estimated PC direction converges to the population direction. In the boundary, $\alpha = 1$, it will be shown that the estimated PC direction is affected by the true PC directions, but not as strong as the

$\alpha > 1$ case. The success of PCA in the HDLSS situation is understood as an example of the $\alpha \geq 1$ cases, as otherwise the estimated principal components are meaningless as shown in (1).

In a multiple spike model with $m > 1$ spikes, where the first m principal components contain the important *signal* of the distribution, the sample PCA can be assessed by simultaneously comparing the first m principal components. In particular, we investigate the limits of *distance* between two subspaces: the subspace generated by the first m sample PC directions $\hat{u}_{1,d}, \ldots, \hat{u}_{m,d}$ and the subspace by the first m population PC directions. The distance is usefully measured by canonical angles and metrics between subspaces, the limiting distributions of which will be investigated for the $\alpha = 1$ case, as well as the cases $\alpha \neq 1$, in Section 3.3. The probability density functions of the limiting distributions for $\alpha = 1$ are also derived and illustrated under a Gaussian assumption, to show the effect of parameters in the distributions.

The HDLSS data set has an interesting geometric representation in the limit $d \to \infty$, as shown in [15]. In Section 4, we extend the result and show that there are three different geometric representations, which coincide with the range of limits depending on α.

2. Range of limits in the single spike model

Suppose we have a data matrix $X_{(d)} = [X_{1,d}, \ldots, X_{n,d}]$, with $d > n$, where the d dimensional random vectors $X_{i,d}$ are independent and identically distributed. We assume for now that $X_{i,d}$ is normally distributed with mean zero and covariance matrix $\mathbf{\Sigma}_{(d)}$, but the Gaussian assumption will be relaxed in the next section. The population covariance matrix $\mathbf{\Sigma}_{(d)}$ is assumed to have one spike, that is, the eigenvalues of $\mathbf{\Sigma}_{(d)}$ are $\lambda_{1,d} = \sigma^2 d^\alpha$, $\lambda_{2,d} = \cdots = \lambda_{d,d} = \tau^2$. The corresponding eigenvectors of $\mathbf{\Sigma}_{(d)}$ are denoted by $u_{i,d}$. The sample covariance matrix is defined as $S_{(d)} = \frac{1}{n} X_{(d)} X'_{(d)}$ with its ith eigenvalue and eigenvector denoted by $\hat{\lambda}_{i,d}$ and $\hat{u}_{i,d}$, respectively.

The following theorem summarizes the spectrum of the limiting distributions of the eigenvalues and eigenvectors of $S_{(d)}$, depending on the different order α of $\lambda_{1,d}$. Note that the angle between the two vectors $u, \hat{u}$ is represented by the inner product through $\text{Angle}(u, \hat{u}) = \cos^{-1}(u'\hat{u})$. For the eigenvectors with common eigenvalues, there are of course an infinite number of choices. The argument in the following theorem assumes that we *choose* a set of population eigenvectors $u_{j,d}$ before obtaining $\hat{u}_{j,d}$.

Theorem 1. *Under the Gaussian assumption and the one spike case above, (i) the limit of the first eigenvalue depends on α:*

$$\frac{\hat{\lambda}_{1,d}}{\max(d^\alpha, d)} \Longrightarrow \begin{cases} \sigma^2 \frac{\chi_n^2}{n}, & \alpha > 1; \\ \sigma^2 \frac{\chi_n^2}{n} + \frac{\tau^2}{n}, & \alpha = 1; \\ \frac{\tau^2}{n}, & \alpha < 1, \end{cases}$$

as $d \to \infty$, where $\Longrightarrow$ denotes the convergence in distribution, and χ_n^2 denotes a random variable with the χ^2 distribution with degrees of freedom n. The rest of the eigenvalues converge to the same quantity when scaled, that is for any $\alpha \in [0, \infty)$, $j = 2, \ldots, n$,

$$\frac{\hat{\lambda}_{j,d}}{d} \to \frac{\tau^2}{n}, \text{ as } d \to \infty,$$

in probability.

(ii) The limit of the first eigenvector depends on α:

$$u_{1,d}' \hat{u}_{1,d} \Longrightarrow \begin{cases} 1 & \alpha > 1; \\ \left(1 + \frac{\tau^2}{\sigma^2 \chi_n^2}\right)^{-\frac{1}{2}} & \alpha = 1; \\ 0, & \alpha < 1, \end{cases}$$

as $d \to \infty$. The rest of the eigenvectors are strongly inconsistent with their population counterpart, for any $\alpha \in [0, \infty)$, $j = 2, \ldots, n$,

$$u_{j,d}' \hat{u}_{j,d} \to 0, \text{ as } d \to \infty,$$

in probability.

The case $\alpha = 1$ bridges the other two cases. In particular, the ratio of the sample and population eigenvalue $\hat{\lambda}_{1,d}/\lambda_{1,d}$ is asymptotically unbiased to 1 when $\alpha > 1$. It is asymptotically biased when $\alpha = 1$, and becomes completely deterministic in the case $\alpha < 1$, where the effect of σ^2 on $\hat{\lambda}_{1,d}$ becomes negligible. Moreover, the angle $\text{Angle}(u_{1,d}, \hat{u}_{1,d})$ to the optimal direction converges to a random quantity which is defined on $(0, \pi/2)$ and depends on σ^2, τ^2, and n. The effect of those parameters on the limiting distribution of $\text{Angle}(u_{1,d}, \hat{u}_{1,d})$ is illustrated in Fig. 1. The ratio σ^2/τ^2 can be understood as a *signal to noise* ratio. A high value of σ^2/τ^2 means that the major variation along the first PC direction is strong. Therefore, for larger values of σ^2/τ^2,

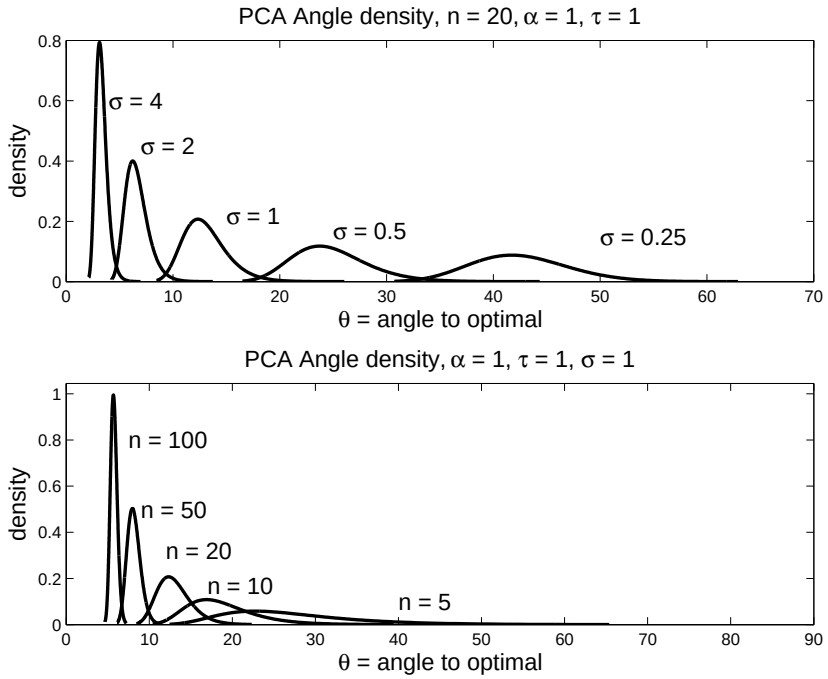

Figure 1: Angle densities for the one spike case. The top panel shows an overlay of the densities with different σ^2, with other parameters fixed. The bottom panel shows an overlay of the densities with different degrees of freedom n of the χ^2 distribution. For a larger signal to noise ratio $\frac{\sigma^2}{\tau^2}$, and for a larger n, the angle to optimal is smaller.

Angle$(u_{1,d}, \hat{u}_{1,d})$ should be closer to zero than smaller values of the ratio, as depicted in the upper panel of Fig. 1. Moreover, the sample PCA with larger sample size n should perform better than with smaller sample size. The sample size n becomes the degrees of freedom of the χ^2 distribution in the limit, and the bottom panel of Fig. 1 shows that the $\hat{u}_{i,d}$ is closer to $u_{i,d}$ for larger values of n.

The Gaussian assumption in the previous theorem appears as a driver of the limiting χ^2 distributions. Under the general non-Gaussian assumption we state in the next section, the χ^2 will be replaced by a distribution that depends heavily on the distribution of the population principal component scores, which may not be Gaussian.

Remark 1. The results in Theorem 1 can be used to estimate the parameters σ^2 and τ^2 in the model with $\alpha = 1$. As a simple example, one can set $\hat{\tau}^2 = \frac{n}{n-1} \sum_{j=2}^{n} \frac{\hat{\lambda}_{j,d}}{d}$ and $\hat{\sigma}^2 = \hat{\lambda}_{1,d}/d - \hat{\tau}^2/n$. Then $\hat{\tau}^2 \to \tau^2$ and $\hat{\sigma}^2 \Longrightarrow \sigma^2 \frac{\chi^2_n}{n}$

as $d \to \infty$ by Theorem 1 and Slutsky's theorem. The estimator $\hat{\sigma}^2$ is not consistent but can evidently be used to construct a confidence interval for σ^2:

$$P\left\{\frac{n\hat{\sigma}^2}{\chi^2_{n,1-(a/2)}} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{\chi^2_{n,a/2}}\right\} \to (1-a) \text{ as } d \to \infty.$$

This can be extended to provide asymptotic confidence intervals for principal component scores. A similar estimation scheme can be found in a related but different setting where $d, n \to \infty$ together; see for example [11] and [13]. These papers do not discuss eigenvector estimation. The sample eigenvector, $\hat{u}_{1,d}$ is difficult to improve upon mainly because the direction of deviation $\hat{u}_{1,d}$ from $u_{1,d}$ is quite random unless a more restrictive assumption (e.g. sparsity) is made.

3. Limits under generalized spiked covariance model

The results in the previous section will be generalized to much broader situations, including a generalized spiked covariance model and a relaxation of the Gaussian assumption. We focus on the $\alpha = 1$ case, and describe the limiting distributions for eigenvalues and eigenvectors.

3.1. Eigenvalues and eigenvectors

In the following, all the quantities depend on the dimension d, but the subscript d is omitted when it does not cause any confusion. We first describe some elementary facts from matrix algebra, that are useful throughout the paper. The dimension of the sample covariance matrix S increases as d grows, so it is challenging to deal with S directly. A useful approach is to use the dual of S, defined as the $n \times n$ symmetric matrix

$$S_D = \frac{1}{n} X'X,$$

by switching the role of rows and columns of X. The (i,j)th element of S_D is $\frac{1}{n} X_i' X_j$. An advantage of working with S_D is that for large d, the finite dimensional matrix S_D is positive definite with probability one, and its n eigenvalues are the same as the non-zero eigenvalues of S. Moreover, the sample eigenvectors $\hat{u}_i$ are related to the eigen-decomposition of S_D, as shown next. Let $S_D = \hat{V}_n \hat{\Lambda}_n \hat{V}_n$, where $\hat{\Lambda}_n = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$ and $\hat{V}_n$ is the $n \times n$ orthogonal matrix of eigenvectors $\hat{v}_i$ corresponding to $\hat{\lambda}_i$. Recall $S = \hat{U}\hat{\Lambda}\hat{U}'$.

Since S is at most rank n, we can write $S = \hat{U}_n \hat{\Lambda}_n \hat{U}_n'$, where $\hat{U}_n = [\hat{u}_1, \ldots, \hat{u}_n]$ consists of the first n columns of $\hat{U}$. The singular value decomposition of X is given by

$$X = \hat{U}_n \hat{\Lambda}_n \hat{V}_n' = \sum_{i=1}^{n} (n\hat{\lambda}_i)^{-\frac{1}{2}} \hat{u}_i \hat{v}_i'.$$

Then the kth sample principal component direction $\hat{u}_k$ for $k \leq n$ is proportional to $X\hat{v}_k$,

$$\hat{u}_k = (n\hat{\lambda}_k)^{-\frac{1}{2}} X \hat{v}_k. \tag{2}$$

Therefore the asymptotic properties of the eigen-decomposition of S, as $d \to \infty$, can be studied via those of the finite dimensional matrix S_D.

It is also useful to represent S_D in terms of the population principal components. Let $Z_{(d)}$ be the standardized principal components of X, defined by

$$Z_{(d)} = \begin{pmatrix} Z_1' \\ \vdots \\ Z_d' \end{pmatrix} = \Lambda_d^{-1/2} U_d' X,$$

where $Z_i' = (Z_{i1}, \ldots, Z_{in})$ is the ith row of $Z_{(d)}$, so that

$$Z_i' = \lambda_i^{-\frac{1}{2}} u_i' X. \tag{3}$$

Under the Gaussian assumption of the previous section, each element of $Z_{(d)}$ is independently distributed as the standard normal distribution. By $X = U_d \Lambda^{1/2} Z_{(d)}$,

$$S_D = \frac{1}{n} X'X = \frac{1}{n} Z_{(d)}' \Lambda Z_{(d)} = \frac{1}{n} \sum_{i=1}^{d} \lambda_i Z_i Z_i'.$$

The Gaussian assumption on X is relaxed as follows. We assume that each column X_i of X follows a d dimensional multivariate distribution with mean zero and covariance matrix Σ. Each entry of the standardized principal components, or the sphered variables $Z_{(d)}$ is assumed to have finite fourth moments, and is uncorrelated but in general dependent with each other. We regulate the dependency of the principal components by a ρ-mixing condition (see [16], [17]). We briefly describe a version of ρ-mixing for our purpose. For $-\infty \leq J \leq L \leq \infty$, let $\mathcal{F}_J^L$ denote the σ-field of events generated by the random variables Z_i, $J \leq i \leq L$. For any σ-field $\mathcal{A}$, let $L_2(\mathcal{A})$ denote the

space of square-integrable, $\mathcal{A}$ measurable real-valued random variables. For each $m \geq 1$, define the maximal correlation coefficient

$$\rho(m) := \sup |\text{corr}(f, g)|, \quad f \in L_2(\mathcal{F}_{-\infty}^{j}),\ g \in L_2(\mathcal{F}_{j+m}^{\infty}),$$

where sup is over all f, g and $j \in \mathbf{Z}$. The sequence $\{Z_i\}$ is said to be ρ-mixing if $\rho(m) \to 0$ as $m \to \infty$.

While the concept of ρ-mixing is useful as a mild condition for the development of laws of large numbers, its formulation is critically dependent on the ordering of variables. Therefore we assume that there is some permutation of the data which is ρ-mixing. In particular, let $\{Z_{ij,(d)}\}_{i=1}^{d}$ be the components of the jth column vector of $Z_{(d)}$. We assume that for each d, there exists a permutation $\pi_d : \{1, \ldots, d\} \longmapsto \{1, \ldots, d\}$ so that the sequence $\{Z_{\pi_d(i)j,(d)} : i = 1, \ldots, d\}$ is ρ-mixing. This assumption makes the results invariant under a permutation of the variables. We denote these distributional assumptions as $(c1)$.

We then define a generalized spiked covariance model. Recall that a simple one spike model was defined on the eigenvalues of the population covariance matrix Σ, for example, $\lambda_1 = \sigma^2 d^\alpha$, $\lambda_2 = \cdots = \lambda_d = \tau^2$. This is generalized by allowing multiple spikes, and by relaxing the uniform eigenvalue assumption in the tail to a decreasing sequence. The tail eigenvalues are regulated by a measure of sphericity ϵ_k in the limit $d \to \infty$. The measure of sphericity ϵ_k, $k = 1, 2, \ldots$, is defined for $\{\lambda_k, \ldots, \lambda_d\}$ as

$$\epsilon_k(d) \equiv \frac{(\sum_{i=k}^{d} \lambda_i)^2}{d \sum_{i=k}^{d} \lambda_i^2},$$

which is away from 0 and close to 1 when $\{\lambda_k, \ldots, \lambda_d\}$ are close to each other. Then we shall assume that the tail eigenvalues do not decrease too fast. In particular, we say the ϵ_k-*condition* is satisfied when $\epsilon_k(d)$ decreases at a rate slower than d^{-1}, i.e.

$$(d\epsilon_k)^{-1} = \frac{\sum_{i=k}^{d} \lambda_i^2}{(\sum_{i=k}^{d} \lambda_i)^2} \to 0 \text{ as } d \to \infty.$$

The ϵ_k condition holds for quite general settings ([3], Sec. 2). As an example, a polynomially decreasing sequence $i^{-\frac{1}{2}}$ of eigenvalues satisfies the condition ϵ_k with $k = 1$. In the generalized spiked model, the eigenvalues are assumed to be of the form $\lambda_1 = \sigma_1^2 d^\alpha, \ldots, \lambda_m = \sigma_m^2 d^\alpha$, for $\sigma_1^2 \geq \cdots \geq \sigma_m^2 > 0$ for some

$m \geq 1$, and the ϵ_{m+1} condition holds for $\lambda_{m+1}, \ldots, \lambda_d$. Also assume that $\frac{1}{d}\sum_{i=m+1}^{d} \lambda_i \to \tau^2$ as $d \to \infty$. These conditions for spike models are denoted by $(c2)$.

The following theorem gives the limits of the sample eigenvalues and eigenvectors under the general assumptions in this section. We use the following notations. Let $\varphi(A)$ be a vector of eigenvalues of a real symmetric matrix A arranged in non-increasing order and let $\varphi_i(A)$ be the ith largest eigenvalue of A. Let $v_i(A)$ denote the ith eigenvector of the matrix A corresponding to the eigenvalue $\varphi_i(A)$ and $v_{ij}(A)$ be the jth loading of $v_i(A)$. Also note that there are many choices of eigenvectors of S including the sign changes. We use the convention that the sign of $\hat{u}_i$ will be chosen so that $\hat{u}_i' u_i \geq 0$. Recall that the vector of the ith standardized principal component scores is $Z_i = (Z_{1i}, \ldots, Z_{ni})'$. Denote the $n \times m$ matrix of the first m principal component scores as $\mathbf{W} = [\sigma_1 Z_1, \cdots, \sigma_m Z_m]$. The limiting distributions heavily depend on the finite dimensional random matrix $\mathbf{W}$.

Theorem 2. *Under the assumptions $(c1)$ and $(c2)$ with fixed $n \geq m \geq 1$, if $\alpha = 1$, then (i) the sample eigenvalues*

$$d^{-1} n \hat{\lambda}_{i,d} \Longrightarrow \begin{cases} \varphi_i(\mathbf{W}'\mathbf{W}) + \tau^2, & i = 1, \ldots, m; \\ \tau^2, & i = m+1, \ldots, n, \end{cases}$$

as $d \to \infty$ jointly for all i.

(ii) The inner products between the sample and population eigenvectors have limiting distributions:

$$\hat{u}_{i,d}' u_{j,d} \Longrightarrow \frac{v_{ij}(\mathbf{W}'\mathbf{W})}{\sqrt{1 + \tau^2/\varphi_i(\mathbf{W}'\mathbf{W})}} \text{ as } d \to \infty \text{ jointly for } i, j = 1, \ldots, m.$$

The rest of eigenvectors are strongly inconsistent with their population counterpart, i.e.

$$\hat{u}_{i,d}' u_{i,d} \to 0 \text{ as } d \to \infty \text{ for } i = m+1, \ldots, n,$$

in probability.

The theorem shows that the first m eigenvectors are neither consistent nor strongly inconsistent to the population counterparts. The limiting distributions of angles $\text{Angle}(\hat{u}_{i,d}, u_{i,d})$ to optimal directions are supported on $(0, \pi/2)$ and depend on the magnitude of the noise τ^2 and the distribution of $\mathbf{W}'\mathbf{W}$. Note that the $m \times m$ symmetric matrix $\mathbf{W}'\mathbf{W}$ is the scaled covariance

matrix of the principal component scores in the first m directions. When the underlying distribution of $\mathbf{X}$ is assumed to be Gaussian, then $\mathbf{W}'\mathbf{W}$ is the Wishart matrix $\mathcal{W}_m(n, \Lambda_m)$, where $\Lambda_m = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. If $m = 1$, the matrix becomes a scalar random variable $\mathbf{W}'\mathbf{W} = \varphi_1(\mathbf{W}'\mathbf{W})$, and is χ_n^2 under the Gaussian assumption, which leads to Theorem 1.

In general, the distribution of $\varphi_i(\mathbf{W}'\mathbf{W})$ is not simply described. We refer to [18, Ch. 9] for the Gaussian case, and [19] for the case $m \to \infty$.

The limiting distributions of the cases $\alpha \neq 1$ can be found in a similar manner, which we only state the result for the case $\alpha > 1$ and for the first m components. We refer to [3] for more general results. For $i, j = 1, \ldots, m$, the eigenvalues $d^{-\alpha} n \hat{\lambda}_{i,d} \Longrightarrow \phi_i(\mathbf{W}'\mathbf{W})$ and the inner products $\hat{u}_{i,d}' u_{j,d} \Longrightarrow v_{ij}(\mathbf{W}'\mathbf{W})$ as $d \to \infty$. In comparison to the $\alpha = 1$ case, if we set τ^2 to be zero, the result becomes identical for all $\alpha \geq 1$.

Remark 2. When the sample size n also grows, consistency of sample eigenvalues and eigenvectors can be achieved. In particular, for $i = 1, \ldots, m$, we have as d grows

$$\frac{\hat{\lambda}_{i,d}}{\lambda_{i,d}} = \frac{d}{n} \frac{d^{-1} n \hat{\lambda}_{i,d}}{\sigma_i^2 d} \Longrightarrow \frac{\varphi_i(\mathbf{W}'\mathbf{W}/n)}{\sigma_i^2} + \frac{\tau^2}{n\sigma_i^2}$$

by Theorem 2. Since $\mathbf{W}'\mathbf{W}/n \to \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ by a law of large numbers, we get the consistency of eigenvalues, i.e. $\hat{\lambda}_{i,d}/\lambda_{i,d} \to 1$ as $d, n \to \infty$, where the limits are applied successively. For the sample eigenvectors, from Theorem 2(ii) and because $v_{ij}(\mathbf{W}^T\mathbf{W}) \to 1$ if $i = j$, 0 otherwise as $n \to \infty$ and $\varphi_i(\mathbf{W}'\mathbf{W}) = O(n)$, we get $\hat{u}_{i,d}' u_{j,d} \to 1$ if $i = j$, 0 otherwise as $d, n \to \infty$. Therefore, the sample PC directions are consistent to the corresponding population PC directions, i.e. $\mathrm{Angle}(\hat{u}_{i,d}, u_{i,d}) \to 0$ as $d, n \to \infty$ (applied successively), for $i \leq m$. Therefore it is conjectured that when d, n grow together with $d \gg n$, a similar conclusion to Theorem 2 can be drawn.

Proof of Theorem 2. The following lemma (Theorem 1 of [3]) shows a version of the law of large numbers for matrices, that is useful in the proof. Recall that $Z_i' \equiv (Z_{1i}, \ldots, Z_{ni})$ is the ith row of $Z_{(d)}$.

Lemma 1. *If the assumption (c*1*) and the ϵ_k-condition holds, then*

$$c_d^{-1} \sum_{i=k}^{d} \lambda_{i,d} Z_i Z_i' \to I_n, \ \textit{as } d \to \infty$$

in probability, where $c_d = n^{-1}\sum_{i=1}^{d}\lambda_{i,d}$ *and* I_n *denotes the* $n \times n$ *identity matrix. In particular, if* $d^{-1}\sum_{i=k}^{d}\lambda_{i,d} \to \tau^2$*, then*

$$\frac{1}{d}\sum_{i=k}^{d}\lambda_{i,d}Z_iZ_i' \to \tau^2 I_n, \text{ as } d \to \infty$$

in probability.

This lemma is used to show that the spectral decomposition of $d^{-1}nS_D$,

$$d^{-1}nS_D = \sum_{i=1}^{m}\sigma_i^2 Z_iZ_i' + d^{-1}\sum_{i=m+1}^{d}\lambda_i Z_iZ_i',$$

can be divided into two parts, and the latter converges to a deterministic part. Applying Lemma 1, we have $d^{-1}nS_D \Longrightarrow S_0$ as $d \to \infty$, where

$$S_0 = \mathbf{W}\mathbf{W}' + \tau^2 I_n.$$

Then since the eigenvalues of a symmetric matrix A are a continuous function of elements of A, we have

$$\varphi(d^{-1}nS_D) \Longrightarrow \varphi(S_0),$$

as $d \to \infty$. Noticing that for $i = 1, \ldots, m$,

$$\varphi_i(S_0) = \varphi_i(\mathbf{W}\mathbf{W}') + \tau^2 = \varphi_i(\mathbf{W}'\mathbf{W}) + \tau^2,$$

and for $i = m+1, \ldots, n$, $\varphi_i(S_0) = \tau^2$ gives the result.

For the eigenvectors, note that the eigenvectors $\hat{v}_i$ of $d^{-1}nS_D$ can be chosen so that they are continuous ([20]). Therefore, we also have that $\hat{v}_i = v_i(d^{-1}nS_D) \Longrightarrow v_i(S_0)$ as $d \to \infty$, for all i. Also note that $v_i(S_0) = v_i(\mathbf{W}\mathbf{W}')$ for $i \leq m$.

Similar to the dual approach for covariance matrices, the eigenvectors of the $n \times n$ matrix $\mathbf{W}\mathbf{W}'$ can be evaluated from the dual of the matrix. In particular, let $\mathbf{W} = U_w\Lambda_w V_w' = \sum_{i=1}^{m}\lambda_{iw}u_{iw}v_{iw}'$, where $\lambda_{iw}^2 = \varphi_i(\mathbf{W}'\mathbf{W})$ and $v_{iw} = v_i(\mathbf{W}'\mathbf{W})$. Then

$$v_1(S_0) = u_{iw} = \frac{\mathbf{W}v_{iw}}{\lambda_{iw}} = \frac{\mathbf{W}v_i(\mathbf{W}'\mathbf{W})}{\sqrt{\varphi_i(\mathbf{W}'\mathbf{W})}}.$$

Now from (2),(3) and the previous equation, for $1 \leq i, j \leq m$,

$$\begin{aligned} u_j'\hat{u}_i &= u_j' \frac{X\hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = \frac{u_j' X \hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = \frac{\sqrt{\sigma_j^2 d} Z_j' \hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = \frac{\sigma_j Z_j' \hat{v}_i}{\sqrt{d^{-1} n \hat{\lambda}_i}} \\ &\Longrightarrow \frac{\sigma_j Z_j' \mathbf{W} v_i(\mathbf{W}'\mathbf{W})}{\sqrt{\varphi_i(\mathbf{W}'\mathbf{W}) + \tau^2}\sqrt{\varphi_i(\mathbf{W}'\mathbf{W})}}. \end{aligned} \tag{4}$$

Note that $\sigma_j Z_j'\mathbf{W} = [\sigma_j\sigma_1 Z_j'Z_1 \cdots \sigma_j\sigma_m Z_j'Z_m]$ is the jth row of $\mathbf{W}'\mathbf{W}$ and $\mathbf{W}'\mathbf{W}v_i(\mathbf{W}'\mathbf{W}) = \varphi_i(\mathbf{W}'\mathbf{W})v_i(\mathbf{W}'\mathbf{W})$. Therefore, the limiting form (4) becomes

$$\frac{\varphi_i(\mathbf{W}'\mathbf{W}) v_{ij}(\mathbf{W}'\mathbf{W})}{\sqrt{\varphi_i(\mathbf{W}'\mathbf{W}) + \tau^2}\sqrt{\varphi_i(\mathbf{W}'\mathbf{W})}}.$$

For $i = m+1, \ldots, n$, again from (2) and (3), we get

$$u_i'\hat{u}_i = \frac{\sqrt{\lambda_i} Z_i' \hat{v}_i}{\sqrt{n\hat{\lambda}_i}} = d^{-\frac{1}{2}} \frac{\tau Z_i' \hat{v}_i}{\sqrt{n\hat{\lambda}_i/d}} = O_p(d^{-\frac{1}{2}}).$$

□

3.2. Asymptotic results for centered data matrix

In practice, the sample covariance matrix is usually derived from the centered data matrix. In such a case, we obtain a weaker result than Theorem 2. Let $\tilde{S}_{(d)} = n^{-1}(X_d - \bar{X})(X_d - \bar{X})'$, where $\bar{X} = n^{-1}\sum_{i=1}^n X_{i,d} 1_n'$ is a $d \times n$ matrix consisting of n columns of the sample mean vector.

For a general mean vector μ, the representation of X in terms of standardized principal components $Z_{(d)}$ is

$$X = \mu 1_n' + U_d \Lambda^{\frac{1}{2}} Z_{(d)}$$

and thus

$$X - \bar{X} = U_d \Lambda^{\frac{1}{2}} (Z_{(d)} - \bar{Z}),$$

where the ith row of $\bar{Z}$ is $\bar{z}_i' = n^{-1}\sum_{j=1}^n Z_{ij} 1_n' = n^{-1} Z_i' J_n$. The symbol J_n represents the $n \times n$ matrix consisting entirely of ones.

The dual covariance matrix of $\tilde{S}_{(d)}$ is then $\tilde{S}_D = n^{-1}\sum_{i=1}^d \lambda_i (Z_i - \bar{z}_i)(Z_i - \bar{z}_i)' = n^{-1}(I_n - n^{-1}J_n)\sum_{i=1}^d \lambda_i Z_i Z_i' (I_n - n^{-1}J_n)'$. Then we have the following result.

Proposition 1. *Let $\tilde{\lambda}_{i,d}$ be the ith largest eigenvalue of $\tilde{S}_{(d)}$. Under the assumptions (c1) and (c2) with fixed $n > m \geq 1$, if $\alpha = 1$, then for any $\epsilon > 0$,*

$$\lim_{d\to\infty} P\left[\bigcap_{i=1}^{m}\{\frac{n}{d}\tilde{\lambda}_{i,d} \in [\varphi_i(\tilde{\mathbf{W}}'\tilde{\mathbf{W}}), \varphi_i(\tilde{\mathbf{W}}'\tilde{\mathbf{W}}) + \tau^2]\} \bigcap_{i=m+1}^{n-1}\{\frac{n}{d}\tilde{\lambda}_{i,d} \in (\tau^2 - \epsilon, \tau^2 + \epsilon)\}\right] = 1,$$

where the $n \times m$ matrix of the first m principal component scores is $\tilde{\mathbf{W}} = [\sigma_1(Z_1 - \bar{z}_1), \cdots, \sigma_m(Z_m - \bar{z}_m)]$.

Proof of Proposition 1. Note that $\tilde{\lambda}_{i,d} = \varphi_i(\tilde{S}_{(d)}) = \varphi_i(\tilde{S}_D)$ for $i = 1, \ldots, n-1$. Similar to the proof of Theorem 2, the spectral decomposition of $d^{-1}n\tilde{S}_D$ is divided into two parts,

$$d^{-1}n\tilde{S}_D = \sum_{i=1}^{m}\sigma_i^2(Z_i - \bar{z}_i)(Z_i - \bar{z}_i)' + (I_n - n^{-1}J_n)\frac{1}{d}\sum_{i=m+1}^{d}\lambda_i Z_i Z_i'(I_n - n^{-1}J_n)',$$

and using Lemma 1, $d^{-1}n\tilde{S}_D$ converges in distribution to $\tilde{S}_0 = \tilde{\mathbf{W}}\tilde{\mathbf{W}}' + \tau^2(I_n - n^{-1}J_n)$. Then the eigenvalues of $d^{-1}n\tilde{S}_D$ jointly converges to the eigenvalues of $\tilde{S}_0$.

For $i = 1, \ldots, m$, Weyl's inequality ([3, p. 4121], [21]) yields that

$$\varphi_i(\tilde{\mathbf{W}}\tilde{\mathbf{W}}') + \varphi_n\{\tau^2(I_n - n^{-1}J_n)\} \leq \varphi_i(\tilde{S}_0) \leq \varphi_i(\tilde{\mathbf{W}}\tilde{\mathbf{W}}') + \varphi_1\{\tau^2(I_n - n^{-1}J_n)\},$$

where $\varphi_j\{\tau^2(I_n - n^{-1}J_n)\} = \tau^2$ for $j = 1, \ldots, n-1$ and 0 for $j = n$, and $\varphi_i(\tilde{\mathbf{W}}\tilde{\mathbf{W}}') = \varphi_i(\tilde{\mathbf{W}}'\tilde{\mathbf{W}})$.

For $i = m+1, \ldots, n-1$, also applying Weyl's inequality gives

$$\varphi_n(\tilde{\mathbf{W}}\tilde{\mathbf{W}}') + \varphi_i\{\tau^2(I_n - n^{-1}J_n)\} \leq \varphi_i(\tilde{S}_0) \leq \varphi_i(\tilde{\mathbf{W}}\tilde{\mathbf{W}}') + \varphi_1\{\tau^2(I_n - n^{-1}J_n)\},$$

and because the rank of $\tilde{\mathbf{W}}\tilde{\mathbf{W}}'$ is at most m, $\varphi_i(\tilde{\mathbf{W}}\tilde{\mathbf{W}}') = 0$ for $i > m$. Thus, $\varphi_i(\tilde{S}_0) = \tau^2$, which leads to $d^{-1}n\tilde{\lambda}_{i,d} \to \tau^2$ in probability as $d \to \infty$. The result is derived by combining the cases $i = 1, \ldots, n-1$. □

When the centered data matrix $X - \bar{X}$ is used, the scaled eigenvalue estimate no longer converges in distribution to $\varphi_i(\tilde{\mathbf{W}}'\tilde{\mathbf{W}}) + \tau^2$. However, the difference becomes smaller for larger n, since the centering matrix $I_n - n^{-1}J_n$ is close to I_n for large n. For the rest of the paper, we assume that the mean is known and zero for the sake of clear presentation.

3.3. Angles between principal component spaces

Under the generalized spiked covariance model with $m > 1$, the first m population principal directions provide a basis of the most important variation. Therefore, it would be more informative to investigate the deviation of each $\hat{u}_i$ from the subspace $\mathcal{L}_1^m(d)$ spanned by $\{u_{1,d}, \ldots, u_{m,d}\}$. Also, denote the subspace spanned by the first m sample principal directions as $\hat{\mathcal{L}}_1^m(d) \equiv \text{span}\{\hat{u}_{1,d}, \ldots, \hat{u}_{m,d}\}$. When performing dimension reduction, it is critical for the sample PC space $\hat{\mathcal{L}}_1^m$ to be close to the population PC space $\mathcal{L}_1^m$. The closeness of two subspaces can be measured in terms of *canonical angles*.

We briefly introduce the notion of canonical angles and metrics between subspaces, detailed discussions of which can be found in [22] and [23]. As a simple case, the canonical angle between a 1-dimensional subspace and an m-dimensional subspace is defined as follows. Let $\hat{\mathcal{L}}_i$ be the 1-dimensional linear space with basis $\hat{u}_i$. Infinitely many angles can be formed between $\hat{\mathcal{L}}_i$ and $\mathcal{L}_1^m$ with $m > 1$. The canonical angle, denoted by $\text{Angle}(\hat{\mathcal{L}}_i, \mathcal{L}_1^m)$, is defined by the smallest angle formed, that is the angle between $\hat{u}_i$ and its projection $\hat{u}_i^P$ onto $\mathcal{L}_1^m$. This angle is represented in terms of an inner product as

$$\text{Angle}(\hat{\mathcal{L}}_i, \mathcal{L}_1^m) = \cos^{-1}\left(\frac{\hat{u}_i' \hat{u}_i^P}{\|\hat{u}_i^P\| \, \|\hat{u}_i\|}\right) = \min_{\mathbf{y} \in \mathcal{L}_1^m} \text{Angle}(\hat{u}_i, \mathbf{y}) \text{ for } \|\mathbf{y}\| > 0. \quad (5)$$

When two multi-dimensional subspaces are considered, multiple canonical angles are defined. Among angles between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$, the first canonical angle is geometrically defined as

$$\theta_1(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \max_{\mathbf{x} \in \hat{\mathcal{L}}_1^m} \min_{\mathbf{y} \in \mathcal{L}_1^m} \text{Angle}(\mathbf{x}, \mathbf{y}) \text{ for } \|\mathbf{x}\|, \|\mathbf{y}\| > 0, \quad (6)$$

where $\text{Angle}(\mathbf{x}, \mathbf{y})$ is the angle formed by the two vectors $\mathbf{x}$, $\mathbf{y}$. One can show that the second canonical angle is defined by the same geometric relation as above with $\hat{\mathcal{L}}_{-\mathbf{x}}$ and $\mathcal{L}_{-\mathbf{y}}$ for $\mathbf{x}$, $\mathbf{y}$ from (6), where $\hat{\mathcal{L}}_{-\mathbf{x}}$ is the orthogonal complement of $\mathbf{x}$ in $\hat{\mathcal{L}}_1^m$. In practice, the canonical angles are found by the singular value decomposition of a matrix. Let $\hat{U}_m$ and U_m be orthonormal bases for $\hat{\mathcal{L}}_1^m$, $\mathcal{L}_1^m$ and γ_i's be the singular values of $\hat{U}_m' U_m$. Then the canonical angles are

$$\theta_i(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \cos^{-1}(\gamma_i)$$

in descending order.

Distances between two subspaces can be defined using the canonical angles. We point out two metrics from Chapter II.4 of [22]:

1. *gap metric* $\rho_g(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \sin(\theta_1)$,
2. *Euclidean sine metric* $\rho_s(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) = \{\sum_{i=1}^m \sin^2(\theta_i)\}^{\frac{1}{2}}$.

The gap metric is simple and only involves the largest canonical angle. The Euclidean sine metric makes use of all canonical angles and thus gives a more comprehensive understanding of the closeness between the two subspaces. We will use both metrics in the following discussion.

At first, we examine the limiting distribution of the angle between the sample PC direction $\hat{u}_i$ and $\mathcal{L}_1^m$.

Theorem 3. *Under the assumptions (c1) and (c2) with fixed $n \geq m \geq 1$, if $\alpha = 1$, then for $i = 1, \ldots, m$, the canonical angle converges in distribution:*

$$\cos\left(Angle(\hat{\mathcal{L}}_i, \mathcal{L}_1^m)\right) \Longrightarrow \frac{1}{\sqrt{1 + \tau^2/\varphi_i(\mathbf{W}'\mathbf{W})}} \text{ as } d \to \infty.$$

Proof. Since $\hat{u}_i^P = \sum_{j=1}^m (\hat{u}_i' u_j) u_j$,

$$\frac{\hat{u}_i' \hat{u}_i^P}{\|\hat{u}_i^P\| \, \|\hat{u}_i\|} = \left\|\hat{u}_i^P\right\| = \sqrt{(\hat{u}_i' u_1)^2 + \cdots + (\hat{u}_i' u_m)^2}.$$

The result follows from (5), Theorem 2(ii) and the fact that $\sum_{j=1}^m \left(v_{ij}(\mathbf{W}'\mathbf{W})\right)^2 = \left\|v_i(\mathbf{W}'\mathbf{W})\right\|^2 = 1$. □

We then investigate the limiting behavior of the distances between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$, in terms of either the canonical angles or the distances. From the fact that $\hat{U}_m = [\hat{u}_1, \ldots, \hat{u}_m]$ and $U_m = [u_1, \ldots, u_m]$ are orthonormal bases of $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$ respectively, cosines of the canonical angles are the singular values of $\hat{U}_m' U_m$. Since the (i, j)th element of $\hat{U}_m' U_m$ is $\hat{u}_i' u_j$, Theorem 2 leads to

$$\hat{U}_m' U_m \Longrightarrow [v_1(\mathbf{W}'\mathbf{W}) \cdots v_m(\mathbf{W}'\mathbf{W})] \begin{pmatrix} \left(1 + \frac{\tau^2}{\varphi_1(\mathbf{W}'\mathbf{W})}\right)^{-\frac{1}{2}} & & 0 \\ & \ddots & \\ 0 & & \left(1 + \frac{\tau^2}{\varphi_m(\mathbf{W}'\mathbf{W})}\right)^{-\frac{1}{2}} \end{pmatrix},$$

as $d \to \infty$. Therefore the canonical angles $(\theta_1, \ldots, \theta_m)$ between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$ converge to the arccosines of

$$\left(\left(1 + \tau^2/\varphi_m(\mathbf{W}'\mathbf{W})\right)^{-\frac{1}{2}}, \ldots, \left(1 + \tau^2/\varphi_1(\mathbf{W}'\mathbf{W})\right)^{-\frac{1}{2}}\right), \tag{7}$$

as $d \to \infty$. Notice that these canonical angles between subspaces converge to the same limit as in Theorem 3, except that the order is reversed. In particular, the limiting distribution of the largest canonical angle θ_1 is the same as that of the angle between $\hat{u}_m$ and $\mathcal{L}_1^m$, and the smallest canonical angle θ_m corresponds to the angle between the first sample PC direction $\hat{u}_1$ and the population PC space $\mathcal{L}_1^m$.

The limiting distributions of the distances between two subspaces are readily derived by the discussions so far. When using the gap metric,

$$\rho_g(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) \Longrightarrow \left(1 + \varphi_m(\mathbf{W}'\mathbf{W})/\tau^2\right)^{-\frac{1}{2}} \text{ as } d \to \infty.$$

And by using the Euclidean sine metric,

$$\rho_s(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m) \Longrightarrow \left(\sum_{i=1}^{m} \frac{1}{1 + \varphi_i(\mathbf{W}'\mathbf{W})/\tau^2}\right)^{\frac{1}{2}} \text{ as } d \to \infty. \tag{8}$$

Remark 3. The convergence of the canonical angles for the case $\alpha > 1$ has been shown earlier. [3] introduced a notion of *subspace consistency*, where the direction $\hat{u}_i$ may not be consistent to u_i but will tend to lie in $\mathcal{L}_1^m$, i.e. $\text{Angle}(\hat{\mathcal{L}}_i, \mathcal{L}_1^m) \to 0$ as $d \to \infty$, for $i \leq m$. In this case, the canonical angles between $\hat{\mathcal{L}}_i^m$ and $\mathcal{L}_1^m$ and the distances will converge to 0 as d grows. In that sense, the empirical PC space $\hat{\mathcal{L}}_i^m$ is consistent to $\mathcal{L}_1^m$. On the other hand, when $\alpha < 1$, all directions $\hat{u}_i$ tend to behave as if they were from the eigen-decomposition of the identity matrix. Therefore, all angles tend to be $\pi/2$ and the distances will converge to their largest possible values, leading to the strong inconsistency.

We now focus back on the $\alpha = 1$ case, and illustrate the limiting distributions of the canonical angles and the Euclidean sine distance, to see the effect of parameters in the distribution. For simplicity and clear presentation, the results corresponding to $m = 2$ are presented under the Gaussian assumption. Note that the limiting distributions depend on the marginal distributions of the first few principal component scores. Therefore no common distribution is evaluated in the limit.

Let σ^2 $(= \sigma_1^2 + \sigma_2^2)$ denote the (scaled) total variance of the first two principal components. The ratio $\frac{\sigma^2}{\tau^2}$ is understood as a signal to noise ratio, similar to the single spike case. Since the ratio of σ_1^2 and σ_2^2 affects the limiting distributions, we use (λ_1, λ_2) with $\lambda_1 + \lambda_2 = 1$ so that $\sigma^2(\lambda_1, \lambda_2) = (\sigma_1^2, \sigma_2^2)$. Note that for $\mathbf{W}'\mathbf{W} \sim \mathcal{W}_2(n, \mathrm{diag}(\sigma_1^2, \sigma_2^2))$, $\varphi(\mathbf{W}'\mathbf{W})$ has the same law as $\sigma^2\varphi(\mathcal{W}_2(n, \mathrm{diag}(\lambda_1, \lambda_2))$.

The joint limiting distribution of the two canonical angles in (7), also in Theorem 3, is illustrated in Fig. 2, with various values of $(\sigma^2/\tau^2, \lambda_1/\lambda_2)$ and fixed n. Note that for large d, the first canonical angle $\theta_1 \approx \mathrm{Angle}(\hat{u}_{2,d}, \mathcal{L}_1^m)$ and $\theta_2 \approx \mathrm{Angle}(\hat{u}_{1,d}, \mathcal{L}_1^m)$, and that $\theta_1 \geq \theta_2$.

- The diagonal shift of the joint densities in Fig. 2 is driven by different σ^2s with other parameters fixed. Both θ_1 and θ_2 are smaller for larger signal to noise ratios.

- For fixed σ^2/τ^2, several values of the ratio between the first and second variances (λ_1/λ_2) are considered, and the overlay of densities according to different λ_1/λ_2 is illustrated as the vertical shift in Fig. 2. When the variation along the first PC direction is much stronger than that along the second, i.e. when λ_1/λ_2 is large, θ_2 becomes smaller but θ_1 tends to be much larger. In other words, $\hat{u}_1$ is a reasonable estimate of u_1, but $\hat{u}_2$ becomes a poor estimate of u_2.

See (A.4) in the appendix for the probability density function of the canonical angles.

The limiting distribution of the Euclidean sine distance between $\hat{\mathcal{L}}_1^m$ and $\mathcal{L}_1^m$ is also depicted in Fig. 3, again with various values of $(\sigma^2, \lambda_1, \lambda_2)$. It can be checked from the top panel of Fig. 3 that the distance to the optimal subspace is smaller when the signal to noise ratio is larger. The bottom panel illustrates the densities corresponding to different ratios of λ_1/λ_2. The effect of λ_1/λ_2 is relatively small compared to the effect of different σ^2s, unless λ_2 is too small.

4. Geometric representation of the HDLSS data

[15] first showed that the HDLSS data has an interesting *geometric representation* in the limit $d \to \infty$. In particular, for large d, the data tend to appear at vertices of a regular simplex and the variability is contained only in the random rotation of the simplex. In the spike model we consider, this

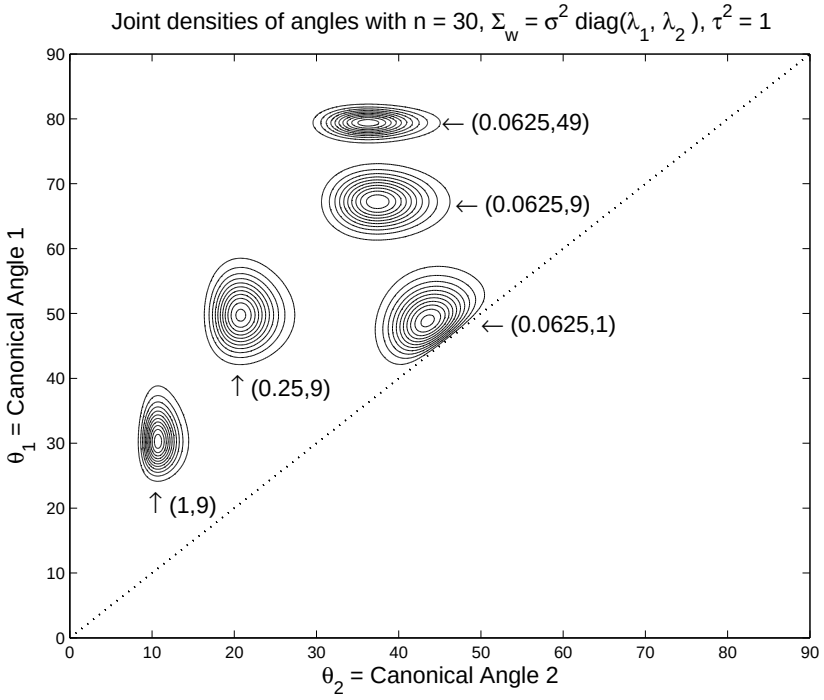

Figure 2: Overlay of contours of densities of canonical angles for the $m = 2$ case, corresponding to different $(\sigma^2/\tau^2, \lambda_1/\lambda_2)$. Larger signal to noise ratios σ^2/τ^2 lead to the diagonal shift of the density function, and both canonical angles will have smaller values. For a fixed σ^2/τ^2, the ratio λ_1/λ_2 between the first and second PC variances is a driver for different distributions, depicted as the vertical shift of the density function.

geometric representation of the HDLSS data holds when $\alpha < 1$, as shown earlier in [3]. The representation in mathematical terms is

$$\|X_i\| = \tau\sqrt{d} + o_p(\sqrt{d}), \|X_i - X_j\| = \tau\sqrt{2d} + o_p(\sqrt{d}), \tag{9}$$

for d dimensional X_i, $i = 1, \ldots, n$. This simplified representation has been used to show some high dimensional limit theory for discriminant analysis, see [2], [24] and [25].

Similar types of representation can be derived in the $\alpha \geq 1$ case. When $\alpha > 1$, where consistency of PC directions happens, we have

$$\|X_i\| / d^{\alpha/2} \Longrightarrow \|Y_i\|, \|X_i - X_j\| / d^{\alpha/2} \Longrightarrow \|Y_i - Y_j\| \tag{10}$$

where $Y_i = (\sigma_1 Z_{1i}, \ldots \sigma_m Z_{mi})'$s are m-dimensional independent random vectors with mean zero and covariance matrix $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. To understand

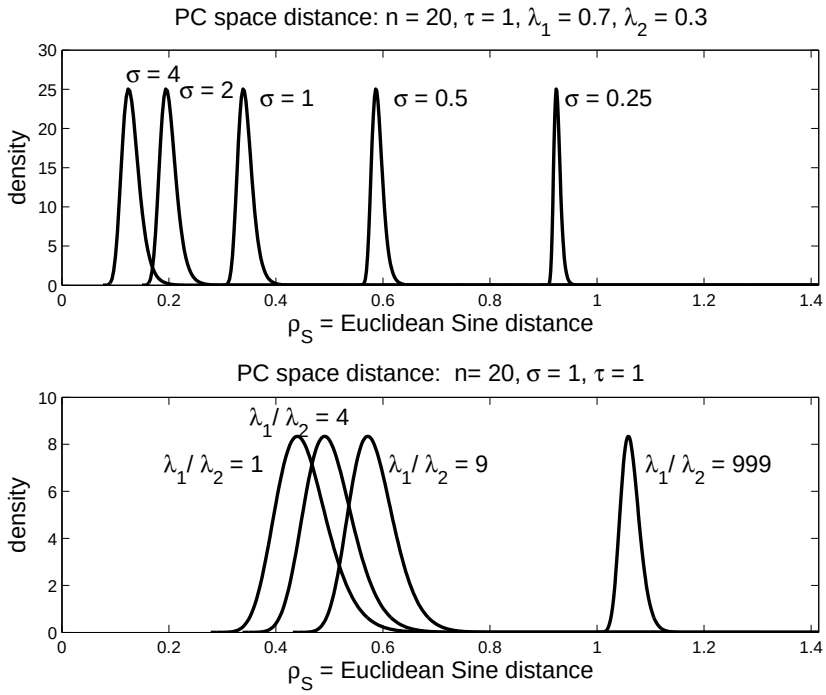

Figure 3: Overlay of densities of the distance between the sample and population PC spaces, measured by ρ_s for the $m = 2$ case. The top panel shows a transition of the density function corresponding to different signal to noise ratios. The bottom panel illustrates the effect of the ratio λ_1/λ_2 between the first two eigenvalues. For a larger signal to noise ratio σ^2/τ^2, and for a smaller value of λ_1/λ_2, the Euclidean sine distance is smaller.

Y_i, let X_i^P be the projection of X_i onto the true PC space $\mathcal{L}_1^m$. It can be checked from $d^{-1/2}X_i^P = \sum_{j=1}^m (\sigma_j Z_{ji})u_j$ that Y_i is the vector of loadings of the scaled X_i^P in the first m principal component coordinates. When $\alpha = 1$, a deterministic term is added:

$$\|X_i\|^2/d^1 \Longrightarrow \|Y_i\|^2 + \tau^2, \|X_i - X_j\|^2/d \Longrightarrow \|Y_i - Y_j\|^2 + 2\tau^2, \qquad (11)$$

These results can be understood geometrically, as summarized and discussed in the following;

$\alpha > 1$**:** The variability of samples is restricted to the true PC space $\mathcal{L}_1^m$ for large d, which coincides with the notion of subspace consistency discussed in Remark 3. The d-dimensional probability distribution degenerates to the m-dimensional subspace $\mathcal{L}_1^m$.

$\alpha = 1$**:** (11) is understood with a help of Pythagorean theorem, that is, the norm of X_i is asymptotically decomposed into orthogonal random and deterministic parts. Thus, data tend to be $\tau\sqrt{d}$ away from $\mathcal{L}_1^m$, and X_is projected on $\mathcal{L}_1^{m\perp} = \text{span}\{u_{m+1}, \ldots, u_d\}$, the orthogonal complement of $\mathcal{L}_1^m$, will follow the representation similar to the $\alpha < 1$ case.

$\alpha < 1$**:** The geometric representation (9) holds.

Note that the case $\alpha = 1$ smoothly bridges the others.

An example elucidating these ideas is shown in Fig. 4. Each panel shows 3d scatterplots of 10 different samples (shown as different symbols) of $n = 3$ Gaussian random vectors in dimensions $d = 3, 30, 3000$ (shown in respective columns of Fig. 4. In the spiked model, we take $\sigma = \tau = 1$ and $m = 1$ for simplicity and investigate three different orders of magnitude $\alpha = \frac{1}{2}, 1, 2$ of the first eigenvalue $\lambda_1 = d^\alpha$. For each pair of (d, α), each sample X_i is projected onto the first true PC direction u_1, shown as the vertical axis. In the orthogonal $d - 1$ dimensional subspace, the 2-dimensional hyperplane that is generated by the data is found, and the data are projected onto that. Within the hyperplane, variation due to rotation is removed by optimally rotating the data onto edges of a regular triangle by a Procrustes method, to give the horizontal axes in each part of Fig. 4. These axes are scaled by dividing by $\max(d^\alpha, d)$ and the 10 samples give an impression of the various types of convergence as a function of d, for each α.

The asymptotic geometric representations summarized above are confirmed by the figure. For $\alpha = \frac{1}{2}$, it is expected from (9) that the data are close to the vertices of the regular triangle, with edge length $\sqrt{2d}$. The vertices of the triangle (in the horizontal plane) with vertical rays (representing u_1, the first PC direction) are shown as the dashed lines in the first two rows of Fig. 4. Note that for $d = 3$, in the top row, the points appear to be quite random, but for $d = 30$, there already is reasonable convergence to the vertices with notable variation along u_1. The case $d = 3000$ shows more rigid representation with much less variation along u_1. On the other hand, for the case $\alpha = 2$ in the last row of Fig. 4, most of the variation in the data is found along u_1, shown as the vertical dotted line, and the variation perpendicular to u_1 becomes negligible as d grows, which confirms the degeneracy to $\mathcal{L}_1^1$ in (10). From these examples, conditions for consistency and strong inconsistency can be checked heuristically. The sample eigenvector $\hat{u}_1$ is consistent with u_1 when $\alpha > 1$, since the variation along u_1 is so strong that $\hat{u}_1$ should

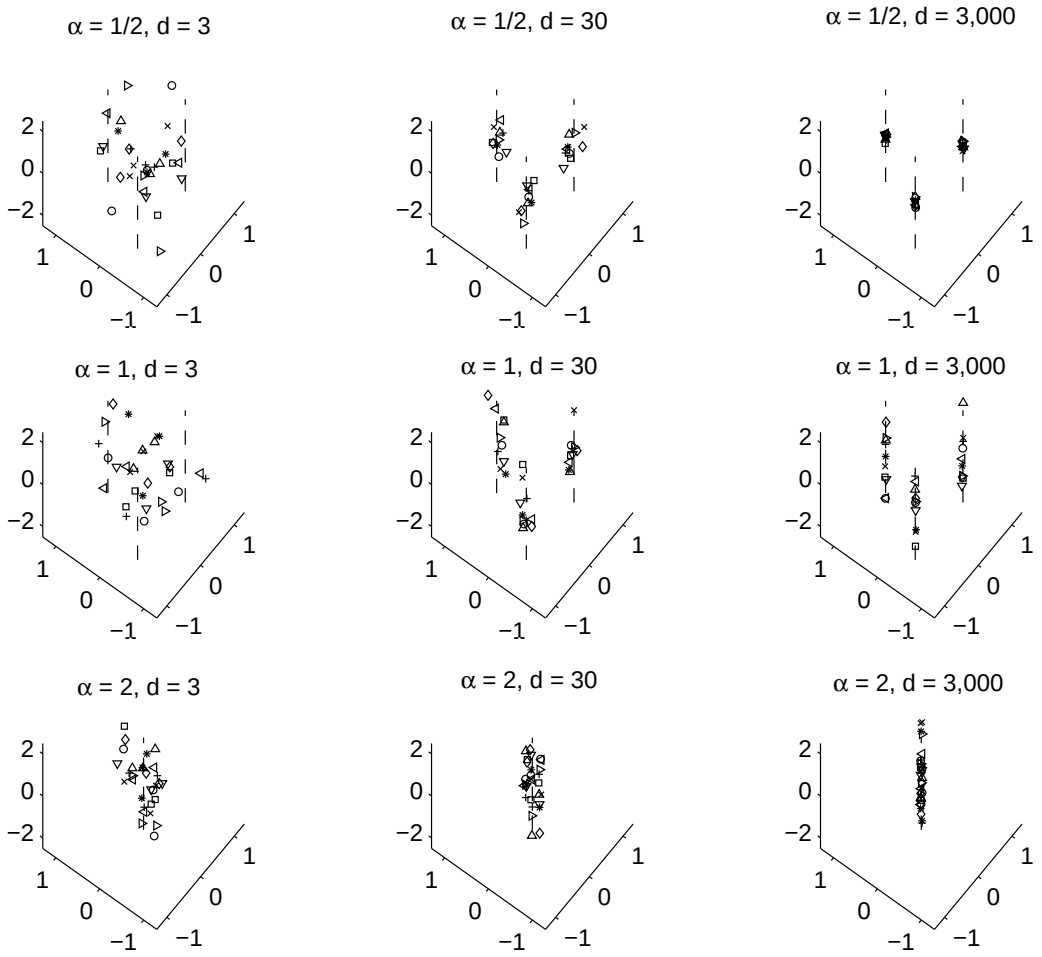

Figure 4: Gaussian toy example, showing the geometric representations of HDLSS data, with $n = 3$, for three different choices of $\alpha = 1/2, 1, 2$ of the spiked model and increasing dimensions $d = 3, 30, 3000$. For $\alpha \neq 1$, data converge to vertices of a regular 3-simplex (case $\alpha < 1$) or to the first PC direction (case $\alpha > 1$). When $\alpha = 1$, data are decomposed into the deterministic part on the horizontal axes and the random part along the vertical axis.

be close to that. $\hat{u}_1$ is inconsistent with u_1 when $\alpha < 1$, since the variation along u_1 becomes negligible so that $\hat{u}_1$ will not be near u_1.

For the $\alpha = 1$ case, in the middle row of Fig. 4, it is expected from (11) that each data point will be asymptotically decomposed into a random and a deterministic part. This is confirmed by the scatterplots, where the order

of variance along u_1 remains comparable to that of horizontal components, as d grows. The convergence to the vertices is noticeable even for $d = 30$, which becomes stronger for larger d, while the randomness along u_1 remains for large d. Also observe that the distance from each X_i to the space spanned by u_1 becomes deterministic for large d, supporting the first part of (11).

Appendix A. Derivation of the density functions

The probability density functions of the limiting distributions in (7) and (8) will be derived for the case $m = 2$, under the normal assumption. The argument is readily generalized to all m, but when non-normal distribution is assumed, such derivation is much challenging.

We first recall some necessary notions for treating the Wishart matrix $\mathbf{W}'\mathbf{W}$ and eigen-decompositions. Most of the results are adopted from [18]. Let $A \sim W_m(n, \Sigma_m)$ and denote its eigen-decomposition as $A = HLH'$ with $L = \mathrm{diag}(l_1, \ldots, l_m)$. Assume Σ_m is positive definite and $n \geq m$ so that $l_1 > l_2 > \cdots > l_m > 0$ with probability 1. Denote $O(m) = \{H_{m\times m} : H'H = I_m\}$ for the set of orthonormal $m \times m$ matrices and (dH) for $H \in O(m)$ as the differential form representing the uniform probability measure on $O(m)$. The multivariate gamma function is defined as

$$\Gamma_m(a) = \pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma\left(a - \frac{1}{2}(i-1)\right),$$

where $\Gamma(\cdot)$ is the usual gamma function. For $H \equiv [\mathbf{h}_1, \ldots, \mathbf{h}_m] \in O(m)$,

$$(dH) \equiv \frac{1}{\mathrm{Vol}(O(m))}(H'dH) = \frac{\Gamma_m(\frac{m}{2})}{2^m \pi^{m^2/2}}(H'dH), \tag{A.1}$$

where $(H'dH) \equiv \prod_{i>j}^{m} \mathbf{h}_i' d\mathbf{h}_j$.

We are now ready to state the density function of $\varphi(\mathbf{W}'\mathbf{W})$ for $m = 2$. Note that under the Gaussian assumption $\mathbf{W}'\mathbf{W}$ is the 2×2 Wishart matrix with degree of freedom n and covariance matrix $\Sigma_W = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$. For simplicity, write $(L_1, L_2) = \varphi(\mathbf{W}'\mathbf{W})$, and $L = \mathrm{diag}(L_1, L_2)$. Then the joint density function of L_1 and L_2 is given by e.g. Theorem 3.2.18 of [18] with $m = 2$, and

$$f_L(l_1, l_2) = \frac{\pi 2^{-n}(\sigma_1^2\sigma_2^2)^{-\frac{n}{2}}}{\Gamma_2(\frac{n}{2})}(l_1 l_2)^{\frac{n-3}{2}}(l_1 - l_2) \int_{O(2)} \exp\left(\mathrm{trace}\left(-\frac{1}{2}\Sigma_W^{-1} HLH'\right)\right)(dH). \tag{A.2}$$

The integral can not be solved analytically but can be simplified by using the special orthogonal group $SO(2) = \{H \in O(2) : \det H = 1\}$. We can parameterize $H \in SO(2)$ as

$$H = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = [\mathbf{h}_1, \mathbf{h}_2] \quad (0 < \theta \leq 2\pi).$$

Then $(H'dH) = \mathbf{h}_2' d\mathbf{h}_1 = d\theta$. Moreover the integral in (A.2) over $O(2)$ is twice as large as the integral over $SO(2)$. This fact and the definition (A.1) together with the parametrization above give

$$\begin{aligned} \int_{O(2)} & \exp\left(\text{trace}\left(-\tfrac{1}{2}\Sigma_W^{-1} HLH'\right)\right)(dH) \\ &= \tfrac{1}{2\pi}\int_{SO(2)} \exp\left(\text{trace}\left(-\tfrac{1}{2}\Sigma_W^{-1} HLH'\right)\right)(H'dH) \\ &= \tfrac{1}{2\pi}\int_0^{2\pi} \exp\left(-\tfrac{1}{2}\left[A\cos^2\theta + B\sin^2\theta\right]\right)d\theta \\ &= \tfrac{1}{2\pi} e^{-\frac{A+B}{4}} \int_0^{2\pi} \exp\left(\tfrac{1}{4}(B-A)\cos t\right)dt, \\ &= e^{-\frac{A+B}{4}} I_0\left(\tfrac{1}{4}(B-A)\right) \end{aligned}$$

where $A = \frac{l_1}{\sigma_1^2} + \frac{l_2}{\sigma_2^2}$, $B = \frac{l_2}{\sigma_1^2} + \frac{l_1}{\sigma_2^2}$ and

$$I_0(x) = \frac{1}{2\pi}\int_0^{2\pi} \exp(x\cos t)\,dt$$

is the modified Bessel function of the first kind. Note that the integral can also be represented by the hypergeometric function of matrix arguments (see Section 7.3 of [18]). We chose to use $I_0(x)$ since it is numerically more stable than the hypergeometric function. Then (A.2) becomes

$$f_L(l_1, l_2) = \frac{\pi 2^{-n}(\sigma_1^2\sigma_2^2)^{-\frac{n}{2}}}{\Gamma_2(\frac{n}{2})}(l_1 l_2)^{\frac{n-3}{2}}(l_1 - l_2)e^{-\frac{A+B}{4}} I_0\left(\frac{1}{4}(B-A)\right). \quad \text{(A.3)}$$

Now the distribution of the canonical angles (in (7) and Theorem 3) is obtained by applying the change of variable on the density (A.3). Let Y_1, Y_2 be the two canonical angles, in the reverse order. Then from

$$\begin{aligned} (Y_1, Y_2) &= \left(\cos^{-1}\{(1+\tau^2/L_1)^{-1/2}\}, \cos^{-1}\{(1+\tau^2/L_2)^{-1/2}\}\right) \\ &= \left(tan^{-1}(\sqrt{\tau^2/L_1}), tan^{-1}(\sqrt{\tau^2/L_2})\right), \end{aligned}$$

the joint density function of Y_1, Y_2 becomes

$$f_{Y_1,Y_2}(y_1, y_2) = f_L(\tau^2 \cot^2 y_1, \tau^2 \cot^2 y_2) \cdot (2\tau^2)^2 \frac{\cos y_1}{\sin^3 y_1} \frac{\cos y_2}{\sin^3 y_2} \tag{A.4}$$

on $0 < y_1 < y_2 < \frac{\pi}{2}$.

The limiting distribution of the distances between the empirical and population principal subspace, measured by the Euclidean sine metric, $\rho_s(\hat{\mathcal{L}}_1^m, \mathcal{L}_1^m)$ in (8) is obtained as follows. Let

$$Z_1 = \sqrt{\frac{\tau^2}{\tau^2 + L_1} + \frac{\tau^2}{\tau^2 + L_2}}, Z_2 = \frac{\tau^2}{\tau^2 + L_2}$$

so that

$$L_1 = \frac{\tau^2}{Z_1^2 - Z_2} - \tau^2, L_2 = \frac{\tau^2}{Z_2} - \tau^2.$$

The distribution of Z_1 is the limiting distribution of interest. Note that the eigenvalues $(L_1, L_2) \sim f_L$ must satisfy $0 < L_2 < L_1 < \infty$. This leads to the support for the joint distribution of (Z_1, Z_2):

$$D = \{Z_1, Z_2 \in \mathbb{R} : Z_2 < Z_1^2, Z_1^2 < 2Z_2, Z_2 < 1\}.$$

By the change of variable on f_L (A.3), we get

$$f_{Z_1,Z_2}(z_1, z_2) = f_L(\frac{\tau^2}{z_1^2 - z_2} - \tau^2, \frac{\tau^2}{z_2} - \tau^2) \cdot 2z_1 \frac{1}{\tau^4} \left(\frac{\tau^2}{z_2} \frac{\tau^2}{z_1^2 - z_2} \right)^2 1_{(z_1,z_2 \in D)}. \tag{A.5}$$

The marginal density of Z_1 can be obtained by numerical integration of f_{Z_1,Z_2}. The support of the density is then $(0, \sqrt{2})$.

References

[1] A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. USA 98(24):13790-5.

[2] J. Ahn, J. S. Marron, K. M. Muller, Y.-Y. Chi, The high-dimension, low-sample-size geometric representation holds under mild conditions, Biometrika 94 (3) (2007) 760–766.

[3] S. Jung, J. S. Marron, Pca consistency in high dimension, low sample size context, Ann. Statist. 37 (6B) (2009) 4104–4130.

[4] K. Yata, M. Aoshima, PCA consistency for non-Gaussian data in high dimension, low sample size context, Comm. Statist. Theory Methods 38 (16-17) (2009) 2634–2652.

[5] K. Yata, M. Aoshima, Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, to appear in Journal of Multivariate Analysis.

[6] G. Casella, J. T. Hwang, Limit expressions for the risk of James-Stein estimators, Canad. J. Statist. 10 (4) (1982) 305–309.
URL http://dx.doi.org/10.2307/3556196

[7] F. Pesarin, L. Salmaso, Finite-sample consistency of combination-based permutation tests with application to repeated measures designs, J. Nonparametr. Stat. 22 (5-6) (2010) 669–684.
URL http://dx.doi.org/10.1080/10485250902807407

[8] F. Pesarin, L. Salmaso, Permutation tests for complex data : theory, applications and software, Chichester, U.K.: Wiley, 2010.

[9] I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist. 29 (2) (2001) 295–327.

[10] J. Baik, J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, J. Multivariate Anal. 97 (6) (2006) 1382–1408.

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

[12] B. Nadler, Finite sample approximation results for principal component analysis: a matrix perturbation approach, Ann. Statist. 36 (6) (2008) 2791–2817.
URL http://dx.doi.org/10.1214/08-AOS618

[13] S. Lee, F. Zou, F. A. Wright, Convergence and prediction of principal component scores in high-dimensional settings, Ann. Statist. 38 (6) (2010) 3605–3629. doi:10.1214/10-AOS821.
URL http://dx.doi.org/10.1214/10-AOS821

[14] N. El Karoui, Spectrum estimation for large dimensional covariance matrices using random matrix theory, Ann. Statist. 36 (6) (2008) 2757–2790.
URL http://dx.doi.org/10.1214/07-AOS581

[15] P. Hall, J. S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (3) (2005) 427–444.

[16] A. N. Kolmogorov, Y. A. Rozanov, On strong mixing conditions for stationary gaussian processes, Theory Probab. Appl. 5 (2) (1960) 204–208.

[17] R. C. Bradley, Basic properties of strong mixing conditions. A survey and some open questions, Probab. Surv. 2 (2005) 107–144 (electronic), update of, and a supplement to, the 1986 original.

[18] R. J. Muirhead, Aspects of multivariate statistical theory, John Wiley & Sons Inc., New York, 1982, wiley Series in Probability and Mathematical Statistics.

[19] Z. Bai, J. W. Silverstein, Spectral analysis of large dimensional random matrices, 2nd Edition, Springer Series in Statistics, Springer, New York, 2010. doi:10.1007/978-1-4419-0661-8.
URL http://dx.doi.org/10.1007/978-1-4419-0661-8

[20] A. F. Acker, Absolute continuity of eigenvectors of time-varying operators, Proc. Amer. Math. Soc. 42 (1974) 198–201.

[21] G. H. Golub, C. F. Van Loan, Matrix computations, 3rd Edition, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996.

[22] G. W. Stewart, J. G. Sun, Matrix perturbation theory, Computer Science and Scientific Computing, Academic Press Inc., Boston, MA, 1990.

[23] T. L. Gaydos, Data representation and basis selection to understand variation of function valued traits, Ph.D. thesis, University of North Carolina at Chapel Hill (2008).

[24] X. Qiao, H. H. Zhang, Y. Liu, M. Todd, J. S. Marron, Weighted Distance Weighted Discrimination and its asymptotic properties, J. Amer. Statist. Assoc. 105 (489) (2010) 401–414.

[25] H. Huang, Y. Liu, J. S. Marron, Bi-Directional Discrimination with application to data visualization, submitted to J. Amer. Statist. Assoc.