Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 2, 255–274
DOI: 10.1214/12-BJPS205

PCA and eigen-inference for a spiked covariance model with largest eigenvalues of same asymptotic order

Addy Bolivar-Cime and Victor Perez-Abreu

Centro de Investigacion en Matematicas

Abstract. In this paper, we work under the setting of data with high dimension d greater than the sample size n (HDLSS). We study asymptotics of the first $p \geq 2$ sample eigenvalues and their corresponding eigenvectors under a spiked covariance model for which its first p largest population eigenvalues have the same asymptotic order of magnitude as d tends to infinity and the rest are constant. We get the asymptotic joint distribution of the nonzero sample eigenvalues when $d \to \infty$ and the sample size n is fixed. We then prove that the p largest sample eigenvalues increase jointly at the same speed as their population counterpart, in the sense that the vector of ratios of the sample and population eigenvalues converges to a multivariate distribution when $d \to \infty$ and n is fixed, and to the vector of ones when both $d, n \to \infty$ and $d \gg n$. We also show the subspace consistency of the corresponding sample eigenvectors when d goes to infinity and n is fixed. Furthermore, using the asymptotic joint distribution of the sample eigenvalues we study some inference problems for the spiked covariance model and propose hypothesis tests for a particular case of this model and confidence intervals for the p largest eigenvalues. A simulation is performed to assess the behavior of the proposed statistical methodologies.

1 Introduction

There is an increasing current interest in the statistical analysis of data arising in problems of genomics, medical image analysis, climatology, finance and functional data analysis, where one frequently observes multivariate data with high dimension greater than the sample size; see, for example, Hall et al. (2005) and Johnstone (2001). An important problem for this kind of data is the inference about the eigen-structure of the population covariance matrix. When the data dimension is greater than the sample size, Principal Component Analysis (PCA) often fails to estimate the population eigenvalues and eigenvectors since the sample covariance matrix is not a good approximation to the population covariance matrix. As pointed out in Johnstone (2001), one often observes one or a small number of large sample eigenvalues well separated from the rest. This case is of special interest, and is called the *spiked covariance model*.

Key words and phrases. Principal Component Analysis, spiked covariance model, eigen-inference, hypothesis test, confidence interval, high dimensional data, HDLSS.

Received February 2012; accepted July 2012.

More specifically, suppose $X = [X_1, X_2, \dots, X_n]$ is a $d \times n$ data matrix where the sample $X_j = (x_{1j}, \dots, x_{dj})^\top$, $j = 1, 2, \dots, n$ are independent and identically distributed random vectors with mean zero and unknown covariance matrix Σ, and X has rank n with probability one (it is not assumed that the X_j's have a multivariate Gaussian distribution). The spiked covariance model considers a covariance matrix of the type

$$\Sigma = O \Lambda O^\top \qquad \text{where } \Lambda = \operatorname{diag}(\tau_1, \tau_2, \dots, \tau_p, \sigma, \dots, \sigma), \tag{1.1}$$

with $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_p > \sigma > 0$, for some $1 \leq p < d$, and O is a $d \times d$ orthogonal matrix.

We assume that the spiked covariance model is such that $\tau_i = \tau_i(d)$ and

$$\frac{\tau_i}{d^{\alpha_i}} \longrightarrow c_i \qquad \text{as } d \to \infty, \tag{1.2}$$

where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p > 1$ and $c_i > 0$, $i = 1, 2, \dots, p$. We say that the p largest population eigenvalues of the spiked covariance model (1.1) have the *same asymptotic order of magnitude in d* if $\alpha_1 = \alpha_2 = \cdots = \alpha_p = \alpha > 1$; and we say that the p largest population eigenvalues have *different asymptotic order of magnitude in d* if $\alpha_1 > \alpha_2 > \cdots > \alpha_p > 1$. In this paper, we focus our attention on the spiked covariance models where the $p \geq 2$ largest eigenvalues have same asymptotic order of magnitude in d.

There are three different contexts in which the study of PCA for the spiked covariance model arises: (i) the Classical case, (ii) the Random Matrix Theory (RMT) context and (iii) the High-Dimensional, Low Sample Size (HDLSS) context. Each context depends on the particular data analytic setting, the modelling features and the way the corresponding asymptotics are considered with respect to the data dimension d and the sample size n.

In the well-known classical case, one considers d fixed and n goes to infinity. In the RMT situation one considers d and n go to infinity simultaneously, in the sense that $d/n \to \gamma$, where $\gamma \in [0, \infty]$; see Bai and Yang (2008), Baik and Silverstein (2006). In this context, the population eigenvalues of the covariance matrix Σ do not depend on d and the basic analytic tool is the so-called Marchenko–Pastur theorem; see Baik and Silverstein (2006).

On the other hand, in the so-called HDLSS context the asymptotic results are developed by letting the data dimension $d \to \infty$ while keeping fixed the sample size n. Some references on this framework are Ahn et al. (2007), Hall et al. (2005), Jung and Marron (2009), Jung et al. (2012) and Yata and Aoshima (2009). One can also consider in this framework the case of letting first the data dimension $d \to \infty$ while keeping fixed the sample size n and in a second step, letting $n \to \infty$; see Ahn et al. (2007), Jung and Marron (2009) and Jung et al. (2012). In other words, d, n tend to infinity successively with d increasing at a much faster rate than n, that is, $d \gg n$. In contrast to the RMT context, in the HDLSS context it may be assumed that the p largest population eigenvalues of the covariance Σ depend also on the data dimension d.

Under a sample Gaussian assumption on X_j, Ahn et al. (2007) show for $p = 1$ and $\Sigma = \text{diag}(d^{\alpha}, 1, \ldots, 1)$ with $\alpha > 1$, that the largest sample eigenvalue increases at the same speed as its population eigenvalue, in the sense that its ratio converges to the distribution $\mathcal{X}_n^2/n$ when $d \to \infty$ and n is fixed, where $\mathcal{X}_n^2$ is a r.v. with chi-square distribution with n degrees of freedom; and converges to one when $d, n \to \infty$ and $d \gg n$. Moreover, they show that the first sample eigenvector is consistent when $d, n \to \infty$ and $d \gg n$.

In the Gaussian case and when the $p \geq 2$ largest sample eigenvalues have different asymptotic order of magnitude, it follows from the results in Jung and Marron (2009) that if $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \cdots \geq \widehat{\tau}_p$ are the p largest sample eigenvalues then

$$\frac{\widehat{\tau}_i}{\tau_i} \xrightarrow{w} \frac{\mathcal{X}_n^2}{n} \qquad \text{as } d \to \infty \tag{1.3}$$

for $i = 1, 2, \ldots, p$. Since $\mathcal{X}_n^2/n \overset{w}{\to} 1$ as $n \to \infty$, in this case, we have that

$$\frac{\widehat{\tau}_i}{\tau_i} \xrightarrow{w} 1 \qquad \text{as } d \to \infty, n \to \infty \tag{1.4}$$

for $i = 1, 2, \ldots, p$, where the limits are applied successively. Thus, the p largest sample eigenvalues increase at the same speed as their population eigenvalues. We give multivariate extensions of these asymptotic results for the non-Gaussian case and when the p largest population eigenvalues have same asymptotic order of magnitude. The work of Jung and Marron (2009) does not address this asymptotic behavior of the p largest sample eigenvalues in those cases, even when they have results for the case of same asymptotic order of magnitude and considering non-Gaussian distributions, only the marginal convergence in distribution of these sample eigenvalues is shown in their Lemma 1 considering a ρ-mixing condition. Moreover, they show the subspace consistency and the consistency of the corresponding sample eigenvectors in the case of same and different asymptotic order of magnitude, respectively. We do not consider ρ-mixing conditions in our assumptions.

Yata and Aoshima (2009) study the asymptotic behavior of the sample eigenvalues and their corresponding eigenvectors for a spiked covariance model where the p largest population eigenvalues may have same asymptotic order of magnitude in d. They prove a result similar to (1.4) with different hypotheses from the considered in the present paper, and without assuming either a Gaussian distributions or a ρ-mixing condition. However, the result (1.4) alone does not contribute to do inference for the p largest population eigenvalues, and therefore it is important to have multivariate extensions of (1.3) as we do in the present paper. Yata and Aoshima (2009) show a kind of central limit theorem for the ratios of the sample and population eigenvalues assuming that the first p largest population eigenvalues are different. In the present paper we prove, under a Gaussian assumption, a kind of multivariate central limit theorem for the vector of these ratios, where the first p largest population eigenvalues have same asymptotic order, and in particular they may be the same.

Under the assumption of same asymptotic order of magnitude in d, we get the asymptotic joint distribution of the nonzero sample eigenvalues, which implies a multivariate extension of (1.3) when $d \to \infty$ and keeping n fixed. We then obtain that the p largest sample eigenvalues increase jointly at the same speed as their population counterpart, in the sense that the vector of ratios of the sample and population eigenvalues converges to a multivariate distribution when $d \to \infty$ and n is fixed, and to the vector of ones when both $d, n \to \infty$ and $d \gg n$. In the work of Jung and Marron (2009), only the marginal convergence of the sample eigenvalues is taken into account. The advantage of considering the joint convergence in distribution of the nonzero sample eigenvalues is that it is possible to derive asymptotic results for functions of them. Furthermore, these asymptotic results are useful to consider some inference problems, as those considered in the present paper. We also show the subspace consistency of the first p sample eigenvectors for our spiked covariance model.

As an important contribution of this article, we also develop some results behind hypothesis tests and confidence intervals in the two asymptotic settings of the HDLSS context. Namely, we apply the above results under a Gaussian assumption, to consider hypothesis tests for our spiked covariance model and confidence intervals for the p largest population eigenvalues. It is seen that some classical statistics are also useful when d goes to infinity and n is fixed, and when d, n go to infinity and $d \gg n$.

The organization of the paper is as follows. In Section 2, we study the asymptotic behavior of the p largest sample eigenvalues in two situations: when $d \to \infty$ and n is fixed; and when first the dimension $d \to \infty$ and then subsequently $n \to \infty$. In Section 3, the subspace consistency of the corresponding eigenvectors is considered. In Section 4, we consider some eigen-inference problems in the case when the sample is taken from a multivariate Gaussian distribution and therefore the sample covariance matrix follows a Wishart distribution. In this section, hypothesis tests for a particular case of our spiked covariance model and confidence interval for the p largest eigenvalues are proposed in the HDLSS context. Finally, in Section 5 a simulation study is conducted to show the good performance of the statistical methodologies proposed in Section 4.

2 Asymptotics of sample eigenvalues

In this section, we consider the spiked covariance model (1.1) where the p largest population eigenvalues have the same asymptotic order of magnitude as d goes to infinity. We consider two situations of the HDLSS framework. We first deal with the case when $d \to \infty$ and n is kept fixed; then we consider the case when $d \to \infty$ first and in a second step $n \to \infty$.

2.1 Sample size n fixed and data dimension $d \to \infty$

We consider the following assumptions for the matrix X:

(a) Let $Z = \Lambda^{-1/2} O^\top X$ and assume that its entries have uniformly bounded fourth moments with respect to d, in the sense that for each $n = p+1, p+2, \ldots$ there exists $K_n > 0$ such that $E(z_{ij}^4) \leq K_n$ for all $i = 1, 2, \ldots, d$, $j = 1, 2, \ldots, n$ and $d = n+1, n+2, \ldots$.

(b) Let Z_i be the i-th row vector of Z and define $\widetilde{Z}_p = [Z_1^\top, \ldots, Z_p^\top]^\top$. Assume that $\widetilde{Z}_p$ converges in distribution to some $p \times n$ matrix Y_n as $d \to \infty$, which has rank p with probability one.

We observe that the columns of Z are independent and identically distributed random vectors with mean zero and identity covariance matrix. These assumptions do not cover all random matrices but are still very general and include some interesting settings. In the case when the independent columns of X have the Gaussian distribution $N_d(0, \Sigma)$, assumptions (a) and (b) are automatically satisfied and the random matrix $W_1 = Z_i^\top Z_i$'s have a Wishart distribution with one degree of freedom. The assumption (b) is also satisfied in the case when the $\widetilde{Z}_p$'s have a stationary distribution in d, that is the distribution of Y_n is the distribution of the $\widetilde{Z}_p$'s for all $d > n$. Assumption (b) also holds in the case considered by Jung and Marron (2009) where a ρ-mixing condition is assumed; see proof of their Lemma 1. We do not assume ρ-mixing conditions, as Yata and Aoshima (2009) mention this kind of conditions are too strict and have obvious shortcoming, since it is needed an ordering of the variables and in some settings as microarray data there is not a natural ordering of the gene expressions and there exists a clear dependence between them.

Denote by $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \cdots \geq \widehat{\tau}_n$ the nonzero sample eigenvalues of the sample covariance matrix $S = n^{-1} X X^\top$. The first result is an analogue of Lemma 1 of Jung and Marron (2009). In the next theorem, we observe the joint convergence in distribution of the vector of nonzero sample eigenvalues when dimension d goes to infinity and the sample size n is fixed. Note that Lemma 1 of Jung and Marron (2009) states only the convergence in distribution of each component of this vector (marginal convergence).

Theorem 2.1. *Suppose that the unknown covariance matrix Σ of the columns of X is given by the spiked covariance model* (1.1), *with $p < n < d$ and where $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_p$ have the same asymptotic order of magnitude in d. Consider the assumptions* (*a*) *and* (*b*) *for the matrix X. Then when n is fixed*

$$d^{-\alpha}(\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_n)^\top \xrightarrow{w} n^{-1}(\ell_1, \ell_2, \ldots, \ell_p, 0, \ldots, 0)^\top$$

as $d \to \infty$, where $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p > 0$ are the eigenvalues of the random matrix $\widetilde{U}_0 = \mathcal{C}_p^{1/2} Y_n Y_n^\top \mathcal{C}_p^{1/2}$, with $\mathcal{C}_p = \mathrm{diag}(c_1, c_2, \ldots, c_p)$.

Proof. The proof is based on the ideas of Section 4.2 of Ahn et al. (2007) where the case $p = 1$ was considered. We have $\Sigma = O \Lambda O^\top$ where $\Lambda = \mathrm{diag}(\tau_1, \ldots, \tau_p, \sigma, \ldots, \sigma)$ is the diagonal matrix of the eigenvalues of Σ and the corresponding eigenvectors are the column vectors of the matrix O. The sample covariance matrix S and the dual sample covariance matrix $S_D = n^{-1} X^\top X$ have the same nonzero eigenvalues. Moreover, the following representation holds

$$nS_D = Z^\top \Lambda Z = \sum_{i=1}^{d} \lambda_i W_i = \sum_{i=1}^{p} \tau_i W_i + \sigma \sum_{i=p+1}^{d} W_i,$$

where $W_i = Z_i^\top Z_i$ and Z_i, $i = 1, 2, \ldots, d$, are the row vectors of Z. Hence

$$d^{-\alpha} n S_D = d^{-\alpha} \sum_{i=1}^{p} \tau_i W_i + d^{-\alpha} \sigma \sum_{i=p+1}^{d} W_i = U + \sigma d^{-\alpha} V, \tag{2.1}$$

where $U = \sum_{i=1}^{p} d^{-\alpha} \tau_i W_i$ and $V = \sum_{i=p+1}^{d} W_i$.

Let $\widetilde{\tau}_p = \mathrm{diag}(\tau_1, \ldots, \tau_p)$ and $\mathcal{C}_p = \mathrm{diag}(c_1, c_2, \ldots, c_p)$. Note that $U = \widetilde{Z}_p^\top (d^{-\alpha} \widetilde{\tau}_p) \widetilde{Z}_p$ converges in distribution to $\widetilde{U} = Y_n^\top \mathcal{C}_p Y_n$ as $d \to \infty$. On the other hand, we can show that $d^{-\alpha} V$ converges to the zero matrix in distribution as $d \to \infty$. In order to see that, consider the norm $\|A\| = [\mathrm{tr}(A^\top A)]^{1/2}$ for the $n \times n$ matrix A. By the Markov's inequality, we have that for any $\varepsilon > 0$

$$P\big(\|d^{-\alpha} V\| > \varepsilon\big) = P\big(\|d^{-\alpha} V\|^2 > \varepsilon^2\big) \leq \big(d^{2\alpha} \varepsilon^2\big)^{-1} E\big(\|V\|^2\big).$$

Using properties of the trace and the fact that the W_i's are symmetric, it can be seen that the right side of the last inequality is equal to

$$\big(d^{2\alpha} \varepsilon^2\big)^{-1} \sum_{i=p+1}^{d} \sum_{j=p+1}^{d} E\big[(Z_i Z_j^\top)^2\big] = \big(d^{2\alpha} \varepsilon^2\big)^{-1} \sum_{i=p+1}^{d} \sum_{j=p+1}^{d} \sum_{k=1}^{n} \sum_{r=1}^{n} E\big(z_{ik}^2 z_{jr}^2\big).$$

Since there exist $K_n > 0$ such that $E(z_{ij}^4) \leq K_n$ for all i, j, and by the Holder's inequality $E(z_{ik}^2 z_{jr}^2) \leq E(z_{ik}^4)^{1/2} E(z_{jr}^4)^{1/2}$, we have that the right side of the last equation is less than or equal to $(d^{2\alpha} \varepsilon^2)^{-1} (d - p)^2 n^2 K_n$, then

$$P\big(\|d^{-\alpha} V\| > \varepsilon\big) \leq \frac{(d-p)^2 n^2 K_n}{d^{2\alpha} \varepsilon^2} = \left(\frac{d-p}{d}\right)^2 \left(\frac{1}{d^{\alpha-1}}\right)^2 \frac{n^2 K_n}{\varepsilon^2} \tag{2.2}$$

and the right side of the inequality tends to zero when $d \to \infty$ because $\alpha > 1$. Thus the second term in the right-hand side of (2.1) goes to the zero matrix in probability, and therefore in distribution, as d increases. Hence,

$$d^{-\alpha} n S_D \xrightarrow{w} \widetilde{U} \qquad \text{as } d \to \infty.$$

Then the vector of the roots of the characteristic polynomial of $d^{-\alpha} n S_D$ converge in distribution to the vector of the roots of the characteristic polynomial of $\widetilde{U}$ as $d \to \infty$.

Since $\widetilde{U} = Y_n^\top \mathcal{C}_p Y_n$, the nonzero eigenvalues of $\widetilde{U}$ are the p nonzero eigenvalues $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$ of $\widetilde{U}_0 = (\mathcal{C}_p^{1/2} Y_n)(\mathcal{C}_p^{1/2} Y_n)^\top = \mathcal{C}_p^{1/2} Y_n Y_n^\top \mathcal{C}_p^{1/2}$. Hence, if $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \cdots \geq \widehat{\tau}_n$ are the nonzero eigenvalues of S_D, or of S, we have

$$d^{-\alpha} n (\widehat{\tau}_1, \widehat{\tau}_2, \dots, \widehat{\tau}_n)^\top \xrightarrow{w} (\ell_1, \dots, \ell_p, 0, \dots, 0)^\top$$

when $d \to \infty$. □

The following consequence of Theorem 2.1 shows the usefulness of the joint convergence in distribution of the sample eigenvalues when the dimension d goes to infinity. The result is a multivariate extension of (1.3). It gives the joint convergence in distribution of the ratios of the sample and population eigenvalues to a random vector of multiples of the eigenvalues corresponding to the random matrix $\widetilde{U}_0$ of Theorem 2.1.

Proposition 2.1. *Under the assumptions of Theorem 2.1 and for n fixed, we have the joint weak convergence*

$$\left(\frac{\widehat{\tau}_1}{\tau_1}, \frac{\widehat{\tau}_2}{\tau_2}, \dots, \frac{\widehat{\tau}_p}{\tau_p}\right)^\top \xrightarrow{w} n^{-1} \left(\frac{\ell_1}{c_1}, \frac{\ell_2}{c_2}, \dots, \frac{\ell_p}{c_p}\right)^\top$$

when $d \to \infty$, where $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p > 0$ are the eigenvalues of the random matrix $\widetilde{U}_0$.

Proof. Note the following

$$\left(\frac{\widehat{\tau_1}}{\tau_1}, \frac{\widehat{\tau_2}}{\tau_2}, \dots, \frac{\widehat{\tau_p}}{\tau_p}\right)^\top = \operatorname{diag}\left(\frac{d^\alpha}{\tau_1}, \frac{d^\alpha}{\tau_2}, \dots, \frac{d^\alpha}{\tau_p}\right) \left(\frac{\widehat{\tau_1}}{d^\alpha}, \frac{\widehat{\tau_2}}{d^\alpha}, \dots, \frac{\widehat{\tau_p}}{d^\alpha}\right)^\top$$

which by Theorem 2.1 tends in distribution to

$$\operatorname{diag}\left(\frac{1}{c_1}, \frac{1}{c_2}, \dots, \frac{1}{c_p}\right) \left(\frac{\ell_1}{n}, \frac{\ell_2}{n}, \dots, \frac{\ell_p}{n}\right)^\top = n^{-1} \left(\frac{\ell_1}{c_1}, \frac{\ell_2}{c_2}, \dots, \frac{\ell_p}{c_p}\right)^\top. \qquad \square$$

Remark 2.1. Suppose $\tau_1 \geq \cdots \geq \tau_p \geq \sigma_{p+1} \geq \cdots \geq \sigma_d > 0$ are functions of d. The two previous results hold if we consider the covariance matrix

$$\Sigma = O \Lambda O^\top \qquad \text{where } \Lambda = \operatorname{diag}(\tau_1, \dots, \tau_p, \sigma_{p+1}, \dots, \sigma_d),$$

where $\tau_1, \dots, \tau_p$ have the same asymptotic order of magnitude in d,

$$\max(\sigma_{p+1}, \dots, \sigma_d)/d^{\alpha-1} \longrightarrow 0 \qquad \text{as } d \to \infty,$$

and O is a $d \times d$ orthogonal matrix. The proof is similar to that of Theorem 2.1; we only need to prove that

$$d^{-\alpha} \sum_{i=p+1}^{d} \sigma_i W_i \xrightarrow{w} \mathbf{0} \qquad \text{as } d \to \infty, \tag{2.3}$$

where W_i is as in the proof of Theorem 2.1 and $\mathbf{0}$ is the $n \times n$ matrix of zeros. We use the result that if A_d, B_d and $A_d - B_d$ are nonnegative definite matrices and $A_d \to \mathbf{0}$ as $d \to \infty$, then $B_d \to \mathbf{0}$ as $d \to \infty$. Let $M_d = \max(\sigma_{p+1}, \ldots, \sigma_d)$ and $V = \sum_{i=p+1}^{d} W_i$. Since W_i is non-negative definite and $M_d - \sigma_i > 0$ for $i = p+1, \ldots, d$, we have that $A_d = d^{-\alpha} M_d V$, $B_d = d^{-\alpha} \sum_{i=p+1}^{d} \sigma_i W_i$ and $A_d - B_d$ are nonnegative definite matrices. Let $\varepsilon > 0$; analogously to the proof of (2.2) it can be seen that

$$P(\|A_d\| > \varepsilon) \leq \frac{(d-p)^2 M_d^2 n^2 K_n}{d^{2\alpha} \varepsilon^2} = \left(\frac{d-p}{d}\right)^2 \left(\frac{M_d}{d^{\alpha-1}}\right)^2 \frac{n^2 K_n}{\varepsilon^2}$$

and the right side of the last inequality tends to zero as $d \to \infty$ since $d^{-(\alpha-1)} \times M_d \to 0$. Therefore $A_d \to \mathbf{0}$ in probability and in distribution as $d \to \infty$. Then we have (2.3).

In particular, the last remark allows us to generalize the previous results to spiked covariance models such that their p largest eigenvalues have same asymptotic order of magnitude and the rest are bounded for a constant as d tends to infinity.

2.2 First $d \to \infty$ and then $n \to \infty$ in a second step

In this section, we study the asymptotic behavior of the sample eigenvalues of our spiked covariance model by letting first the data dimension $d \to \infty$ and in a second step letting the sample size $n \to \infty$. The next theorem is a generalization of the result given in Section 4.2 of Ahn et al. (2007) which considers the case $p = 1$.

Theorem 2.2. *Suppose that the unknown covariance matrix of the columns of X is given by the spiked covariance model* (1.1), *with $p < n < d$ and where $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_p$ have the same asymptotic order of magnitude in d. Suppose that X satisfies* (*a*) *and the following assumption*:

(b′) *Let Z_i be the i-th row of Z and define $\widetilde{Z}_p = [Z_1^\top, \ldots, Z_p^\top]^\top$. Assume that $\widetilde{Z}_p$ converges in distribution to some $p \times n$ matrix $Y_n = (y_{ij,n})$ as $d \to \infty$, which has rank p with probability one and its entries have uniformly bounded fourth moments with respect to n, that is for some $M > 0$ we have $E(y_{ij,n}^4) \leq M$ for all $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, n$ and $n = p+1, p+2, \ldots$. Furthermore, suppose that the matrix distribution of $Y_n Y_n^\top$ is continuous.*

Then we have

$$\left(\frac{\widehat{\tau}_1}{\tau_1}, \frac{\widehat{\tau}_2}{\tau_2}, \dots, \frac{\widehat{\tau}_p}{\tau_p}\right)^\top \xrightarrow{w} (1, 1, \dots, 1)^\top \qquad as\ d \to \infty, n \to \infty, \tag{2.4}$$

where the limits are applied successively.

For the proof of this theorem, we first have the following Law of Large Numbers for random matrices and vector of eigenvalues. It also gives an extension of the one-dimensional fact that if χ_n^2 is a chi-square random variable with n degrees of freedom, then χ_n^2/n converges to 1 in probability (almost surely and in distribution), as $n \to \infty$.

Proposition 2.2. *Let Y_n be a sequence of $p \times n$ random matrices with $p < n$, such that its columns are independent with mean zero and identity covariance matrix. Assume that the rank of $Y_n = (y_{ij,n})$ is p with probability one and its entries have uniformly bounded fourth moments with respect to n, that is $E(y_{ij,n}^4) \le K$ for all $i = 1, 2, \dots, p$, $j = 1, 2, \dots, n$ and $n = p + 1, p + 2, \dots$. Let $A_n = Y_n Y_n^\top$. Then we have*:

(i)

$$n^{-1} A_n \xrightarrow{w} I_p \qquad as\ n \to \infty.$$

(ii) *Assume that $\Sigma = O \Lambda O^\top$ is a $p \times p$ positive definite matrix, where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ is the diagonal matrix of its eigenvalues and O is the $p \times p$ orthogonal matrix of its eigenvectors. Let $\ell_1 \ge \ell_2 \ge \dots \ge \ell_p$ be the eigenvalues of $W_n = O \Lambda^{1/2} A_n \Lambda^{1/2} O^\top$. Then*

$$n^{-1}\left(\frac{\ell_1}{\lambda_1}, \frac{\ell_2}{\lambda_2}, \dots, \frac{\ell_p}{\lambda_p}\right)^\top \xrightarrow{w} (1, 1, \dots, 1)^\top \qquad as\ n \to \infty.$$

Proof. (i) We have that $Y_n Y_n^\top = (\sum_{k=1}^n y_{ik,n} y_{jk,n})$; therefore

$$n^{-1} Y_n Y_n^\top - I_p = \left(n^{-1} \sum_{k=1}^n y_{ik,n} y_{jk,n} - \delta_{i,j}\right),$$

where $\delta_{i,j}$ is one if $i = j$ and zero otherwise. It is sufficient to prove that for all $\varepsilon > 0$

$$P\left(\left|\sum_{k=1}^n n^{-1} y_{ik,n} y_{jk,n} - \delta_{i,j}\right| > \varepsilon\right) \longrightarrow 0 \qquad \text{as } n \to \infty. \tag{2.5}$$

For the case $i = j$, by the Chebyshev's inequality and the assumptions for Y_n we have that

$$\begin{aligned} P\left(\left|n^{-1}\sum_{k=1}^{n} y_{ik,n}^2 - 1\right| > \varepsilon\right) &\leq \varepsilon^{-2}\operatorname{Var}\left(n^{-1}\sum_{k=1}^{n} y_{ik,n}^2\right) \\ &= (n\varepsilon)^{-2}E\left[\left(\sum_{k=1}^{n} y_{ik,n}^2 - n\right)^2\right] \\ &= (n\varepsilon)^{-2}E\left(\sum_{k=1}^{n} y_{ik,n}^4 + 2\sum_{k_1<k_2} y_{ik_1,n}^2 y_{ik_2,n}^2 - 2n\sum_{k=1}^{n} y_{ik,n}^2 + n^2\right) \\ &= (n\varepsilon)^{-2}\left(\sum_{k=1}^{n} E\big(y_{ik,n}^4\big) - n\right). \end{aligned} \tag{2.6}$$

Since $E(y_{ik,n}^4) \leq M$ for all i, k and $n = p+1, p+2, \ldots$, the last expression of (2.6) is less than or equal to $(n\varepsilon)^{-2}(nM - n) = n^{-1}\varepsilon^{-2}(M-1)$ which tends to zero as $n \to \infty$. Thus, we have (2.5).

Analogously, for the case $i \neq j$, by the Chebyshev's inequality and the assumptions for Y_n we have

$$\begin{aligned} P\left(\left|n^{-1}\sum_{k=1}^{n} y_{ik,n}^2\right| > \varepsilon\right) &\leq (n\varepsilon)^{-2}\operatorname{Var}\left(\sum_{k=1}^{n} y_{ik,n}y_{jk,n}\right) \\ &= (n\varepsilon)^{-2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n} E(y_{ik_1,n}y_{jk_1,n}y_{ik_2,n}y_{jk_2,n}) \\ &= (n\varepsilon)^{-2}\sum_{k=1}^{n} E\big(y_{ik,n}^2 y_{jk,n}^2\big). \end{aligned} \tag{2.7}$$

By the Holder's inequality, we have $E(y_{ik,n}^2 y_{jk,n}^2) \leq E(y_{ik,n}^4)^{1/2}E(y_{jk,n}^4)^{1/2} \leq M$, thus the last expression of (2.7) is less than or equal to $n^{-1}\varepsilon^{-2}M$ which tends to zero as $n \to \infty$.

(ii) Suppose that $W_n = V_n \ell_n V_n^\top$, where $\ell_n = \operatorname{diag}(\ell_1, \ldots, \ell_p)$ is the diagonal matrix of the eigenvalues of W_n and V_n is the orthogonal matrix of its eigenvectors. Since $n^{-1}A_n \overset{w}{\to} I_p$ as $n \to \infty$ by (i), we have that $V_n(n^{-1}\ell_n)V_n^\top = n^{-1}W_n \overset{w}{\to} \Sigma = O\Lambda O^\top$ and therefore $n^{-1}\ell_n \overset{w}{\to} \Lambda$ as $n \to \infty$. It follows that $n^{-1}\Lambda^{-1}\ell_n \overset{w}{\to} I_p$ as $n \to \infty$. □

Proof of Theorem 2.2. Let $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p > 0$ be the eigenvalues of the matrix $\widetilde{U}_0 = \mathcal{C}_p^{1/2} Y_n Y_n^\top \mathcal{C}_p^{1/2}$ with $\mathcal{C}_p = \operatorname{diag}(c_1, \ldots, c_p)$. Let $F_{\mathbf{1}_p}$, $F_{\widehat{\tau}/\tau}$ and $F_{n^{-1}\ell/\mathcal{C}_p}$ be the distribution functions of $\mathbf{1}_p = (1, 1, \ldots, 1)^\top$, $\widehat{\tau}/\tau = (\frac{\widehat{\tau}_1}{\tau_1}, \frac{\widehat{\tau}_2}{\tau_2}, \ldots, \frac{\widehat{\tau}_p}{\tau_p})^\top$ and $n^{-1}\ell/\mathcal{C}_p = n^{-1}(\frac{\ell_1}{c_1}, \frac{\ell_2}{c_2}, \ldots, \frac{\ell_p}{c_p})^\top$, respectively. Since $Y_n Y_n^\top$ has continuous matrix distribution then $\widetilde{U}_0$ and $n^{-1}\ell/\mathcal{C}_p$ have continuous distributions. Therefore, the continuity set of $F_{n^{-1}\ell/\mathcal{C}_p}$ is given by $\mathcal{C}(F_{n^{-1}\ell/\mathcal{C}_p}) = \mathbb{R}^p$. By Proposition 2.1

$$\lim_{d\to\infty} |F_{\widehat{\tau}/\tau}(t) - F_{n^{-1}\ell/\mathcal{C}_p}(t)| = 0$$

for all $t \in \mathbb{R}^p$. Therefore,

$$\lim_{d\to\infty} |F_{\widehat{\tau}/\tau}(t) - F_{\mathbf{1}_p}(t)| = |F_{n^{-1}\ell/\mathcal{C}_p}(t) - F_{\mathbf{1}_p}(t)| \qquad \forall t \in \mathbb{R}^p.$$

Since $\widetilde{Z}_p$ has independent column vectors and it converges in distribution to Y_n, the column vectors of Y_n are also independent. Because $\widetilde{Z}_p$ has uniformly bounded fourth moment with respect to d, by Theorem 4.5.2 of Chung (2001) we have $E(z_{ij}) = 0 \to E(y_{ij,n})$, $E(z_{ij}^2) = 1 \to E(y_{ij,n}^2)\,\forall i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, n$, and $E(z_{ik}z_{jk}) = 0 \to E(y_{ik,n}y_{jk,n})\,\forall k = 1, 2, \ldots, n$, $i \neq j$, as $d \to \infty$. Therefore, Y_n has mean zero and its column vectors have identity matrix. Thus by Proposition 2.2(ii)

$$\lim_{n\to\infty} \left|F_{n^{-1}\ell/\mathcal{C}_p}(t) - F_{\mathbf{1}_p}(t)\right| = 0$$

for all t in the continuity set of $F_{\mathbf{1}_p}$, namely $\mathcal{C}(F_{\mathbf{1}_p})$. Thus

$$\lim_{n\to\infty} \lim_{d\to\infty} \left|F_{\widehat{\tau}/\tau}(t) - F_{\mathbf{1}_p}(t)\right| = 0$$

for all $t \in \mathcal{C}(F_{\mathbf{1}_p})$. □

Thus by Proposition 2.1 and Thereore 2.2, we conclude that the p largest sample eigenvalues of the considered spiked covariance model increase jointly at the same speed as their population counterpart, generalizing in this way the results of Section 4.2 of Ahn et al. (2007) and Section 4.5 of Jung and Marron (2009).

3 Subspace consistency of sample eigenvectors

As mentioned in Jung and Marron (2009), in the case when several population eigenvalues indexed by J are similar, their corresponding sample eigenvectors may not be distinguishable. Therefore, for $j \in J$ the sample eigenvector v_j, corresponding to the j-th sample eigenvalue, will not be consistent for its corresponding population eigenvector o_j but rather may asymptotically be in $E_J = \operatorname{span}\{o_j : j \in J\}$, the linear span generated by $\{o_j : j \in J\}$. We define

$$\begin{aligned}\mathrm{Angle}(v_j, E_J) &= \arccos\left(\frac{v_j^\top [\mathrm{Proj}_{E_J} v_j]}{\|v_j\| \| \mathrm{Proj}_{E_J} v_j\|}\right)\\ &= \arccos\left(\frac{v_j^\top (\sum_{i\in J}(o_i^\top v_j)o_i)}{\|v_j\| \| \sum_{i\in J}(o_i^\top v_j)o_i\|}\right),\end{aligned}$$

the second equality being true when the o_j's are mutually orthogonal.

We say that

- v_i is *consistent* if
$$\mathrm{Angle}(v_i, o_i) \xrightarrow{P} 0 \qquad \text{as } d \to \infty.$$
- v_i is *strongly inconsistent* if
$$\mathrm{Angle}(v_i, o_i) \xrightarrow{P} \frac{\pi}{2} \qquad \text{as } d \to \infty.$$
- v_i is *subspace consistent* if
$$\mathrm{Angle}(v_i, E_J) \xrightarrow{P} 0 \qquad \text{as } d \to \infty$$
for some set of indices J with $i \in J$.

From the results of Jung and Marron (2009), under our spiked covariance model the first p sample eigenvectors $v_1, v_2, \ldots, v_p$ are subspace consistent and the sample eigenvectors $v_{p+1}, v_{p+2}, \ldots, v_n$ are strongly inconsistent, when $d \to \infty$ and n is fixed. We give a similar proof of the subspace consistency of the first p sample eigenvectors using the results of our Section 2 when $d \to \infty$ and n is fixed. We recall that the population eigenvectors of the spiked covariance model (1.1) are the column vectors, $o_1, o_2, \ldots, o_d$, of the matrix O.

Theorem 3.1. *Under the same assumptions of Theorem 2.2, let $v_1, v_2, \ldots, v_p$ be the sample eigenvectors corresponding to the first p sample eigenvalues $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \cdots \geq \widehat{\tau}_p$. Then for $i = 1, 2, \ldots, p$,*

$$\mathrm{Angle}(v_i, E_J) \xrightarrow{w} 0 \qquad \text{as } d \to \infty, \tag{3.1}$$

where $E_J = \mathrm{span}\{o_1, o_2, \ldots, o_p\}$.

Proof. We follow closely the ideas in Ahn et al. (2007) and Jung and Marron (2009). Consider the eigenvalue decomposition of the sample covariance matrix $S = VLV^\top$, where $L = \mathrm{diag}(\widehat{\tau}_1, \ldots, \widehat{\tau}_n, 0, \ldots, 0)$ is the diagonal matrix of the sample eigenvalues and $V = [v_1, v_2, \ldots, v_d]$ is the matrix of the sample eigenvectors $v_j = (v_{1j}, \ldots, v_{dj})^\top$, $j = 1, 2, \ldots, d$. We assume that V is orthogonal, that is $V^\top V = I_d$. We have $\Sigma = O\Lambda O^\top$, where $\Lambda = \mathrm{diag}(\tau_1, \ldots, \tau_p, \sigma, \ldots, \sigma)$ is the diagonal matrix of eigenvalues of Σ and $O = [o_1, \dots, o_d]$ the $d \times d$ orthogonal matrix of its eigenvectors. A standardized version of the sample covariance matrix S is given by

$$\widetilde{S} = \Lambda^{-1/2} O^\top S O \Lambda^{-1/2} = \Lambda^{-1/2} O^\top V L V^\top O \Lambda^{-1/2}. \tag{3.2}$$

Thus we have $S = n^{-1} X X^\top = n^{-1} O \Lambda^{1/2} Z Z^\top \Lambda^{1/2} O^\top$ and

$$\widetilde{S} = n^{-1} \Lambda^{-1/2} O^\top O \Lambda^{1/2} Z Z^\top \Lambda^{1/2} O^\top O \Lambda^{-1/2} = n^{-1} Z Z^\top. \tag{3.3}$$

From (3.2), we have that the j-th diagonal entry of $\widetilde{S}$ is given by $\widetilde{s}_{jj} = \lambda_j^{-1} \sum_{i=1}^n \widehat{\tau}_i (v_i^\top o_j)^2$, where λ_j is the j-th diagonal entry of Λ, for $j = 1, 2, \dots, d$. Therefore $\lambda_j^{-1} \widehat{\tau}_i (v_i^\top o_j)^2 \le \widetilde{s}_{jj}$, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, d$. Furthermore, from (3.3) we also have $\widetilde{s}_{jj} = n^{-1} Z_j Z_j^\top = n^{-1} \sum_{k=1}^n z_{jk}^2$. Thus for $i = 1, 2, \dots, n$

$$\sum_{j=p+1}^{d} (v_i^\top o_j)^2 \le \sum_{j=p+1}^{d} \frac{\lambda_j}{\widehat{\tau}_i} \widetilde{s}_{jj} = \frac{\sigma}{n \widehat{\tau}_i} \sum_{j=p+1}^{d} \sum_{k=1}^{n} z_{jk}^2 = \frac{\sigma}{n} \frac{d^\alpha}{\widehat{\tau}_i} \sum_{k=1}^{n} \sum_{j=p+1}^{d} \frac{z_{jk}^2}{d^\alpha}. \tag{3.4}$$

By Theorem 2.1, we have $\widehat{\tau}_i / d^\alpha \xrightarrow{w} \ell_i / n$ as $d \to \infty$, for $i = 1, 2, \dots, p$. Since the entries of Z have uniformly bounded fourth moments in d, we have that there exist $K_n^* > 0$ such that $E(z_{jk}^2) \le K_n^*$ for all $j = 1, 2, \dots, d$, $k = 1, 2, \dots, n$ and $d = n+1, n+2, \dots$. Let $\varepsilon > 0$ and observe that

$$P\left(\left| \sum_{j=p+1}^{d} \frac{z_{jk}^2}{d^\alpha} \right| > \varepsilon \right) \le \frac{E(\sum_{j=p+1}^{d} z_{jk}^2)}{d^\alpha \varepsilon} \le \frac{(d-p) K_n^*}{d^\alpha \varepsilon} \longrightarrow 0 \qquad \text{as } d \to \infty,$$

that is $\sum_{j=p+1}^{d} d^{-\alpha} z_{jk}^2 \xrightarrow{P} 0$ as $d \to \infty$. Hence, it follows from (3.4) that

$$\sum_{j=p+1}^{d} (v_i^\top o_j)^2 \xrightarrow{w} 0 \qquad \text{as } d \to \infty \tag{3.5}$$

for $i = 1, 2, \dots, p$. Since $V^\top O O^\top V = I_d$ we have $\sum_{j=1}^{d} (v_i^\top o_j)^2 = 1$, and thus (3.5) implies

$$\sum_{j=1}^{p} (v_i^\top o_j)^2 \xrightarrow{w} 1 \qquad \text{as } d \to \infty \tag{3.6}$$

for $i = 1, 2, \dots, p$.

Finally, following the arguments in Section 5.2.2 of Jung and Marron (2009), we have that for $i = 1, 2, \dots, p$,

$$\mathrm{Angle}(v_i, E_J) = \arccos\left(\left[\sum_{j=1}^{p} (v_i^\top o_j)^2 \right]^{1/2} \right).$$

Then from (3.6) it follows that

$$\operatorname{Angle}(v_i, E_J) \xrightarrow{w} 0 \qquad \text{as } d \to \infty$$

for $i = 1, 2, \ldots, p$. □

Remark 3.1. The result of Theorem 3.1 holds if we consider that the population covariance matrix is as in Remark 2.1. The proof is similar to that of Theorem 3.1.

4 The Gaussian case and some statistical eigen-inference

In this section, we assume that the data matrix X comes from a sample of multivariate Gaussian distribution $N(0, \Sigma)$ where the matrix Σ is a spiked covariance matrix under the assumption that the p largest eigenvalues have same order of magnitude in d, with $c_1 = \cdots = c_p = c > 0$ in (1.2). In this case the matrix $\widetilde{U}_0$ of Theorem 2.1 follows a Wishart random matrix distribution $\mathcal{W}(n, cI_p)$.

We now use the asymptotic results in Section 2, in particular the joint convergence in distribution of the nonzero sample eigenvalues, to consider some inference problems for the population eigenvalues and to show that some of the classical statistics are also useful in the cases when d goes to infinity and n is fixed, and when d, n go to infinity and $d \gg n$.

We first point out three asymptotic results. The first one is a kind of central limit theorem for the vector of the ratios of the sample and population eigenvalues under our model and when d and n go to infinity successively.

Theorem 4.1. *Under the same assumptions as in Theorem 2.1, suppose $c_1 = c_2 = \cdots = c_p = c > 0$ in (1.2) and the columns of X are Gaussian. Let $\frac{\widehat{\tau}}{\tau} = (\frac{\widehat{\tau}_1}{\tau_1}, \ldots, \frac{\widehat{\tau}_p}{\tau_p})^\top$ and let $\varphi = (\varphi_1, \ldots, \varphi_p)$ be the vector of eigenvalues of a standard $p \times p$ Gaussian matrix with density function*

$$f_\varphi(\varphi_1, \ldots, \varphi_p) = \frac{\pi^{p(p-1)/4}}{2^{p/2}\Gamma_p(p/2)} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\varphi_i^2\right)\prod_{i<j}^{p}(\varphi_j - \varphi_i), \qquad \varphi_p > \cdots > \varphi_1. \tag{4.1}$$

Then we have that

$$n^{1/2}\left(\frac{\widehat{\tau}}{\tau} - \mathbf{1}_p\right)^\top \xrightarrow{w} \varphi \qquad \text{as } d \to \infty, n \to \infty, \tag{4.2}$$

where the limits are applied successively.

Proof. Without lost of generality, we can assume $c = 1$. Let $L = n^{-1}(\ell_1, \ldots, \ell_p)^\top$, where $\ell_1 \geq \cdots \geq \ell_p > 0$ are the eigenvalues of the matrix $\widetilde{Z}_p\widetilde{Z}_p^\top$ with distribution $\mathcal{W}(n, I_p)$. By Proposition 2.1 we have $\widehat{\tau}/\tau \xrightarrow{w} L$ as $d \to \infty$ and by Corollary 13.3.2 in Anderson (2003) we have $n^{1/2}(L - \mathbf{1}_p)^\top \xrightarrow{w} \varphi$ as $n \to \infty$, where the random vector φ has density function given by (4.1); see Theorem 13.3.5 in Anderson (2003). Thus, we have (4.2). □

The next two propositions are consequences of the joint convergence in distribution of the nonzero sample eigenvalues given in Theorem 2.1, and they are useful to study some inference problems in the context of data with dimension greater than the sample size.

Proposition 4.1. *Under the assumptions of Theorem 2.1 and considering $c_1 = c_2 = \cdots = c_p = c > 0$ in (1.2), let $T = \operatorname{diag}(\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_p)$ be the diagonal matrix of the p largest sample eigenvalues and $\ell = \operatorname{diag}(\ell_1, \ell_2, \ldots, \ell_p)$, where $\ell_1, \ell_2, \ldots, \ell_p$ are the nonzero eigenvalues of a Wishart matrix with distribution $\mathcal{W}(n, cI_p)$. Then we have the following when n is fixed*:

(i) $\operatorname{tr}(T)/\tau_i \overset{w}{\to} \mathcal{X}^2_{np}/n$ *as* $d \to \infty$, *for* $i = 1, 2, \ldots, p$.

(ii) $\widetilde{V} = \det(T)/[\operatorname{tr}(T)/p]^p \overset{w}{\to} V = \det(\ell)/[\operatorname{tr}(\ell)/p]^p$; *furthermore* $\widetilde{V}$ *is asymptotically independent of* $\operatorname{tr}(T)/d^\alpha$ *as* $d \to \infty$.

(iii) $\det(T)/\tau_i^p \overset{w}{\to} (\prod_{j=1}^p \mathcal{X}^2_{n-j+1})/n^p$ *as* $d \to \infty$ *for* $i = 1, 2, \ldots, p$, *where* $\mathcal{X}^2_{n-j+1}$ *are independent random variables with chi-square distribution with* $n - j + 1$ *degrees of freedom, for* $j = 1, 2, \ldots, p$.

Proof. Using the continuity of the trace and determinant, from the joint weak convergence of the eigenvalues in Theorem 2.1 and the assumption of same asymptotic order of magnitude in d we have that for n fixed

$$\operatorname{tr}(T)/\tau_i = [\operatorname{tr}(T)/(cd^\alpha)][cd^\alpha/\tau_i] \xrightarrow{w} \operatorname{tr}(\ell)/cn, \tag{4.3}$$

$$\widetilde{V} = \frac{\det(T/d^\alpha)}{[\operatorname{tr}(T/d^\alpha)/p]^p} \xrightarrow{w} \frac{\det(\ell/n)}{[\operatorname{tr}(\ell/n)/p]^p} = V, \tag{4.4}$$

$$\det(T)/\tau_i^p = [\det(T)/(cd^\alpha)^p][cd^\alpha/\tau_i]^p \xrightarrow{w} \det(\ell)/(nc)^p, \tag{4.5}$$

as $d \to \infty$. From Theorem 3.2.20 in Muirhead (1982), we have that $\operatorname{tr}(\ell)/cn \sim \mathcal{X}^2_{np}/n$ as $d \to \infty$ and $\det(\ell)/[\operatorname{tr}(\ell)/p]^p$ is independent of $\operatorname{tr}(\ell)/n$. Thus, using (4.3) and (4.4) we have (i) and (ii). It follows from Theorem 3.2.15 in Muirhead (1982) that $\det(\ell)/(nc)^p$ is equal in distribution to $(\prod_{j=1}^p \mathcal{X}^2_{n-j+1})/n^p$, where $\mathcal{X}^2_{n-j+1}$ for $j = 1, 2, \ldots, p$, are independent random variables with chi-square distribution with $n - j + 1$ degrees of freedom, thus from (4.5) we have (iii). □

Proposition 4.2. *Under the assumptions of Theorem 2.1 and considering $c_1 = c_2 = \cdots = c_p = c > 0$ in (1.2), let $T = \operatorname{diag}(\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_p)$ be the diagonal matrix of the p largest sample eigenvalues. Then we have the following*:

(i) $(np/2)^{1/2}[\operatorname{tr}(T)/p - \tau_i]/\tau_i \overset{w}{\to} N(0, 1)$ *as* $d \to \infty$, $n \to \infty$, *where the limits are applied successively, for* $i = 1, 2, \ldots, p$.

(ii) *Let* $\widetilde{V} = \det(T)/[\mathrm{tr}(T)/p]^p$ *and* $\rho = 1 - (2p^2 + p + 2)/(6np)$, *then* $\widetilde{U} = -n\rho \ln(\widetilde{V}) \overset{w}{\to} \mathcal{X}_r^2$ *as* $d \to \infty$, $n \to \infty$, *where the limits are applied successively and* $\mathcal{X}_r^2$ *is a chi-square r.v. with* $r = (p+2)(p-1)/2$ *degrees of freedom.*

Proof. It follows from Proposition 4.1(i) that

$$\left(\frac{np}{2}\right)^{1/2}\left(\frac{\mathrm{tr}(T)/p - \tau_i}{\tau_i}\right) = \frac{n\,\mathrm{tr}(T)/\tau_i - np}{(2np)^{1/2}} \xrightarrow{w} \frac{\mathcal{X}_{np}^2 - np}{(2np)^{1/2}} \tag{4.6}$$

as $d \to \infty$, where $\mathcal{X}_{np}^2$ is a chi-square r.v. with np degrees of freedom. Since $\mathcal{X}_{np}^2$ is equal in distribution to $\sum_{j=1}^n \mathcal{X}_{p,j}^2$, where $\mathcal{X}_{p,j}^2$ for $j = 1, 2, \ldots, n$ are independent r.v.'s with chi-square distribution with p degrees of freedom, we have by the CLT that

$$\frac{\mathcal{X}_{np}^2 - np}{(2np)^{1/2}} \xrightarrow{w} N(0,1) \tag{4.7}$$

as $n \to \infty$. Thus, from (4.6) and (4.7) we have (i). From Proposition 4.1(ii) and Theorem 8.3.7 in Muirhead (1982), we obtain (ii). □

4.1 Hypothesis test for the p largest population eigenvalues

Let M_d be the maximum of the $d - p$ smaller population eigenvalues and suppose that we have evidence that the sequence $\{M_d\}_{d\in\mathbb{N}}$ is bounded by a constant number M, that is $0 < M_d \le M$ for all $d > n$ and $d \in \mathbb{N}$. Consider the null hypothesis

$$H_0 : \tau_i/d^\alpha \to c \qquad \text{for all } i = 1, 2, \ldots, p, \tag{4.8}$$

where $\alpha > 1$ and $c > 0$ are unspecified numbers. Under H_0 we have a population covariance matrix as in Remark 2.1, therefore all the results of Section 2 are valid in this case.

In order to test the null hypothesis H_0 that the first p largest population eigenvalues have the same asymptotic order of magnitude and $c_1 = c_2 = \cdots = c_p = c > 0$, we can use the classical *ellipticity statistic* $\widetilde{V} = \det(T)/[\mathrm{tr}(T)/p]^p$, see Muirhead (1982, p. 336), where $T = \mathrm{diag}(\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_p)$ is the diagonal matrix of the p largest sample eigenvalues. The null hypothesis (4.8) can be tested in the following two situations:

- *When* $d \to \infty$ *and* n *is fixed.* By Proposition 4.1(ii) $\widetilde{V} \overset{w}{\to} V = \det(\ell)/[\mathrm{tr}(\ell)/p]^p$ as $d \to \infty$, where $\ell = \mathrm{diag}(\ell_1, \ell_2, \ldots, \ell_p)$ and $\ell_1, \ell_2, \ldots, \ell_p$ are the eigenvalues of a Wishart matrix with distribution $\mathcal{W}(n, cI_p)$. Therefore, if $\widetilde{V}_0$ is the observed value of $\widetilde{V}$, a test of asymptotic significance level β is to reject H_0 if $\widetilde{V}_0 \le k_\beta$, where k_β is the lower $100\beta\%$ point of the distribution of V. We expect that this rejection region works very well, because if A is a $p \times p$ random matrix with distribution $\mathcal{W}(n, \Psi)$ and if η_0 is the observed value of

$\eta = \det(A)/(\mathrm{tr}(A)/p)^p$, then the test that rejects $H_1 : \Psi = cI_p$ if $\eta_0 \leq k_\beta$ is unbiased, see Muirhead (1982, p. 336). Explicit expressions for the density function of V are given in Consul (1967) and Consul (1969), and tables of percentage points of V for some values of p and various values of n can be found in Nagarsenker and Pillai (1973).

- *When $d, n \to \infty$ and $d \gg n$.* By Proposition 4.2(ii), the statistic $\widetilde{R} = -n\rho \times \ln(\widetilde{V}) \overset{w}{\to} \mathcal{X}_r^2$, where $\mathcal{X}_r^2$ is a chi-square r.v. with $r = (p+2)(p-1)/2$ degrees of freedom. Thus, if $\widetilde{R}_0$ is the observed value of $\widetilde{R}$, a test of asymptotic significance level β is to reject H_0 if $\widetilde{R}_0 > u_\beta$, where u_β is the upper $100\beta\%$ point of the chi-square distribution with r degrees of freedom.

4.2 Confidence intervals for the p largest population eigenvalues

Under the hypothesis H_0 given in (4.8) we have, by Proposition 2.1 and Theorem 2.2, that the p largest sample eigenvalues increases at the same speed as their population counterpart, however this does not guarantee that these sample eigenvalues are good approximation for their population counterpart. We may be interested in a confidence interval for the population eigenvalue τ_i, for $i = 1, 2, \ldots, p$. Again, we have two situations in which we may address this problem:

- *When $d \to \infty$ and n is fixed.* From Proposition 4.1(i), for $0 < \beta < 1$ and d is large enough

$$P\left(\frac{k_{\beta/2}}{n} \leq \frac{\mathrm{tr}(T)}{\tau_i} \leq \frac{u_{\beta/2}}{n}\right) \approx 1 - \beta,$$

where $k_{\beta/2}$ and $u_{\beta/2}$ are the lower and upper $100(\beta/2)\%$ point of the chi-square distribution with np degrees of freedom, respectively. Therefore, a confidence interval with asymptotic confidence level $1 - \beta$ for τ_i is

$$\left[\frac{n\,\mathrm{tr}(T)}{u_{\beta/2}}, \frac{n\,\mathrm{tr}(T)}{k_{\beta/2}}\right]. \tag{4.9}$$

- *When $d, n \to \infty$ and $d \gg n$.* From Proposition 4.2(i), for $0 < \beta < 1$ and d, n sufficiently large with $d \gg n$ we have

$$P\left(-z_{\beta/2} \leq \left(\frac{np}{2}\right)^{1/2}\left(\frac{\mathrm{tr}(T)/p - \tau_i}{\tau_i}\right) \leq z_{\beta/2}\right) \approx 1 - \beta,$$

where $z_{\beta/2}$ is the upper $100(\beta/2)\%$ point of the standard normal distribution. Thus, a confidence interval with asymptotic confidence level $1 - \beta$ for τ_i is

$$\left[\frac{\mathrm{tr}(T)/p}{1 + z_{\beta/2}[2/(np)]^{1/2}}, \frac{\mathrm{tr}(T)/p}{1 - z_{\beta/2}[2/(np)]^{1/2}}\right]. \tag{4.10}$$

5 Simulations

In this section, we present some simulation results to show the performance of the hypothesis tests and the confidence intervals proposed in Section 4. For the simulation study, we consider d-multivariate Gaussian data with mean zero and covariance matrix

$$\Sigma = \operatorname{diag}(\tau_1, \ldots, \tau_p, 1, \ldots, 1),$$

where $\tau_i = d^\alpha$ for $i = 1, 2, \ldots, p$, and $p = 2, 4$. We take $\alpha = 1.5, 3$, and sample sizes $n = 25, 50, 100$ for each value of α. This is because we want to assess the performance of the methodologies varying the order of magnitude of the largest eigenvalues and increasing the sample size. We take $d = 200$ and $d = 1000$ for each pair (α, n) to consider the case when $d > n$ and $d \gg n$, respectively.

For each setting, $M = 10{,}000$ replications of the data have been obtained, and for each replication the two hypothesis tests of Section 4.1 have been performed with significance level 5% and taking the corresponding value of p. In Table 1 are shown the empirical probabilities of the Type I error ($\varrho = P(\text{reject } H_0 | H_0 \text{ is true})$) of the two tests, given by

$$\widehat{\varrho}_1 = \frac{\#\{\widetilde{V}_0 \le k_\beta\}}{M}$$

for the hypothesis test based on the statistic $\widetilde{V}$, where k_β is the lower $100\beta\%$ point of the distribution of V; and

$$\widehat{\varrho}_2 = \frac{\#\{\widetilde{R}_0 > u_\beta\}}{M}$$

Table 1 *Empirical probabilities of Type I error of proposed hypothesis tests*

			$p = 2$	$p = 4$	
α	n	d	$\widehat{\varrho}_1$ $\widehat{\varrho}_2$	$\widehat{\varrho}_1$	$\widehat{\varrho}_2$
1.5	25	200	0.0469	0.0491	0.0495
		1000	0.0506	0.0488	0.0492
	50	200	0.0494	0.0481	0.0483
		1000	0.0508	0.0459	0.0461
	100	200	0.0451	0.0470	0.0471
		1000	0.0519	0.0487	0.0487
3	25	200	0.0482	0.0502	0.0510
		1000	0.0466	0.0518	0.0526
	50	200	0.0513	0.0491	0.0492
		1000	0.0494	0.0488	0.0489
	100	200	0.0460	0.0485	0.0486
		1000	0.0492	0.0542	0.0543

for the hypothesis test based on the statistic $\widetilde{R}$, where u_β is the upper $100\beta\%$ point of the chi-square distribution with $r=(p+2)(p-1)/2$ degrees of freedom. The values of k_β were calculated using the expressions of the distribution function of V given in Consul (1967).

For the case $p=2$ the values of $\widehat{\varrho}_1$ and $\widehat{\varrho}_2$ were exactly the same, this is because for this value of p the chi-square distribution $\mathcal{X}_r^2$ is a very good approximation to the distribution of $\widetilde{R}$, and therefore the two tests are equivalent. For the case $p=4$, we observe that $\widehat{\varrho}_1$ is slightly smaller than $\widehat{\varrho}_2$, and they tend to be similar as n increases. All the empirical probabilities of Type I error are close and around 5% as expected, thus we conclude that the two proposed tests perform very well and they can be used to test the null hypothesis (4.8) for HDLSS data.

These simulation results also show the usefulness of the asymptotic setting $d, n \to \infty$ and $d \gg n$ in the HDLSS context, since the results for this asymptotic setting are very similar to that of $d \to \infty$ and n is fixed, and the hypothesis test based on $\widetilde{R}$ is the easiest to perform.

Similarly, for $M=10{,}000$ replications of the data the confidence intervals (4.9) and (4.10) have been calculated with confidence level 95% and taking the corresponding value of p. In Table 2 is shown the empirical coverage of the two classes of intervals for each setting. C_1 and C_2 denote the empirical coverage of the confidence intervals (4.9) and (4.10), respectively.

We observe that C_1 is always smaller than C_2. This is because the intervals (4.10) are slightly wider than the intervals (4.9) and cover larger values. We also see that all the empirical coverages are near to 95%. Therefore, we conclude that these proposed intervals have good performance as confidence intervals for the p

Table 2 *Empirical coverages of proposed confidence intervals*

			$p=2$		$p=4$	
α	n	d	C_1	C_2	C_1	C_2
1.5	25	200	0.9507	0.9509	0.9513	0.9514
		1000	0.9486	0.9508	0.9510	0.9527
	50	200	0.9504	0.9545	0.9484	0.9492
		1000	0.9533	0.9548	0.9515	0.9518
	100	200	0.9502	0.9506	0.9491	0.9493
		1000	0.9534	0.9544	0.9516	0.9522
3	25	200	0.9509	0.9541	0.9516	0.9517
		1000	0.9478	0.9520	0.9512	0.9523
	50	200	0.9507	0.9540	0.9474	0.9478
		1000	0.9512	0.9516	0.9512	0.9513
	100	200	0.9524	0.9525	0.9511	0.9514
		1000	0.9468	0.9488	0.9475	0.9470

largest eigenvalues under the null hypothesis (4.8), in both asymptotic settings of the HDLSS context.

Acknowledgments

The authors thank to an anonymous referee for his/her valuable suggestions and comments to improve this manuscript. They are also grateful to Steve Marron for his feedback on an early version of this paper. The first author wants to thank the Centro de Investigacion en Matematicas (CIMAT) for the support provided during the elaboration of this paper and she also is grateful to CONACYT for her Ph.D. scholarship to study in CIMAT (2007–2010).

References

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94**, 760–766. MR2410023

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken, NJ: Wiley. MR1990662

Bai, Z. D. and Yang, J. (2008). Central limit theorems for eigenvalues in spiked population model. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques* **44**, 447–474. MR2451053

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382–1408. MR2279680

Chung, K. L. (2001). *A Course in Probability Theory*, 3rd ed. San Diego: Academic Press. MR1796326

Consul, P. C. (1967). On the exact distribution of the criterion W for testing sphericity in a p-variate normal distribution. *The Annals of Mathematical Statistics* **38**, 1170–1174. MR0212907

Consul, P. C. (1969). *The Exact Distribution of Likelihood Criteria for Different Hypothesis*. *Multivariate Analysis* **2**. New York: Academic Press. MR0260077

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Ser. B* **67**, 427–444. MR2155347

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29**, 295–327. MR1863961

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics* **37**, 4104–4130. MR2572454

Jung, S., Sen, A. and Marron, J. S. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis* **109**, 190–203. MR2922863

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, 2nd ed. New York: Wiley. MR0652932

Nagarsenker, B. N. and Pillai, K. C. S. (1973). The distribution of the sphericity test criterion. *Journal of Multivariate Analysis* **3**, 226–235. MR0329126

Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context. *Communications in Statistics—Theory and Methods*, **38**, 2634–2652. MR2568176

Centro de Investigacion en Matematicas
Jalisco S/N, Col. Valenciana, CP 36240
Guanajuato, Gto
Mexico
E-mail: addy@cimat.mx
pabreu@cimat.mx