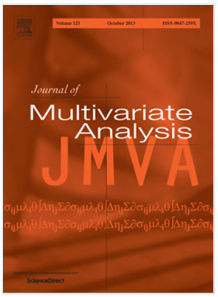
Accepted Manuscript



Likelihood ratio test for partial sphericity in high and ultra-high dimensions

Liliana Forzani, Antonella Gieco, Carlos Tolmasky

PII: S0047-259X(17)30199-9
DOI: http://dx.doi.org/10.1016/j.jmva.2017.04.001
Reference: YJMVA 4238

To appear in: *Journal of Multivariate Analysis*

Received date: 5 July 2016

Please cite this article as: L. Forzani, A. Gieco, C. Tolmasky, Likelihood ratio test for partial sphericity in high and ultra-high dimensions, *Journal of Multivariate Analysis* (2017), http://dx.doi.org/10.1016/j.jmva.2017.04.001

Likelihood ratio test for partial sphericity in high and ultra-high dimensions

Liliana Forzani[†], Antonella Gieco[†], Carlos Tolmasky[*]

[†] *Universidad Nacional del Litoral, Santa Fe de la Vera Cruz, Argentina*

[*] *Institute for Mathematics and Its Applications and MCFAM, University of Minnesota, Minneapolis, MN 55455.*

Abstract

We consider, in the setting of p and n large, sample covariance matrices whose population counterparts follow a spiked population model, i.e., with the exception of the first (largest) few, all the population eigenvalues are equal. We study the asymptotic distribution of the partial maximum likelihood ratio statistic and use it to test for the dimension of the population spike subspace. Furthermore, we extend this to the ultra-high-dimensional case, i.e., $p > n$. A thorough study of the power of the test gives a correction that allows us to test for the dimension of the population spike subspace even for values of the limit of p/n close to 1, a setting where other approaches have proved to be deficient.

Keywords: Sample covariance matrix, spiked population model, high-dimensional statistics, principal component analysis, random matrix theory

1. Introduction

In many applications involving high-dimensional data, a few of the dimensions contain most of the relevant information. Identifying how many dimensions should be kept in the analysis is of paramount importance in representing and modeling data efficiently. Even though this issue has attracted much attention from practitioners as well as researchers, there is still

Email addresses: liliana.forzani@gmail.com (Liliana Forzani[†]), antogieco@gmail.com (Antonella Gieco[†]), tolmasky@ima.umn.edu (Carlos Tolmasky[*])

no clear consensus on how to proceed in a systematic way. Among practitioners, a popular approach amounts to checking how many of the transformed variables explain a large part of the variance in the data and little (if any) attention is paid to the nature of what is discarded. An exception to this simplified approach is presented in [20], in which the authors compare the bulk of the eigenvalues to the typical bulk found in random matrix theory.

Systems of this sort, in which a small number of variables contain all the relevant information, appear in various fields. In an effort to understand these type of systems, Johnstone [11] introduced the spiked population model. In this model, all the population eigenvalues are equal to 1 except for a few fixed, larger eigenvalues that carry the relevant information. The behavior of the sample eigenvalues of the spiked population model in the high-dimensional case has been thoroughly studied in the past decade; see, e.g., [2, 3, 19]. In a remarkable result, Baik et al. [3] proved that the asymptotic behavior of the sample eigenvalues experiences a phase transition. If a population eigenvalue from the spike is not big enough, its value cannot be recovered from the samples: the estimated eigenvalue gets pulled towards the bulk, the noisy section of the matrix. On the other hand, if the spike population eigenvalue is bigger than a certain threshold, its value can be recovered from the limit of the estimates, which are, however, biased.

The same question about how many components should be kept was long ago answered in the traditional p fixed, n growing paradigm (here p indicates the dimension of the data $\mathbf{X}$ and n indicates the sample size). One of the most common tests assumes that the data follow a normal distribution and uses the maximum likelihood ratio statistics $LRT_d = L_d/L_p$, where L_d indicates the maximum likelihood under the null hypothesis (that d components should be kept) while L_p is the maximum likelihood under the full model [15]. This maximum likelihood ratio test is used sequentially, starting with $d = 0$ and estimating d as the first hypothesized value that is not rejected. In the fixed p and n growing paradigm, under the null hypothesis, $\ln(LRT_d)$ has a known asymptotic distribution—a fact used by Bartlett [4] and by Lawley [13] to build the rejection region of the test. Another common approach, which has the advantage of requiring no subjective judgments, is based on the application of information theoretic criteria. Wax and Kailath [26] presented an estimator in this direction using the minimum description length (MDL) principle [21, 22]. In both cases, sequential testing or information criteria, a crucial ingredient is the knowledge of the asymptotic distribution of the maximum likelihood ratio statistic under the null hypothesis.

In the high-dimensional case, the dimensionality of the data can be relatively large compared to the sample size and traditional statistical theory cannot be easily adapted. Under the assumption that there exist $q_0 < p < n$ fixed components, Kritchman and Nadler [12] considered the MDL estimator developed in [26]. They show that MDL fails to detect the signal at low signal-to-noise ratios and hence underestimates the signal at small sample sizes; they then present a new estimator that improves the detection rate. Nevertheless, they only prove the consistency of their estimator under the scenario in which p is fixed and $n \to \infty$.

One of the contributions of our paper is the study of the asymptotic distribution of the partial maximum likelihood ratio statistic for the case in which $p, n \to \infty, p/n \to y \in (0, 1)$. This allows us to present a sequential test to determine the dimension of the population spike subspace. Also, as a bonus, it paves the way to correct the penalty term in Wax–Kailath's MDL estimator of the true dimension and then prove its consistency in this high-dimensional scenario.

We also address the problem for $p > n$. In some applications one can find situations in which the number of variables exceeds the number of observations ($y > 1$). Suppose that we have multiple time series and, given a window in time, we look for a small number of factors that contain most of the relevant information. In principle, we could take a big window (large n) to estimate the covariance matrix. Financial time series, for example, change frequently (they could even be non-stationary) leading us to believe that bigger time windows do not help in the understanding of the current structure. To attack a situation of this sort we would need to develop a similar test for the case $p \geq n, p/n \to y \in [1, \infty)$. In this case the maximum likelihood ratio statistic is not defined; see [7]. However, we motivate a new definition by switching the rows and columns in the data matrix. We find its asymptotic distribution and extend the definition and consistency of the MDL criteria to this case. It should be noted that the case $d = 0$ was already done by Srivastava [23].

This paper is organized as follows: Section 2 presents the asymptotic distribution of the maximum likelihood ratio statistic which is used in Section 3 to define the sequential test. Section 4 illustrates the results using simulated scenarios. The power of the test is found in Section 5. Finally, Section 6 builds on the analysis from Section 5 to improve on the way to estimate the true dimension in a consistent way and Section 7 concludes. All proofs are relegated to the Appendix.

The following notation and definitions will be used in our exposition. For positive integers m and n, $\mathbb{R}^{m\times n}$ stands for the class of all matrices of dimension $m \times n$. For a square matrix $\mathbf{A}$, $|\mathbf{A}|$ indicates its determinant. We will use the operator vec : $\mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ which vectorizes an arbitrary matrix by stacking its columns. Let $\mathbf{A} \otimes \mathbf{B}$ denote the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$. We will use $\mathbf{S} \sim \mathcal{W}_p(m, \mathbf{\Sigma})$ to denote that $\mathbf{S}$ follows a Wishart distribution with m degrees of freedom and scale matrix $\mathbf{\Sigma}$, i.e., $\mathbf{S} = \mathbf{X}^\top\mathbf{X}$ where $\mathbf{X} \in \mathbb{R}^{m\times p}$ has independent rows following a normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}$. We write $\chi^2(f)$ for the chi-square distribution with f degrees of freedom. The multivariate Gamma function is defined as $\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\{x - 1/2(j-x)\}$ for a complex number x with $\mathrm{Re}(x) > 1/2(p-1)$, where $\Gamma(x)$ is the ordinary Gamma function; see p. 62 of [15].

2. Asymptotic distribution of the maximum likelihood ratio statistic for partial sphericity

For $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, with $\mathbf{X} \in \mathbb{R}^p$, the *sphericity test* is given by

$$\mathcal{H}_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}_p \quad \text{vs.} \quad \mathcal{H}_a : \mathbf{\Sigma} \neq \sigma^2\mathbf{I}_p \tag{1}$$

with unknown σ. The maximum likelihood ratio test statistic to test the null hypothesis (1) was first derived by Mauchly [14] as the power $n/2$ of

$$LRT_0 = |\widehat{\mathbf{\Sigma}}|\{\,\mathrm{tr}(\widehat{\mathbf{\Sigma}})/p\}^{-p}, \tag{2}$$

where $\widehat{\mathbf{\Sigma}}$ is the sample covariance matrix of the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$, defined as $\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top/(n-1)$. Gleser [7] shows that the maximum likelihood ratio statistic exists only when $p \leq n-1$ and that the test with the rejection region $\{LRT_0 \leq c_\alpha\}$ (where c_α is chosen so that the test has a significance level of α) is unbiased. The choice of c_α follows from the classical asymptotic result [see 15, Theorem 8.3.7] to the effect that under $\mathcal{H}_0$ with p fixed

$$-(n-1)\rho \ln(LRT_0) \rightsquigarrow \chi^2(f)$$

as $n \to \infty$, where $\rightsquigarrow$ denotes convergence in distribution. Here

$$\rho = 1 - \frac{2p^2 + p + 2}{6(n-1)p} \quad \text{and} \quad f = \frac{1}{2}(p-1)(p+2).$$

The quantity $\rho = \rho_n \to 1$ is a correction term to improve the convergence rate when the sample is finite. For the high-dimensional case (p, n big) it was proved in [10] that the probability of wrongly rejecting the null hypothesis goes to 1 as p increases and therefore the classical test can fail completely. To overcome this problem, in their Theorem 1 they found the asymptotic distribution of (2) under the null hypothesis $\mathcal{H}_0$ when p as well as n grow in such a way that $p < n - 1$ and $p/n \to y \in (0, 1]$. Based on these results they find a rejection region for the test that has asymptotic significance level α for a given α.

If the null hypothesis is not rejected, we conclude that $\mathbf{\Sigma}$ is a constant times the identity or, equivalently, in terms of principal components, that no reduction in dimension can be achieved by transforming to principal components with lower dimension. If this null hypothesis is rejected, it is still possible, for example, for the $p - 1$ smallest eigenvalues to be equal. In this case, if their common value is small compared to the largest eigenvalue, most of the variation in the sample is explained by just the first principal component, giving a substantial reduction in dimension. Hence, it is reasonable to consider the null hypothesis that the $p - 1$ smallest eigenvalues of $\mathbf{\Sigma}$ are equal. If this is rejected, we can test whether the $p-2$ smallest eigenvalues are equal, and so on. Then in practice we test sequentially the null hypotheses

$$\mathcal{H}_d : \lambda_{d+1} = \cdots = \lambda_p, \tag{3}$$

for all $d \in \{0, \ldots, p-2\}$, where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{\Sigma}$. The null hypothesis $\mathcal{H}_d$ is equivalent to having $\mathbf{\Sigma} = (\mathbf{\Psi}, \mathbf{\Psi}_0)\mathbf{\Lambda}(\mathbf{\Psi}, \mathbf{\Psi}_0)^\top = \mathbf{\Psi}\mathbf{\Lambda}_d\mathbf{\Psi}^\top + \sigma^2\mathbf{\Psi}_0\mathbf{\Psi}_0^\top$, where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \overset{s_1}{\ldots}, \lambda_1, \ldots, \lambda_h, \overset{s_h}{\ldots}, \lambda_h, \sigma^2, \overset{p-d}{\ldots}, \sigma^2)$, $\mathbf{\Lambda}_d$ is the truncated matrix obtained by deleting the last $p - d$ rows and columns of $\mathbf{\Lambda}$, $d = s_1 + \cdots + s_h$, $\lambda_1 > \cdots > \lambda_h > \sigma^2$, $\mathbf{\Psi} = (\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_h) \in \mathbb{R}^{p\times d}$ semi-orthogonal with $\mathbf{\Psi}_i \in \mathbb{R}^{p\times s_i}$ and $\mathbf{\Psi}_0$ the semi-orthogonal complement of $\mathbf{\Psi}$ of dimension $p \times (p-d)$. If $\mathcal{H}_d$ is true, we say that the population covariance matrix $\mathbf{\Sigma}$ has d spike eigenvalues, or that the dimension of the spike subspace, the span of the columns of $\mathbf{\Psi}$, is d. As in the case of sphericity, this kind of test was much studied in the multivariate literature for p fixed and n growing. More specifically, the test called partial sphericity for the $\mathcal{H}_d$ hypothesis is based on the statistic [see 15, Theorem 9.6.1]

$$LRT_d \;=\; \frac{\hat{\lambda}_{d+1} \times \cdots \times \hat{\lambda}_p}{\left(\frac{1}{p-d}\sum_{i=d+1}^{p} \hat{\lambda}_i\right)^{p-d}}, \tag{4}$$

where $\hat{\lambda}_i$ are the eigenvalues (in decreasing order) of $\hat{\mathbf{\Sigma}}$ (the sample covariance matrix). Let us remark that $0 < LRT_d \leq LRT_{d+1} \leq 1$, which is easy to see from the fact that $LRT_d^{1/(p-d)}$ is the ratio between the geometric and arithmetic means. It is known that the maximum likelihood ratio test (4) is well defined only when $m = n - 1 \geq p$, as in the case of $d = 0$, and it was proved by Lawley [13] (later improved by James [9]) that the asymptotic distribution of (4) under $\mathcal{H}_d$ is

$$-\rho \ln(LRT_d) \rightsquigarrow \chi^2_{(p-d+2)(p-d-1)/2}, \tag{5}$$

as n increases with p fixed, where $\hat{\sigma}^2 = \sum_{i=d+1}^{p} \hat{\lambda}_i/(p-d)$ and

$$\rho = \; m - d - \frac{2(p-d)^2 + (p-d) + 2}{6(p-d)} + \sum_{i=1}^{d} \frac{\hat{\sigma}^2}{(\hat{\lambda}_i - \hat{\sigma}^2)^2}.$$

This asymptotic distribution makes it possible to define a test that has asymptotic significance level α using the rejection region $\{LRT_d < c_\alpha\}$. As in the case of the sphericity test, the result is no longer true when p/m is large.

Going back to the case $\mathcal{H}_0 : d = 0$, even when the maximum likelihood ratio test is not well defined when $p > m$, it was pointed out in [23] that we can still build a sphericity test. Namely, for $p > m$, under $\mathcal{H}_0$, $\mathbf{W} = m\widehat{\mathbf{\Sigma}} \sim \mathcal{W}_p(m, \sigma^2 \mathbf{I}_p)$ and therefore $\mathbf{W} = \mathbf{Y}^\top \mathbf{Y}$ with $\mathbf{Y} \in \mathbb{R}^{m \times p}$ independent normals with mean 0 and variance σ^2. Then $\widetilde{\mathbf{W}} = \mathbf{Y}\mathbf{Y}^\top \sim \mathcal{W}_m(p, \sigma^2 I_m)$ with $p > m$, and one can build the maximum likelihood ratio test for $\widetilde{\mathbf{W}}$ as was done for $\mathbf{W}$ for the case $p < m$. Since the non-zero eigenvalues of $\widetilde{\mathbf{W}}$ and $\mathbf{W}$ coincide, we get the maximum likelihood ratio test, under the null hypothesis $\mathcal{H}_0$, for $\widetilde{\mathbf{W}}$ as

$$LRT_0 = \frac{\hat{\lambda}_1 \times \cdots \times \hat{\lambda}_m}{\left(\frac{1}{m}\sum_{i=1}^{m} \hat{\lambda}_i\right)^m}.$$

Using [10, Theorem 1] we get the asymptotic distribution for LRT_0, under $\mathcal{H}_0$, with m and p exchanging their roles. As a consequence, a test with asymptotic significance level α can be built. Let us note that in the case of [23], the approximation is given in terms of a χ^2 distribution, while in [10] and here, a normal approximation is given.

This gives the motivation to define the maximum likelihood ratio test for partial sphericity, for the case of $p > m + d$, as

$$LRT_d = \frac{\hat{\lambda}_{d+1} \times \cdots \times \hat{\lambda}_m}{\left(\frac{1}{m-d}\sum_{i=d+1}^{m} \hat{\lambda}_i\right)^{m-d}}.$$

In order to build a test that has asymptotic level α for the case of p and m increasing under partial sphericity, and with $p < m$ or $p > m + d$, we first need to find the asymptotic distribution of the LRT_d in these cases. The following proposition gives the asymptotic distribution when p, m grow to infinity and $p/m \to y$ with $y \in (0, \infty)$. In the rest of paper we will assume the condition Q_0:

Condition Q_0 : There exists $q_0 \ll \min(p, m)$ independent of p and m such that $\mathcal{H}_d$ defined in (3) is true for $d \leq q_0$.

Proposition 1. *Let $W = m\hat{\mathbf{\Sigma}} \sim \mathcal{W}_p(m, \mathbf{\Sigma})$ and let us assume Condition Q_0. Under the null hypothesis that the true number of spikes is d fixed, i.e.,*

$$\mathcal{H}_d : \lambda_{d+1} = \cdots = \lambda_p,$$

the asymptotic distribution of LRT_d (when m and p grow and $p/m \to y > 0$) is given by

a) Case $p < m$:

$$\frac{\ln LRT_d - \mu_{m,p,d}}{\sigma_{m,p,d}} \rightsquigarrow \mathcal{N}(0, 1),$$

where

$$\mu_{m,p,d} = \widetilde{\mu}_{m,p} + \ln A_{m,p,d} + \ln B_{m,p,d}, \quad \sigma^2_{m,p,d} = -2\left\{\frac{p-d}{m} + \ln\left(1 - \frac{p}{m}\right)\right\},$$

with

$$\begin{aligned} \widetilde{\mu}_{m,p} &= -p - (m - p - 1/2)\ln(1 - p/m), \\ A_{m,p,d} &= \prod_{i=1}^{h} \lambda_i^{s_i} / \prod_{i=1}^{d} \hat{\lambda}_i, \\ B_{m,p,d} &= \left(1 + \frac{\sum_{i=1}^{d} \hat{\lambda}_i - \sum_{i=1}^{h} s_i \lambda_i}{\sum_{i=d+1}^{p} \hat{\lambda}_i}\right)^{p-d}. \end{aligned}$$

b) Case $p > m + d$:

$$\frac{\ln(LRT_d) - \mu_{m,p,d}}{\sigma_{m,p,d}} \rightsquigarrow \mathcal{N}(0,1),$$

where

$$\mu_{m,p,d} = \widetilde{\mu}^*_{m,p,d} + \ln B^*_{m,p,d} + \ln C^*_{m,p,d} + \ln D^*_{m,p,d}$$

and

$$\sigma^2_{m,p,d} = -2\left\{\frac{m}{p-d} + \ln\left(1 - \frac{m}{p-d}\right)\right\},$$

with

$$\begin{aligned}
\widetilde{\mu}^*_{m,p,d} &= -m - (p-d-m-1/2)\ln\{1 - m/(p-d)\}, \\
B^*_{m,p,d} &= \left(1 + \frac{\sum_{i=1}^{d}\hat{\lambda}_i - \sum_{i=1}^{h} s_i\lambda_i}{\sum_{i=d+1}^{m}\hat{\lambda}_i}\right)^{m-d}\left(\frac{m-d}{m}\right)^{m-d}, \\
C^*_{m,p,d} &= \left\{\frac{\sigma^2(p-d)}{m}\right\}^d, \\
D^*_{m,p,d} &= \prod_{i=1}^{h}\left(1 + \frac{\lambda_i}{\sigma^2}\frac{m}{p-d-m-1}\right)^{s_i} / \prod_{i=1}^{d}\hat{\lambda}_i.
\end{aligned}$$

The proof of Proposition 1 can be found in Appendix A.

Remark 1. Proposition 1 contains the sphericity test studied by Jiang and Yang [10] for $m > p$. Moreover, if $p > m$, under the hypothesis of sphericity we get that the new maximum likelihood ratio test has the same asymptotic distribution as in the case in [10], except that p and m change their roles. This should not be a surprise, since $\mathbf{W} \sim \mathcal{W}_p(m, \sigma^2\mathbf{I}_p)$ can be written as $\mathbf{W} = \sigma^2\mathbf{X}^\top\mathbf{X}$ with $\mathbf{X} \in \mathbb{R}^{m\times p}$ filled with independent standard normals. From here, defining $\widetilde{\mathbf{W}} = \sigma^2\mathbf{X}\mathbf{X}^\top$, we get $\widetilde{\mathbf{W}} \sim \mathcal{W}_m(p, \sigma^2\mathbf{I}_m)$. The non-zero eigenvalues of $\mathbf{W}$ and $\widetilde{\mathbf{W}}$ are the same, and therefore the usual definition of the likelihood ratio test for $\widetilde{\mathbf{W}}$ coincides with the definition of the maximum likelihood ratio test for $\mathbf{W}$. As a consequence, the asymptotic distribution of the LRT for $p > m$ under $d = 0$ follows from the result of [10], exchanging the roles of m and p.

Remark 2. Let us note that $\mu_{m,p,d}$ is a random variable that depends on the true values of σ^2, $\lambda_1, \ldots, \lambda_d$ and of the first d sample eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$

of $\hat{\mathbf{\Sigma}}$. To be able to use the asymptotic distribution of Proposition 1 to test for the dimension of the spike subspace, we need to replace the true values by consistent estimators. The parameter σ^2 can be replaced by its consistent estimator $\hat{\sigma}^2 = \sum_{i=d+1}^{p} \hat{\lambda}_i/(p-d)$. For λ_i it is well known that, in the limit, the estimates of the spike eigenvalues experience a phase transition, [3]. Indeed, if $\lambda > \sigma^2(1+\sqrt{y})$, then

$$\hat{\lambda} \to \lambda \left(1 + \frac{y\sigma^2}{\lambda - \sigma^2}\right), \tag{6}$$

whereas for eigenvalues λ which are in the range $(\sigma^2, \sigma^2(1+\sqrt{y})]$, the limit becomes $\sigma^2(1+\sqrt{y})^2$, making them invisible, i.e., indistinguishable from the bulk since the sample eigenvalues corresponding to a fixed number of eigenvalues of the bulk go to the same $\sigma^2(1+\sqrt{y})^2$; see [2]. Therefore we cannot directly replace λ_i by $\hat{\lambda}_i$ in $\mu_{m,p,d}$ since, even if λ_i is greater than the threshold $\sigma^2(1+\sqrt{y})$, the $\hat{\lambda}_i$ are biased estimators of λ_i. We do know, however, the bias of the estimator exactly from (6). Therefore we can substitute the $\lambda_i \geq \sigma^2(1+\sqrt{y})$ in the expression for $\mu_{m,p,d}$ using the equation suggested by Eq. (6):

$$\hat{\lambda}_i = \lambda_i \left(1 + \frac{p}{m} \frac{\hat{\sigma}^2}{\lambda_i - \hat{\sigma}^2}\right)$$

to get

$$\widetilde{\lambda}_i = \frac{1}{2}\left\{\hat{\lambda}_i + \hat{\sigma}^2 - \hat{\sigma}^2 \frac{p}{m} + \sqrt{-4\hat{\lambda}_i\hat{\sigma}^2 + \left(\hat{\lambda}_i + \hat{\sigma}^2 - \hat{\sigma}^2 \frac{p}{m}\right)^2}\right\}, \tag{7}$$

a consistent estimator for λ_i. In the limit, the discriminant will be non-negative if and only if $\lambda_i \geq \sigma^2(1+\sqrt{y})$. Now, the sample version of the discriminant can be negative when the true eigenvalue is close to the threshold (or less than the threshold). In that case we will consider $\widetilde{\lambda}_i = \hat{\sigma}^2(1+\sqrt{p/m})$ since that is the value of $\widetilde{\lambda}_i$ for $\hat{\lambda}_i$ that makes the discriminant be zero. Replacing λ_i in $\mu_{m,p,d}$ by $\widetilde{\lambda}_i$ and σ^2 by $\hat{\sigma}^2$, we get a new approximation for the asymptotic distribution, when all the spike eigenvalues are greater than the threshold, that can be used for testing. On the other hand, if one or more spike eigenvalues are less than the threshold, this new asymptotic distribution will give a test with asymptotic level not greater than α. (See Lemma 1).

Summarizing, we have the following corollary.

Corollary 1. *Under the hypothesis of Proposition 1, if the spike eigenvalues* $\lambda_1, \ldots, \lambda_d$ *are all greater than the threshold* $\sigma^2(1+\sqrt{y})$,

$$\frac{\ln(LRT_d) - \hat{\mu}_{m,p,d}}{\sigma_{m,p,d}} \rightsquigarrow \mathcal{N}(0,1),$$

where $\hat{\mu}_{m,p,d}$ *is obtained from* $\mu_{m,p,d}$ *by replacing* λ_i *with* $\tilde{\lambda}_i$ *defined on* (7) *nad* σ^2 *with* $\hat{\sigma}^2$.

3. Test to find the dimension d of the spike subspace

As we pointed out in Section 2, the test will be done sequentially as in the usual case. We consider the null hypothesis that $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$. If this is rejected, we can test whether the $\min(p-1, m-1)$ smallest eigenvalues are equal, and so on, i.e., we test sequentially the null hypotheses for each $d \in \{0, \ldots, q_0\}$ when Condition Q_0 is true.

For the test to have significance level α, the rejection region will be the set $\{LRT_d < C\}$ where C is chosen such that $\Pr_{\mathcal{H}_d}(LRT_d < C) = \alpha$ and where C depends on α and can depend on the sample. Using Corollary 1 we can build the asymptotic test with rejection region

$$\left\{ \frac{\ln(LRT_d) - \hat{\mu}_{m,p,d}}{\sigma_{m,p,d}} < z_\alpha \right\}, \tag{8}$$

where $\hat{\mu}_{m,p,d}$ was defined in Corollary 1, $\sigma_{m,p,d}$ was defined in Proposition 1, and z_α is the α quantile of the normal distribution. This test will have asymptotic significance level α when all the true spike eigenvalues are greater than $\sigma^2(1+\sqrt{y})$. On the other hand, if one or more spike eigenvalues are less than the threshold, the test with rejection region (8) will have significance level not greater than α and therefore will be a conservative test. Summarizing, we have the following lemma, whose proof is in Appendix B.

Lemma 1. *The test for the hypothesis*

$$\mathcal{H}_d : \lambda_{d+1} = \cdots = \lambda_p \quad vs. \quad \mathcal{H}_1 : \lambda_{d+2} = \ldots = \lambda_p$$

with rejection region defined in (8) *has significance level* $\Phi(z_\alpha + L)$ *with*

$$L = \begin{cases} \frac{1}{\sqrt{-2\{y+\ln(1-y)\}}} \sum_{i \in J_2} s_i \left\{ \frac{\lambda_i}{\sigma^2} - (1+\sqrt{y}) - \ln \frac{\lambda_i}{\sigma^2(1+\sqrt{y})} \right\} & \text{when } p < m, \\ \frac{1}{\sqrt{-2\{1/y+\ln(1-1/y)\}}} \sum_{i \in J_2} s_i \left\{ \frac{\lambda_i}{y\sigma^2} - \frac{1+\sqrt{y}}{y} - \ln \frac{\sigma^2(y-1)+\lambda_i}{\sqrt{y}\sigma^2(\sqrt{y}+1)} \right\} & \text{when } p > m+d, \end{cases}$$

where $J_1 = \{i \leq h : \lambda_i > \sigma^2(1+\sqrt{y})\}$, $J_2 = \{i \leq h : \lambda_i \in (\sigma^2, \sigma^2(1+\sqrt{y})]\}$, Φ *is the cumulative distribution function of* $\mathcal{N}(0,1)$ *and* z_α *is its* α*th quantile.*

Remark 3. Since $L \leq 0$, without information about the population spike eigenvalues, the test defined in Lemma 1 has asymptotic significance level smaller than α. When all the spike eigenvalues are greater than the threshold, $L = 0$ and the test will have asymptotic significance level α.

4. Simulations

4.1. Set of simulations to show the behavior of the asymptotic approximations

In this section we show the behavior of the asymptotic distribution given in Proposition 1 as well as the asymptotic approximation distribution presented in Corollary 1 that we used for testing when the null hypothesis is true. We do this for a variety of m, p as well as d. Let us recall that these two distributions are asymptotically equivalent when all the spike eigenvalues are greater than the threshold. In order to do this we have chosen the scenarios used in [17]. They consider models with spike subspace dimensions $d = 4$ and $d = 5$ and spike eigenvalues $(7, 6, 5, 4)$ and $(259.72, 17.97, 11.04, 7.88, 4.82)$, respectively. In both cases σ^2 is chosen to be 1. We note that in both of these scenarios, all the spike eigenvalues are bigger than the threshold $\sigma^2\,(1+\sqrt{y})$. In all the simulations, we assume that as p and m grow, their ratio is constant and therefore equal to its limit y.

In Figure 2 we show (red lines) the asymptotic distributions from Proposition 1 and (blue lines) the approximation distributions described in Section 3 for the above settings for $p/m = .3$ and $p/m = .6$ (i.e., for $m > p$). We can see how the behavior of the asymptotic distributions improves as p and m increase in both cases $p/m = .3$ and $p/m = .6$.

We have also run similar simulations for some cases in which $p > m$ (the results can be found in the supplementary material). The results are essentially the same. The rates of convergence to the true distribution, however, slow down as p/m increases. Further simulation results reported in the supplementary material show that the approximating distributions improve when the spike eigenvalues are further away from the critical value, $\sigma^2\,(1+\sqrt{y})$, the exact and the asymptotic distributions becoming almost indistinguishable.

An interesting point already noticed by Jiang and Yang [10] for the case $d = 0$ is that the classical chi-square approximation (5) becomes poorer as p

becomes large relative to m. An illustration is provided in the supplementary material.

In addition to the those presented, we have run simulations using non-normal distributions (Student's, chi, and uniform) and obtained unsurprising results. In the fat-tailed cases (Student's with 4 degrees of freedom, for example) the test has a slight tendency to overshoot, whereas in the non-fat-tailed cases (uniform) the deviations in the results are on the undershooting side.

4.2. Sequential test for the dimension of the spike subspace. Simulations

Several methods have been recently proposed using random matrix theory for determining the number of factors for high-dimensional data. These contributions come from different fields. Among others, we can cite [8] or [16] in economics, and [12] in the array processing or chemometrics literature. A review and an up to date method for the high-dimensional case is presented by [17], based on recent results from the theory of random matrices [1, 5, 19].

As we did in the previous subsection, we consider the models they use to check our results. In both, the ratios used by them are $p/m = .3$ and $p/m = .6$. We present the results of our iterative procedure in Tables 1–4. Tables 1–2 should be compared with their Tables 1–2 and Tables 3–4 to their Table 3. In every case we run 1000 replications. In addition, the variance is not assumed to be known and it is estimated as the average of the remainder eigenvalues.

As we can see our method is very competitive when $p/m = .3$ but its performance deteriorates for higher ratios. We have run the same scenarios for lower values of p/m which, for the sake of space, we have included in the supplementary material. Based on those we can confirm that this behavior (performance getting worse as p/m increases) persists. The problem is that even if the asymptotic distribution under the null hypothesis is almost perfect, as we saw in Section 4.1, the sequential likelihood ratio test underestimates the dimension of the spike subspace. Nevertheless, the maximum likelihood ratio test chooses, in close to 95% of the cases, either the true dimension or a value that is lower than the true dimension.

Clearly, if we were to test $\mathcal{H}_d$: true dimension $= d$ vs. $\mathcal{H}_a$: true dimension $> d$ in all the cases, the maximum likelihood ratio test would not reject the null hypothesis 95% of the cases, as expected. However, since the test is sequential, the problem is that it can get stuck in a dimension smaller than the true one. As we will see in the next section, this is due to the fact that the power

of the likelihood ratio test decreases when p/m grows to 1 (case $p < m$) or decreases to 1 (case $p > m$), a phenomenon already seen by Jiang and Yang [10] in their Table 1, for $p < m$ and p/m growing to 1. To overcome the problem of underestimating the dimension, we will present, in Section 6, a potential solution based on a more detailed study of the behavior of the maximum likelihood ratio statistic under the alternative hypothesis.

	1	2	3	4	5	6	7	8	9
$(30, 100)$	0.000	0.000	0.000	0.007	0.981	0.010	0.002	0.000	0.000
$(60, 200)$	0.000	0.000	0.000	0.000	0.983	0.013	0.003	0.001	0.000
$(120, 400)$	0.000	0.000	0.000	0.000	0.981	0.016	0.003	0.000	0.000
$(240, 800)$	0.000	0.000	0.000	0.000	0.959	0.027	0.010	0.002	0.002

Table 1: Values of d picked for spikes $= (259.72, 17.97, 11.04, 7.88, 4.82)$ (true dimension $= 5$) and $p/m = .3$

	1	2	3	4	5	6	7	8	9
$(60, 100)$	0.000	0.000	0.000	0.216	0.735	0.031	0.008	0.003	0.006
$(120, 200)$	0.000	0.000	0.000	0.184	0.762	0.033	0.012	0.004	0.003
$(240, 400)$	0.000	0.000	0.000	0.167	0.798	0.025	0.006	0.003	0.001
$(480, 800)$	0.000	0.000	0.000	0.126	0.841	0.022	0.007	0.003	0.001

Table 2: Values of d picked for spikes $= (259.72, 17.97, 11.04, 7.88, 4.82)$ (true dimension $= 5$) and $p/m = .6$

	1	2	3	4	5	6	7	8	9
$(30, 100)$	0.000	0.000	0.039	0.901	0.036	0.014	0.002	0.000	0.001
$(60, 200)$	0.000	0.000	0.014	0.932	0.041	0.009	0.004	0.000	0.000
$(120, 400)$	0.000	0.000	0.006	0.935	0.042	0.011	0.004	0.001	0.000
$(240, 800)$	0.000	0.000	0.006	0.949	0.035	0.009	0.001	0.000	0.000

Table 3: Values of d picked for spikes $= (7, 6, 5, 4)$ (true dimension $= 4$) and $p/m = .3$

	1	2	3	4	5	6	7	8	9
$(60, 100)$	0.000	0.018	0.482	0.452	0.028	0.011	0.004	0.003	0.001
$(120, 200)$	0.000	0.011	0.427	0.518	0.028	0.008	0.006	0.000	0.001
$(240, 400)$	0.000	0.002	0.368	0.583	0.027	0.012	0.007	0.001	0.000
$(480, 800)$	0.000	0.002	0.352	0.605	0.024	0.013	0.003	0.001	0.000

Table 4: Values of d picked for spikes $= (7, 6, 5, 4)$ (true dimension $= 4$) and $p/m = .6$

5. Asymptotic distribution of LRT_d when the true dimension is d_1

We have studied the asymptotic distribution of the statistic under the null hypothesis. To go more deeply into the understanding of its behavior, we would like to know the distribution of the statistic when we are not considering the correct (true) number of spikes. Throughout this section we will assume that the true number of spikes is d_1 and the statistic under consideration is LRT_d where d is less than or greater than d_1.

Proposition 2. *Let us assume Condition Q_0. The asymptotic distribution as $p/m \to y$ of the maximum likelihood ratio test LRT_d when the true model is spiked of dimension $d_1 \neq d$ is given by*

$$\frac{\ln(LRT_d) - \mu_{m,p,d,d_1}}{\sigma_{m,p,d_1}} \rightsquigarrow \mathcal{N}(0,1),$$

where

a) for $p < m$, μ_{m,p,d,d_1} is equal to μ_{m,p,d_1} plus

$$\begin{cases} \sum_{i=h_0+1}^{h_1} s_i(\ln k_i - k_i + 1) \text{ when } d < d_1, \\ (d - d_1)\left\{(1+\sqrt{y})^2 - \ln(1+\sqrt{y})^2 - 1\right\} \text{ when } d_1 < d \leq q_0, \end{cases} \tag{9}$$

b) for $p > m + d_1$, μ_{m,p,d,d_1} is equal to μ_{m,p,d_1} plus

$$\begin{cases} \sum_{i=h_0+1}^{h_1} s_i\left(\ln k_i/y - k_i/y + 1\right) \text{ when } d < d_1, \\ (d - d_1)\left\{\frac{(1+\sqrt{y})^2}{y} - \ln\frac{(1+\sqrt{y})^2}{y} - 1\right\} \text{ when } d_1 < d \leq q_0. \end{cases} \tag{10}$$

Anytime, μ_{m,p,d_1} is defined as in Proposition 1 but replacing d by d_1, $k_i = \lambda_i/\sigma^2\{1 + y\sigma^2/(\lambda_i - \sigma^2)\}$ if $i \in J_{1,k}$ and $k_i = (1+\sqrt{y})^2$ if $i \in J_{2,k}$ or when $\lambda_i = \sigma^2$ where for $k = 0, 1$, $J_{1,k} = \{i \leq h_k : \lambda_i > \sigma^2(1+\sqrt{y})\}$, $J_{2,k} = \{i \leq h_k : \lambda_i \in (\sigma^2, \sigma^2(1+\sqrt{y})]\}$, and h_0, h_1 are such that $d = s_1 + \cdots + s_{h_0}$ and $d_1 = s_1 + \cdots + s_{h_1}$.

The proof can be found in Appendix C.

5.1. The power of the test

The next proposition gives the asymptotic power of the maximum likelihood ratio test for the hypothesis $\mathcal{H}_d$: the spike subspace has dimension d vs. H_a : the spike subspace has dimension greater than d for the case that the specific alternative hypothesis true model has a spike subspace of dimension $d_1 > d$ and all the spike eigenvalues are greater than the threshold.

Proposition 3. *Let us assume Condition Q_0. The asymptotic power of the maximum likelihood ratio test for the hypothesis $\mathcal{H}_d$: spike subspace of dimension d vs. H_a : spike subspace has dimension greater than d, for the case that the specific alternative hypothesis true model has $d_1 \in (d, q_0]$ spikes and all the spike eigenvalues are greater than the threshold, is given as follows:*

1. *Case $p < m$*

$$\psi(d_1) = \Phi\left[\frac{\sum_{i=h_0+1}^{h_1} s_i\{\frac{\lambda_i}{\sigma^2} - \ln(\frac{\lambda_i}{\sigma^2}) - 1\} + z_\alpha \sigma_{m,p,d}}{\sigma_{m,p,d_1}}\right]$$

2. *Case $p > m + d_1$*

$$\psi(d_1) = \Phi\left[\frac{\sum_{i=h_0+1}^{h_1} s_i\{\frac{\lambda_i}{y\sigma^2} - \frac{1}{y} - \ln(1 - \frac{1}{y} + \frac{\lambda_i}{y\sigma^2})\} + z_\alpha \sigma_{m,p,d}}{\sigma_{m,p,d_1}}\right]$$

where Φ is the cumulative standard normal distribution and z_α is the α quantile of the standard normal.

The proof of this proposition can be found in Appendix D.

	1	2	3
$(30, 100)$	1.000	1.000	0.884
$(60, 200)$	1.000	1.000	0.978
$(120, 400)$	1.000	1.000	0.997
$(240, 800)$	1.000	1.000	0.996

Table 5: Probability of rejecting, spikes $= (7, 6, 5, 4)$ (true dimension $= 4$) and $p/m = .3$

	1	2	3
$(30, 100)$	1.000	0.976	0.408
$(60, 200)$	1.000	0.996	0.543
$(120, 400)$	1.000	0.995	0.613
$(240, 800)$	1.000	0.998	0.595

Table 6: Probability of rejecting, spikes $= (7, 6, 5, 4)$ (true dimension $= 4$) and $p/m = .6$

	1	2	3	4
$(30, 100)$	1.000	1.000	1.000	0.987
$(60, 200)$	1.000	1.000	1.000	1.000
$(120, 400)$	1.000	1.000	1.000	1.000
$(240, 800)$	1.000	1.000	1.000	1.000

Table 7: Probability of rejecting, spikes $= (259.72, 17.97, 11.04, 7.88, 4.82)$ (true dimension $= 5$) and $p/m = .3$

	1	2	3	4
$(30, 100)$	1.000	1.000	1.000	0.703
$(60, 200)$	1.000	1.000	1.000	0.784
$(120, 400)$	1.000	1.000	1.000	0.848
$(240, 800)$	1.000	1.000	1.000	0.872

Table 8: Probability of rejecting, spikes $= (259.72, 17.97, 11.04, 7.88, 4.82)$ (true dimension $= 5$) and $p/m = .6$

Let us note that for fixed y, as m and p grow, $\sigma_{m,p,d}/\sigma_{m,p,d_1} \to 1$ when $p/m \to y$. Now, the bigger the eigenvalues λ_i with $i \in \{h_0 + 1, \ldots, h_1\}$, the bigger $\sum_{i=h_0+1}^{h_1} s_i\{\lambda_i/\sigma^2 - \ln(\lambda_i/\sigma^2) - 1\} > 0$ and $\sum_{i=h_0+1}^{h_1} s_i[\lambda_i/(y\sigma^2) - 1/y - \ln\{1 - 1/y + \lambda_i/(y\sigma^2)\}] > 0$. As a consequence, larger values of λ_i imply a greater power. Moreover $\lambda_i \to \infty$ for $i \in \{h_0 + 1, \ldots, h_1\}$ implies that the power goes to 1. On the other hand, for fixed λ_i with $i \in \{h_0 + 1, \ldots, h_1\}$ and $y \to 1$, we have that $\sigma_{m,p,d_1} \to \infty$, $\sigma_{m,p,d}/\sigma_{m,p,d_1} \to 1$ and, therefore, the power decreases to α as $y \to 1$. For $y \to 0$, $\sigma_{m,p,d_1} \to 0$, $\sigma_{m,p,d}/\sigma_{m,p,d_1} \to 1$ and in this case the power goes to 1.

Tables 5–8 show the probabilities of rejection for each of the values under the true dimension for the two models considered in our simulation runs. We see how, not surprisingly, the power decreases as p/m grows closer to 1.

In light of Proposition 3 and the explanation above, the results obtained in Section 4.2 (and in the supplementary material) should not be surprising. We saw that the maximum likelihood ratio test underestimates the true dimension of the spike subspace when the limit of p/m is close to 1. This is confirmed by the behavior of the power of the test when $y \approx 1$. The consequence of this is that the sequential test stops earlier than it is supposed to. In spite of this, we do know the asymptotic behavior of the statistic as a function of the null hypothesis (thanks to Proposition 2). We can then use this knowledge to modify the statistic by penalizing the number of spikes chosen.

Remark 4. Passemier and Yao [17] present a method to test for the dimension of the spike subspace that is based on eigenvalue spacings. Due to the nature of the technique, dealing with multiple spike eigenvalues can be tricky. In [18] they prove that due to the different speeds of convergence of the spacings between the spike and the bulk eigenvalues, their method still works for matrices with repeated eigenvalues, even if at a different, slower,

rate of convergence. The maximum likelihood ratio does not run into these problems since, as will become clear in the results from the next section, the important quantity we look at is related to the value of the eigenvalue itself and not to their spacings.

6. A penalized version of the maximum likelihood ratio test

As we mentioned in Section 2, the maximum likelihood ratio statistic (4) is an increasing function of d. The same phenomenon was observed in Proposition 2 for the asymptotic mean. But, also in Proposition 4, we have seen that the growth rate of the asymptotic mean changes from $d < d_1$ to $d > d_1$. This will allow us to define a new consistent estimator of the dimension of the spike subspace via information criteria, as was done in [26] for the fixed-p case.

Proposition 4. *Suppose that the spike subspace of the true model has dimension d_1. Given μ_{m,p,d,d_1} defined in Proposition 2 and $\epsilon \geq 0$, we consider, for $\tilde{y} = \max(1, y)$, the function $g(d) = \mu_{m,p,d,d_1} - (d - d_1)\left[h\{\sigma^2(1+\sqrt{y})\} + \epsilon\right]$ with*

$$h(\lambda) = \frac{\lambda}{\tilde{y}\sigma^2}\left(1 + \frac{y\sigma^2}{\lambda - \sigma^2}\right) - \ln\left\{\frac{\lambda}{\tilde{y}\sigma^2}\left(1 + \frac{y\sigma^2}{\lambda - \sigma^2}\right)\right\} - 1.$$

Let $\lambda^* > \sigma^2(1+\sqrt{y})$ be such that $\epsilon = h(\lambda^*) - h\{\sigma^2(1+\sqrt{y})\}$. Then, if all the spike eigenvalues are greater than λ^*, we have that g has a global maximum at $d = d_1$.

The proof can be found in Appendix E.

Remark 5. First, $h\{\sigma^2(1+\sqrt{y})\}$ is only dependent on y. Second, for $\epsilon = 0$, the function g is increasing for $d < d_1$ and constant for $d \geq d_1$. Moreover, if all the eigenvalues in the spiked part are greater than $\sigma^2(1+\sqrt{y})$, then g is strictly increasing for $d \leq d_1$.

Proposition 4 gives us a clear intuition of the behavior of the mean of the distribution of the different statistics used in the sequential test. In the population, if $\epsilon > 0$ and all the spike eigenvalues are greater than λ^*, then the mean of the maximum likelihood ratio test plus the penalty term has a

global maximum at $d = d_1$. Inspired by Proposition 4, we will define a new estimator for the dimension $\hat{d}_\epsilon$. For $\epsilon \geq 0$ fixed, we define

$$\hat{d}_\epsilon = \min \left[\arg\max_{0 \leq j \leq q_0} \left\{ \ln(LRT_j) - j[h\{\sigma^2 (1 + \sqrt{y})\} + \epsilon] \right\} \right], \qquad (11)$$

where q_0 is an upper bound for d. Note that if we knew what the true value of the dimension of the spike subspace (d_1) is, we could replace j by $j - d_1$ in the definition of $\hat{d}_\epsilon$, leaving us with an expression that is very closely related to the function g defined above. Clearly we cannot do that, since the point of defining $\hat{d}_\epsilon$ is exactly to estimate the value of d_1. For our purposes this is, however, inconsequential.

Now, since the function $h(\lambda)$ is strictly increasing, when $\lambda > \sigma^2 (1 + \sqrt{y})$, there exists a value λ^* such that $\epsilon = h(\lambda^*) - h\{\sigma^2 (1 + \sqrt{y})\}$. The idea is that this new estimator will miss the eigenvalues located between the threshold and λ^* but, with high probability, will pick up all the eigenvalues bigger than λ^* as $m, p \to \infty$. As a consequence of Proposition 4, we have

Proposition 5. *Suppose that the true dimension of the spike subspace is d_1. If all the spike eigenvalues are greater than λ^*, then $\hat{d}_\epsilon$ defined in (11) is a consistent estimator of d_1 in the sense that*

$$\Pr(\hat{d}_\epsilon = d_1) \to 1 \quad \text{as } p, m \to \infty, \quad \frac{p}{m} \to y > 0.$$

The proof can be found in Appendix F.

Remark 6. As was discussed above, the choice of ϵ will determine which eigenvalues will be detected. The method will miss any eigenvalues λ which are smaller than λ^* where λ^* satisfies $\epsilon = h(\lambda^*) - h\{\sigma^2 (1 + \sqrt{y})\}$. Therefore, the strategy for picking ϵ should be as follows: first pick a λ^* slightly bigger than $\sigma^2 (1 + \sqrt{y})$, then define ϵ as the mentioned difference. As the proposition shows, this method is consistent as long as we have chosen a λ^* which sits to the left of the smallest of the eigenvalues which are bigger than the threshold. Otherwise, if our chosen λ^* turns out to be larger than some of the relevant eigenvalues, those eigenvalues will not be picked up and the dimension estimated will be smaller than d_1.

6.1. Simulations and comparison with Kritchmaker–Nadler's method

The procedure defined in (11) gives us another way to estimate the true dimension of the spike subspace. To illustrate this estimator we have replicated simulations for the examples already presented. Tables 9–12 show the

results for the cases corresponding to Tables 1–4. It can be seen that the performance is greatly improved. Using results from the theory of random matrices pays off since they allow us to pick the penalty function in a meaningful way. More results from this fact can be found in the supplementary material.

	1	2	3	4	5	6	7	8	9
$(30, 100)$	0.000	0.000	0.000	0.000	0.995	0.005	0.000	0.000	0.000
$(60, 200)$	0.000	0.000	0.000	0.000	0.999	0.001	0.000	0.000	0.000
$(120, 400)$	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
$(240, 800)$	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000

Table 9: Values of d picked via HD-MDL for spikes $= (259.72, 17.97, 11.04, 7.88, 4.82)$ (true dimension $= 5$) and $p/m = .3$

	1	2	3	4	5	6	7	8	9
$(60, 100)$	0.000	0.000	0.000	0.000	0.968	0.032	0.000	0.000	0.000
$(120, 200)$	0.000	0.000	0.000	0.000	0.993	0.007	0.000	0.000	0.000
$(240, 400)$	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
$(480, 800)$	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000

Table 10: Values of d picked via HD-MDL for spikes $= (259.72, 17.97, 11.04, 7.88, 4.82)$ (true dimension $= 5$) and $p/m = .6$

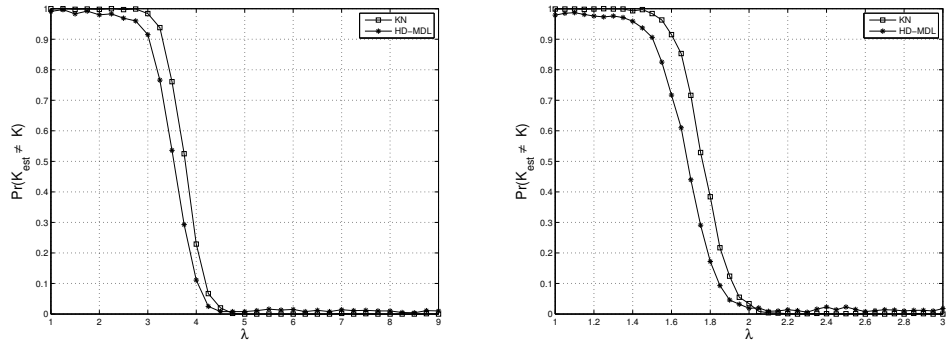
	1	2	3	4	5	6	7	8	9
$(30, 100)$	0.000	0.000	0.000	0.992	0.008	0.000	0.000	0.000	0.000
$(60, 200)$	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
$(120, 400)$	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
$(240, 800)$	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000

Table 11: Values of d picked via HD-MDL for spikes $= (7, 6, 5, 4)$ (true dimension $= 4$) and $p/m = .3$

	1	2	3	4	5	6	7	8	9
$(60, 100)$	0.000	0.000	0.002	0.971	0.027	0.000	0.000	0.000	0.000
$(120, 200)$	0.000	0.000	0.000	0.995	0.005	0.000	0.000	0.000	0.000
$(240, 400)$	0.000	0.000	0.000	0.999	0.001	0.000	0.000	0.000	0.000
$(480, 800)$	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000

Table 12: Values of d picked via HD-MDL for spikes $= (7, 6, 5, 4)$ (true dimension $= 4$) and $p/m = .6$

In addition, we run the scenarios shown in Figures 7 and 8 in [12]. The results are presented in Figure 1. In both cases we plot the probability of misdirection when there is only one spike eigenvalue. The spike value appears on the x axis. The new estimator defined is denoted by HD-MDL (high-dimensional MDL). We see how a detailed analysis of the growth of the maximum likelihood ratio statistics allows us to improve the detection performance. One caveat of our approach, however, is that we have to choose a value for ϵ. For large values of ϵ the estimator will miss eigenvalues that are close to the threshold but will minimize the probability of missing larger ones. On the other hand, very low values will increase, slightly, the probability of missing larger eigenvalues (when p and m are not sufficiently large).



$p = 2000, n = 500, d_1 = 1, \sigma^2 = 1, \epsilon = .01$ $\quad$ $p = 200, n = 8000, d_1 = 1, \sigma^2 = 1, \epsilon = .025$

Figure 1: Probabilities of misdetection in two cases.

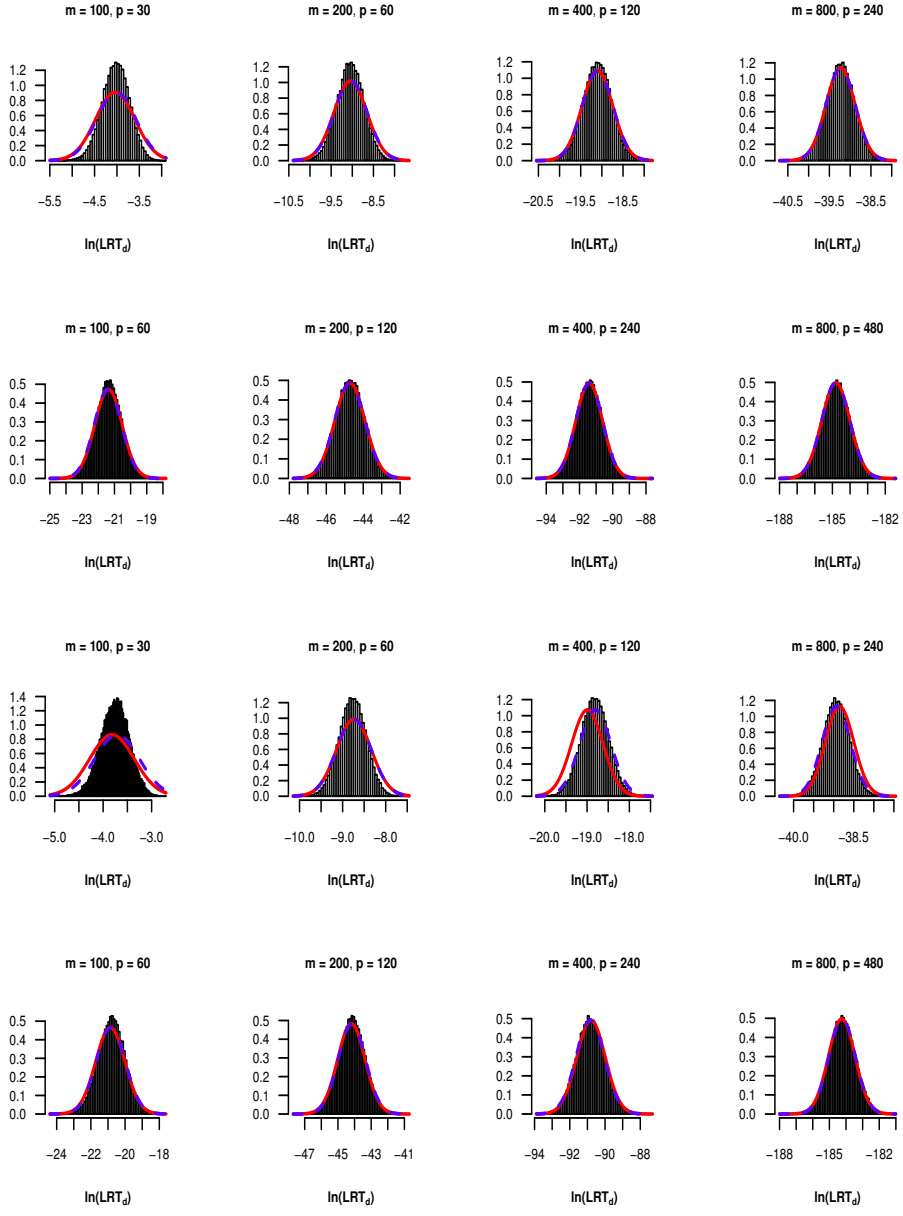


Figure 2: Simulation with $\sigma^2 = 1$, $p = 100, 200, 300, 400$, for $d = 4$, $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (7, 6, 5, 4)$, $p/m = .3$ in the first row, $p/m = .6$ in the second row, $d = 5$, $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = (259.72, 17.97, 11.04, 7.88, 4.82)$, $p/m = .3$ in the third row and $p/m = .6$ in the fourth row. The red curve is the asymptotic distribution of $\ln LRT_d$ given in Proposition 1 and the blue curve the lower approximation given in Corollary 1.

7. Conclusion

For the high-dimensional case ($p < n$ and $p, n \to \infty$) we study the asymptotic distribution of the maximum likelihood ratio statistics for partial sphericity in high-dimensional settings for the case of a spiked covariance model as introduced by Johnstone [11]. In addition, we consider the ultra high-dimensional case ($p > n$ and $p, n \to \infty$) and study the asymptotic distribution of the maximum likelihood ratio statistics where the roles of p and n are reversed. Knowledge of these asymptotic distributions allows us to develop a test to choose the dimension of the spike subspace that focuses on the non-spiked portion of the covariance matrix. One nice feature of this approach is that no knowledge of the variance of the non-spiked part is required. The study of the power of the maximum likelihood ratio test leads us to refine the test, adding a penalty term to the likelihood. The idea of a penalty term is connected to the *elbow method* used in cluster analysis to choose the number of clusters [24] and, also, to the information theoretic approaches such as AIC and MDL [26]. By studying the change of behavior of the distribution of the maximum likelihood ratio statistic for values below and over the true dimension of the spike subspace, we are able to modify it and to prove that the resulting estimator is consistent.

Acknowledgments

We would like to gratefully thank the Editor and an anonymous referee for her/his comments which have greatly improved our manuscript. This work was supported by the SECTEI grant 2010–072–14, by the UNL grants 500–040, 501–499 and 500–062; by the CONICET grant PIP 742 and by the ANPCYT grant PICT 2012–2590.

Appendix: Proofs

Through the proofs we will use the fact that in all spiked models with spike subspace of dimension d, $\hat{\sigma}^2 = \sum_{i=d+1}^{p} \hat{\lambda}_i/(p-d)$ converges almost surely to σ^2.

A. Proof of Proposition 1

We consider separately the cases $p < m$ and $p > m + d$,

1. **Case** $p < m$. This case follows using the technique developed by Muirhead [15] and the approximations given in [10]. For each $i \in \{0, \ldots, h\}$, let us call $\hat{\mathbf{\Psi}}_i$ the sample version of $\mathbf{\Psi}_i \in \mathbb{R}^{p \times s_i}$, where $s_0 = p - d$. Remember that, under the null hypothesis $\mathcal{H}_d$, the matrices $\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_h$ correspond to the d spike eigenvalues of the covariance matrix $\mathbf{\Sigma}$. We have $\widehat{\mathbf{\Sigma}} = (\hat{\mathbf{\Psi}}, \hat{\mathbf{\Psi}}_0)\widehat{\mathbf{\Lambda}}(\hat{\mathbf{\Psi}}, \hat{\mathbf{\Psi}}_0)^\top$ with $\widehat{\mathbf{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$. It should be noted that since $m > p$ all the eigenvalues of the sample covariance matrix are different and not 0 with probability 1. Now,

$$LRT_d = \frac{\hat{\lambda}_{d+1} \times \cdots \times \hat{\lambda}_p}{\left(\frac{1}{p-d}\sum_{i=d+1}^{p} \hat{\lambda}_i\right)^{p-d}} = LRT_* A_{\mathbf{\Psi}_1} \cdots A_{\mathbf{\Psi}_h} A_{\mathbf{\Psi}_0}$$

with

$$\begin{aligned}
LRT_* &= \frac{|\hat{\mathbf{\Sigma}}|}{\left\{\frac{\mathrm{tr}(\mathbf{\Psi}_1^\top \hat{\mathbf{\Sigma}} \mathbf{\Psi}_1)}{s_1}\right\}^{s_1} \cdots \left\{\frac{\mathrm{tr}(\mathbf{\Psi}_h^\top \hat{\mathbf{\Sigma}} \mathbf{\Psi}_h)}{s_h}\right\}^{s_h} \left\{\frac{\mathrm{tr}(\mathbf{\Psi}_0^\top \hat{\mathbf{\Sigma}} \mathbf{\Psi}_0)}{p-d}\right\}^{p-d}}, \\
A_{\mathbf{\Psi}_i} &= \frac{\left\{\frac{\mathrm{tr}(\mathbf{\Psi}_i^\top \hat{\mathbf{\Sigma}} \mathbf{\Psi}_i)}{s_i}\right\}^{s_i}}{|\hat{\mathbf{\Psi}}_i^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{\Psi}}_i|}, \quad i = 1, \ldots, h, \text{ and} \\
A_{\mathbf{\Psi}_0} &= \left\{\frac{\mathrm{tr}(\mathbf{\Psi}_0^\top \hat{\mathbf{\Sigma}} \mathbf{\Psi}_0)}{\mathrm{tr}(\hat{\mathbf{\Psi}}_0^\top \hat{\mathbf{\Sigma}} \hat{\mathbf{\Psi}}_0)}\right\}^{p-d}.
\end{aligned}$$

Since $\sigma_{m,p,d} \not\to 0$, the pieces required to get the result are

$$\frac{\ln(LRT_*) - \widetilde{\mu}_{m,p}}{\sigma_{m,p,d}} \rightsquigarrow \mathcal{N}(0,1), \tag{12}$$

$$\ln(\Pi_{i=1}^h A_{\mathbf{\Psi}_i}) - \ln(A_{m,p,d}) \to 0, \tag{13}$$

$$\ln(A_{\mathbf{\Psi}_0}) - \ln(B_{m,p,d}) \to 0. \tag{14}$$

Proof of (12)*:* To follow Muirhead's proof we need to compute $\mathrm{E}(LRT_*^t)$, the moment generating function of $\ln(LRT_*)$, under the null hypothesis for t in a neighborhood of 0. We find

$$\begin{aligned}
LRT_* &= \frac{|\hat{\mathbf{\Sigma}}|}{\prod_{i=0}^h \{\mathrm{tr}(\Psi_i^\top \hat{\mathbf{\Sigma}} \Psi_i / s_i)\}^{s_i}} \\
&= \frac{|\hat{\mathbf{\Sigma}}||\mathbf{\Sigma}^{-1}|}{\prod_{i=0}^h \{\lambda_i^{-1} \mathrm{tr}(\Psi_i^\top \hat{\mathbf{\Sigma}} \Psi_i / s_i)\}^{s_i}} = \frac{|\mathbf{\Sigma}^{-1/2} \hat{\mathbf{\Sigma}} \mathbf{\Sigma}^{-1/2}|}{\prod_{i=0}^h \{\mathrm{tr}(\Psi_i^\top \mathbf{\Sigma}^{-1/2} \hat{\mathbf{\Sigma}} \mathbf{\Sigma}^{-1/2} \Psi_i / s_i)\}^{s_i}}
\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{E}(LRT_*^t) &= \mathrm{E}\left[\frac{|\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}|^t}{\prod_{i=0}^h\{\,\mathrm{tr}(\Psi_i^\top\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\Psi_i/s_i)\}^{s_it}}\right]\\ &= \mathrm{E}\left\{\frac{|\mathbf{A}|^t}{\prod_{i=0}^h(\,\mathrm{tr}\mathbf{A}_{ii}/s_i)^{s_it}}\right\},\end{aligned}$$

where the expectation in the last expectation is taken with respect to a matrix $\mathbf{A} = m\ \boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} \sim \mathcal{W}_p(m, \mathbf{I}_p)$. Using Proposition 8.1 of [6], $\mathbf{A}_{ii} = m\ \Psi_i^\top\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\Psi_i \sim \mathcal{W}_{s_i}(m, \mathbf{I}_{s_i})$ and are independent; see Theorem 3.2.6 in [15]. As a consequence,

$$\begin{aligned}\mathrm{E}(LRT_*^t) &= \frac{1}{2^{\frac{mp}{2}}\Gamma_p(\frac{m}{2})}\int\frac{|\mathbf{A}|^{t+\frac{m-p-1}{2}}}{\Pi_{i=0}^h(\frac{\mathrm{tr}\mathbf{A}_{ii}}{s_i})^{s_it}}\exp\left(-\frac{1}{2}\,\mathrm{tr}\mathbf{A}\right)d\mathbf{A}\\ &= \frac{2^{\frac{(m+2t)p}{2}}\Gamma_p(\frac{m+2t}{2})}{2^{\frac{mp}{2}}\Gamma_p(\frac{m}{2})}\mathrm{E}\left\{\frac{1}{\Pi_{i=0}^h\left(\frac{\mathrm{tr}\mathbf{A}_{ii}}{s_i}\right)^{s_it}}\right\},\end{aligned} \tag{15}$$

where the expectation in the last expectation is taken with respect to a matrix $\mathbf{A} \sim \mathcal{W}_p(m+2t, \mathbf{I}_p)$ and $\mathbf{A}_{ii} \sim \mathcal{W}_{s_i}(m+2t, \mathbf{I}_{s_i})$ independent using the definition of Wishart distribution with non-integer degrees of freedom; see Section 3.2 in [15]. Now, $\mathrm{E}\{(\,\mathrm{tr}\mathbf{A}_{ii}/s_i)^{-s_it}\}$ is equal to

$$\begin{aligned}&\frac{1}{\Gamma_{s_i}(\frac{m+2t}{2})2^{\frac{(m+2t)s_i}{2}}}\int\left(\frac{\mathrm{tr}S}{s_i}\right)^{-s_it}|S|^{\frac{m+2t-s_i-1}{2}}\exp(-\frac{1}{2}\,\mathrm{tr}S)dS\\ =\ &\frac{1}{\Gamma_{s_i}(\frac{m+2t}{2})2^{\frac{(m+2t)s_i}{2}}}\int|S|^t\left(\frac{\mathrm{tr}S}{s_i}\right)^{-s_it}|S|^{\frac{m-s_i-1}{2}}\exp(-\frac{1}{2}\,\mathrm{tr}S)dS\\ =\ &\frac{\Gamma_{s_i}(\frac{m}{2})2^{\frac{ms_i}{2}}}{\Gamma_{s_i}(\frac{m+2t}{2})2^{\frac{(m+2t)s_i}{2}}}\mathrm{E}\left[\left\{\frac{|S|}{(\frac{1}{s_i}\,\mathrm{tr}S)^{s_i}}\right\}^t\right],\end{aligned}$$

where the last expectation is considering $S \sim \mathcal{W}_{s_i}(m, \mathbf{I}_{s_i})$. Using Corollary 8.3.6 of [15],

$$\mathrm{E}\left\{\left(\frac{\mathrm{tr}\mathbf{A}_{ii}}{s_i}\right)^{-s_it}\right\} = \frac{s_i^{s_it}\Gamma(\frac{1}{2}s_im)}{2^{ts_i}\Gamma(\frac{1}{2}s_im + s_it)}.$$

Consequently, replacing in (15),

$$\begin{aligned}\mathrm{E}(LRT_*^t) &= \frac{\Gamma_p(\frac{m}{2}+t)}{\Gamma_p(\frac{m}{2})}(p-d)^{(p-d)t}\frac{\Gamma\{\frac{1}{2}(p-d)m\}}{\Gamma\{\frac{1}{2}(p-d)m+(p-d)t\}} \quad (16)\\ &\quad \Pi_{i=1}^h \frac{s_i^{s_i t}\Gamma(\frac{1}{2}s_i m)}{\Gamma(\frac{1}{2}s_i m + s_i t)}.\end{aligned}$$

We now use Lemma 5.1 from [10], which is a consequence of Stirling's expansion for Gamma functions. We take $b(x) = 2tx/m$ with $x = m(p-d)/2$. Since $b(x) = \mathcal{O}(\sqrt{x}) = \mathcal{O}(m)$ for t finite, $-\ln[\Gamma\{m(p-d)/2 + t(p-d)\}/\Gamma\{m(p-d)/2\}]$ is equal to

$$\begin{aligned}&-t(p-d)\ln\left\{\frac{m}{2}(p-d)\right\} - \frac{t^2(p-d)^2 - t(p-d)}{m(p-d)} + \mathcal{O}(1/m)\\ =\ &-t(p-d)\ln\left\{\frac{m}{2}(p-d)\right\} - \frac{t^2(p-d)}{m} + \mathcal{O}(1/m)\end{aligned}$$

and taking $b(x) = s_i t$, for t finite and $x = m s_i/2$, $b(x) = \mathcal{O}(1)$,

$$\begin{aligned}-\ln\frac{\Gamma(\frac{m}{2}s_i + s_i t)}{\Gamma(\frac{m}{2}s_i)} &= -s_i t\ln\left(\frac{m}{2}s_i\right) - \frac{s_i^2 t^2 - s_i t}{m s_i} + \mathcal{O}(1/m^2)\\ &= -s_i t\ln\left(\frac{m}{2}s_i\right) + \mathcal{O}(1/m).\end{aligned}$$

Taking logarithms in (16) and using the above approximations plus Lemma 5.4 from [10], we have for $r_m^2 = -\ln(1-p/m)$ and $t = \mathcal{O}(1/r_m)$:

$$\ln \mathrm{E}(LRT_*^t) = 2\left\{-\frac{(p-d)}{m} + r_m^2\right\}\frac{t^2}{2} + \left\{-p + r_m^2(m-p-\frac{1}{2})\right\}t + o(1),$$

which leads to (12).

Proof of (13)*:* Since $m\boldsymbol{\Psi}_i^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Psi}_i \sim \mathcal{W}_{s_i}(m, \lambda_i\mathbf{I}_{s_i})$, we have when $m\to\infty$ $\mathrm{tr}(\boldsymbol{\Psi}_i^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Psi}_i)/s_i \to \lambda_i$; seeTheorem 3.2.20 in [15]. Therefore,

$$\sum_{i=1}^h \ln(A_{\boldsymbol{\Psi}_i}) - \ln(A_{m,p,d}) = \sum_{i=1}^h s_i\left\{\ln\frac{\mathrm{tr}(\boldsymbol{\Psi}_i^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Psi}_i)}{s_i} - \ln\lambda_i\right\} \to 0.$$

Proof of (14)*:* By definition of $A_{\mathbf{\Psi}_0}$ we need to compute

$$\frac{\mathrm{tr}(\mathbf{\Psi}_0^\top \widehat{\mathbf{\Sigma}} \mathbf{\Psi}_0)}{\mathrm{tr}(\hat{\mathbf{\Psi}}_0^\top \widehat{\mathbf{\Sigma}} \hat{\mathbf{\Psi}}_0)} = \frac{\sum_{i=1}^p \hat{\lambda}_i - \mathrm{tr}(\mathbf{\Psi}^\top \widehat{\mathbf{\Sigma}} \mathbf{\Psi})}{\mathrm{tr}(\hat{\mathbf{\Psi}}_0^\top \widehat{\mathbf{\Sigma}} \hat{\mathbf{\Psi}}_0)}.$$

Therefore,

$$\ln\left(A_{\mathbf{\Psi}_0}/B_{m,p,d}\right) = (p-d)\ln\left\{1 + \frac{1}{p-d}\frac{\sum_{i=1}^h s_i\lambda_i - \mathrm{tr}(\mathbf{\Psi}^\top \widehat{\mathbf{\Sigma}} \mathbf{\Psi})}{\frac{1}{p-d}(\sum_{i=1}^p \hat{\lambda}_i - \sum_{i=1}^h s_i\lambda_i)}\right\},$$

and the proof of (14) follows if

$$a = \frac{\sum_{i=1}^h s_i\lambda_i - \mathrm{tr}(\mathbf{\Psi}^\top \widehat{\mathbf{\Sigma}} \mathbf{\Psi})}{\frac{1}{p-d}(\sum_{i=1}^p \hat{\lambda}_i - \sum_{i=1}^h s_i\lambda_i)} \to 0. \tag{17}$$

In fact, it was proven above that the numerator goes to 0. On the other hand, for the denominator we have

$$\mathrm{E}\left\{\frac{1}{p-d}\left(\sum_{i=1}^p \hat{\lambda}_i - \sum_{i=1}^h s_i\lambda_i\right)\right\} = \sigma^2$$

$$\mathrm{var}\left\{\frac{1}{p-d}\left(\sum_{i=1}^p \hat{\lambda}_i - \sum_{i=1}^h s_i\lambda_i\right)\right\} = \frac{2}{m(p-d)^2}\left\{\sum_{i=1}^h s_i\lambda_i^2 + \sigma^4(p-d)\right\} \to 0$$

from which (17) follows.

2. **Case** $p > m + d$. Since $m\widehat{\mathbf{\Sigma}} \sim \mathcal{W}_p(m, \mathbf{\Sigma})$ then $m\widehat{\mathbf{\Sigma}} = \mathbf{Z}^\top\mathbf{Z}$ with $\mathbf{Z} \in \mathbb{R}^{m\times p} \sim \mathcal{N}(0, \mathbf{I}_m \otimes \mathbf{\Sigma})$ and $m\widetilde{\mathbf{\Sigma}} = \mathbf{Z}\mathbf{Z}^\top = \mathbf{Z}(\mathbf{\Psi}, \mathbf{\Psi}_0)(\mathbf{\Psi}, \mathbf{\Psi}_0)^\top\mathbf{Z}^\top \in \mathbb{R}^{m\times m}$ with $\mathbf{Z}(\mathbf{\Psi}, \mathbf{\Psi}_0) \in \mathbb{R}^{m\times p}$ and $\mathbf{Z}(\mathbf{\Psi}, \mathbf{\Psi}_0) \sim \mathcal{N}(0, \mathbf{I}_m \otimes \mathbf{\Lambda})$. As a consequence,

$$m\widetilde{\mathbf{\Sigma}} = \mathbf{Z}\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{Z} + \mathbf{Z}\mathbf{\Psi}_0\mathbf{\Psi}_0^\top Z.$$

Now, $\mathbf{Z}\mathbf{\Psi} \in \mathbb{R}^{m\times d}$ and $\mathbf{Z}\mathbf{\Psi} \sim \mathcal{N}(0, \mathbf{I}_m \otimes \mathbf{\Lambda}_d)$ and therefore $\mathbf{Z}\mathbf{\Psi} = \mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}$ for some $\mathbf{Z}_d \sim \mathcal{N}(0, \mathbf{I}_m \otimes \mathbf{I}_d)$. Analogously, $\mathbf{Z}\mathbf{\Psi}_0 = \sigma\tilde{\mathbf{Z}}$ for some $\tilde{\mathbf{Z}} \sim \mathcal{N}(0, \mathbf{I}_m \otimes \mathbf{I}_{p-d})$ and moreover $\mathbf{Z}_d$ and $\tilde{\mathbf{Z}}$ are independent. Therefore,

$$\widetilde{\mathbf{\Sigma}} = \frac{1}{m}\mathbf{Z}_d\mathbf{\Lambda}_d\mathbf{Z}_d^\top + \frac{\sigma^2}{m}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top, \tag{18}$$

with $\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \sim \mathcal{W}_m(p-d, \mathbf{I}_m)$ independent of $\mathbf{Z}_d \sim N(0, \mathbf{I}_m \otimes \mathbf{I}_d)$. Now,

$$|\widetilde{\mathbf{\Sigma}}| = \left|\frac{\sigma^2}{m}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\right| \left|\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}\right|$$

Let us note that the eigenvalues of $\widetilde{\mathbf{\Sigma}}$ are the non-zeros eigenvalues of $\widehat{\mathbf{\Sigma}}$: $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$. Therefore,

$$\begin{aligned}
LRT_d &= \frac{|\widetilde{\mathbf{\Sigma}}|}{\left(\frac{1}{m-d}\sum_{i=d+1}^m \hat{\lambda}_i\right)^{m-d}} \frac{1}{\hat{\lambda}_1 \times \cdots \times \hat{\lambda}_d} \\
&= \frac{|\frac{\sigma^2}{m}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top||\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}|}{\left(\frac{1}{m-d}\sum_{i=d+1}^m \hat{\lambda}_i\right)^{m-d}} \frac{1}{\hat{\lambda}_1 \times \cdots \times \hat{\lambda}_d} \\
&= \frac{|\sigma^2\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top|}{\left\{\frac{1}{m}\operatorname{tr}(\sigma^2\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)\right\}^m} \left\{\frac{\frac{1}{m}\operatorname{tr}(\sigma^2\frac{\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top}{m})}{\frac{1}{m-d}\sum_{i=d+1}^m \hat{\lambda}_i}\right\}^{m-d} \left\{\frac{1}{m}\operatorname{tr}\left(\sigma^2\frac{\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top}{m}\right)\right\}^d \\
&\quad \frac{|\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}|}{\hat{\lambda}_1 \times \cdots \times \hat{\lambda}_d} \\
&= LRT^* BCD,
\end{aligned}$$

where LRT^*, B, C and D are the factors in exactly the same order that in the third line of the previous equation. Since $\sigma_{m,p,d} \not\to 0$, the pieces to get the result are

$$\frac{\ln(LRT^*) - \widetilde{\mu}^*_{m,p,d}}{\sigma_{m,p,d}} \rightsquigarrow \mathcal{N}(0,1), \tag{19}$$

$$\ln B - \ln B^*_{m,p,d} \to 0, \tag{20}$$

$$\ln C - \ln C^*_{m,p,d} \to 0, \tag{21}$$

$$\ln D - \ln D^*_{m,p,d} \to 0. \tag{22}$$

Proof of (19)*:* By definition,

$$LRT^* = \frac{|\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top|}{\left\{\frac{1}{m}\operatorname{tr}(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)\right\}^m}.$$

Now, the result follows from [10] since $\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \sim \mathcal{W}_m(p-d, \mathbf{I}_m)$.

Proof of (20)*:* By definition of B and (18),

$$B = \left\{1 + \frac{\sum_{i=1}^d \hat{\lambda}_i - \frac{1}{m}\operatorname{tr}\left(\mathbf{Z}_d\mathbf{\Lambda}_d\mathbf{Z}_d^\top\right)}{\operatorname{tr}(\hat{\mathbf{\Psi}}_0^\top\widehat{\mathbf{\Sigma}}\hat{\mathbf{\Psi}}_0)}\right\}^{m-d}\left(\frac{m-d}{m}\right)^{m-d}$$

Then

$$\ln\left(B/B^*_{m,p,d}\right) = (m-d)\ln\left\{1 + \frac{\sum_{i=1}^h s_i\lambda_i - \frac{1}{m}\operatorname{tr}\left(\mathbf{Z}_d\mathbf{\Lambda}_d\mathbf{Z}_d^\top\right)}{\operatorname{tr}(\hat{\mathbf{\Sigma}}) - \sum_{i=1}^h s_i\lambda_i}\right\}$$

and the proof of (20) follows if

$$a = \frac{\sum_{i=1}^h s_i\lambda_i - \frac{1}{m}\operatorname{tr}\left(\mathbf{Z}_d\mathbf{\Lambda}_d\mathbf{Z}_d^\top\right)}{\frac{1}{m-d}(\operatorname{tr}\hat{\mathbf{\Sigma}} - \sum_{i=1}^h s_i\lambda_i)} \to 0. \tag{23}$$

In fact, since $\mathbf{Z}_d^T\mathbf{Z}_d \sim \mathcal{W}_d(m, I_d)$, $\mathrm{E}\{\operatorname{tr}\left(\mathbf{Z}_d\mathbf{\Lambda}_d\mathbf{Z}_d^\top/m\right)\} = \sum_{i=1}^h s_i\lambda_i$ and $\operatorname{var}\{\operatorname{tr}\left(\mathbf{Z}_d\mathbf{\Lambda}_d\mathbf{Z}_d^\top/m\right)\} = 2\sum_{i=1}^d s_i\lambda_i^2/m = O(m^{-1})$. Therefore the numerator goes to 0. On the other hand, the denominator goes to $y\sigma^2 < \infty$ since $\operatorname{tr}\widehat{\mathbf{\Sigma}}/p \to \sigma^2$, from what follows (23).

Proof of (21)*:* Since

$$\mathrm{E}\left\{\frac{\operatorname{tr}(\frac{\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top}{m})}{p-d}\right\} = \frac{1}{m(p-d)}\operatorname{tr}\{\mathrm{E}(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)\} = \frac{\sigma^2}{m}\operatorname{tr}(\mathbf{I}_m) = 1$$

and

$$\begin{aligned}\operatorname{var}\left\{\frac{\operatorname{tr}(\frac{\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top}{m})}{p-d}\right\} &= \frac{1}{(p-d)^2m^2}\operatorname{var}\{\operatorname{tr}(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)\} \\ &= \frac{2}{m(p-d)} \to 0,\end{aligned}$$

it follows that $\ln(C) - \ln(C^*_{m,p,d}) \to 0$.

Proof of (22)*:* By definition of D and $D^*_{m,p,d}$ it is enough to prove that

$$\ln\left|\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}\right| - \sum_{i=1}^h s_i\ln\left(1 + \frac{\lambda_i}{\sigma^2}\frac{m}{p-d-m-1}\right) \to 0.$$

But this follows directly from the fact that since $\mathbf{Z}_d$ and $\widetilde{\mathbf{Z}}$ are independent and $\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \sim \mathcal{W}_m(p-d, \mathbf{I}_m)$ using the moments of the inverse Wishart distribution [25],

$$\mathrm{E}\left\{\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}\right\} = \mathbf{I}_d + \frac{1}{\sigma^2}\frac{m}{p-d-m-1}\mathbf{\Lambda}_d > 0,$$

and

$$\mathrm{var}\left[\mathrm{vec}\left\{\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}\right\}\right] \to 0.$$

To check the last statement, we use the variance decomposition formula, the fact that

$$\mathrm{var}\left[\mathrm{vec}\left\{\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\mathbf{Z}_d\mathbf{\Lambda}_d^{1/2}\right\}|\mathbf{Z}_d\right] = \mathbf{K}\mathrm{var}\{\mathrm{vec}(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top)^{-1}\}\mathbf{K}^\top,$$

with $\mathbf{K} = (\mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top \otimes \mathbf{\Lambda}_d^{1/2}\mathbf{Z}_d^\top)$ and the moments of the inverse Wishart distribution [25]. The proof of (22) follows from the fact that the determinant is a continuous function.

B. Proof of Lemma 1

The lemma follows if we prove that

$$\frac{\hat{\mu}_{m,p,d} - \mu_{m,p,d}}{\sigma_{m,p,d}} \to L \qquad \text{when } p, m \to \infty \tag{24}$$

We will use repetitively that $\tilde{\lambda}_i \to \lambda_i$ when $\lambda_i \in J_1$ and $\tilde{\lambda}_i \to \sigma^2(1+\sqrt{y})$ for $\lambda_i \in J_2$. Let us consider the cases $p < m$ and $p > m + d$ separately.

1. Case $p < m$. The difference between the term $\ln(A_{m,p,d})$ for $\hat{\mu}_{m,p,d}$ and $\mu_{m,p,d}$ converges to

$$\sum_{i\in J_2} s_i \ln \sigma^2(1+\sqrt{y}) - \sum_{i\in J_2} s_i \ln \lambda_i. \tag{25}$$

To compare the term $\ln(B_{m,p,d})$ for $\hat{\mu}_{m,p,d}$ and $\mu_{m,p,d}$, let us note before that $\ln(B_{m,p,d})$ from Proposition 1 is asymptotically equivalent to

$$\widetilde{B}_{m,p,d} = y\sum_{i\in J_1} s_i \frac{\lambda_i}{\lambda_i - \sigma^2} + \sum_{i\in J_2} s_i(1+\sqrt{y})^2 - \sum_{i\in J_2} s_i \frac{\lambda_i}{\sigma^2}\,.$$

In fact, since $(p-d)\ln\{1+a_{p,m}/(p-d)\} = a + o(1)$ when $p-d \to \infty$ if $a = \lim a_{p,m}$ is finite, it is enough to prove that

$$\ln(B_{m,p,d}) - (p-d)\ln\left(1+\frac{1}{p-d}\widetilde{B}_{m,p,d}\right) \to 0$$

as p and m go to infinity. But by definition of $\ln(B_{m,p,d})$ and $\widetilde{B}_{m,p,d}$ we have

$$\begin{aligned}
& \ln(B_{m,p,d}) - (p-d)\ln\left(1+\frac{1}{p-d}\widetilde{B}_{m,p,d}\right) = \\
= \;& (p-d)\ln\left(1+\frac{\sum_{i=1}^{d}\hat{\lambda}_i - \sum_{i=1}^{h} s_i\lambda_i}{\sum_{i=d+1}^{p}\hat{\lambda}_i}\right) \\
& -(p-d)\ln\left[1+\frac{1}{p-d}\left\{y\sum_{i\in J_1} s_i\frac{\lambda_i}{\lambda_i-\sigma^2} + \sum_{i\in J_2} s_i(1+\sqrt{y})^2 - \sum_{i\in J_2} s_i\frac{\lambda_i}{\sigma^2}\right\}\right] \\
= \;& (p-d)\ln\left(1+\frac{1}{p-d}\,H\right).
\end{aligned}$$

The proof of the statement then follows if we prove that

$$H = \frac{-\left\{y\sum_{i\in J_1}\frac{s_i\lambda_i}{\lambda_i-\sigma^2} + \sum_{i\in J_2} s_i(1+\sqrt{y})^2 - \sum_{i\in J_2}\frac{s_i\lambda_i}{\sigma^2}\right\} + \frac{\sum_{i=1}^{d}\hat{\lambda}_i - \sum_{i=1}^{h} s_i\lambda_i}{\frac{1}{p-d}\sum_{i=d+1}^{p}\hat{\lambda}_i}}{1+\frac{1}{p-d}\left\{y\sum_{i\in J_1}\frac{s_i\lambda_i}{\lambda_i-\sigma^2} + \sum_{i\in J_2} s_i(1+\sqrt{y})^2 - \sum_{i\in J_2} s_i\frac{\lambda_i}{\sigma^2}\right\}} \to 0.$$

This can be seen through the fact that the numerator goes to 0 and the denominator goes to 1 using again [2]. As a consequence, the difference between the term $\widetilde{B}_{m,p,d}$ for $\hat{\mu}_{m,p,d}$ and $\mu_{m,p,d}$ is in the terms involving J_2 and it converges to

$$-\sum_{i\in J_2} s_i(1+\sqrt{y}) + \sum_{i\in J_2} s_i\frac{\lambda_i}{\sigma^2}\,. \tag{26}$$

Eqs. (25) and (26) give us (24) since $\sigma_{m,p,d} \to \sqrt{-2\{y+\ln(1-y)\}}$.

2. Case $p > m+d$. In this case, in part b) of the Proposition 1 $\ln(B^*_{m,p,d})$ can be replaced by

$$\sum_{i\in J_1} s_i\frac{\lambda_i}{\lambda_i-\sigma^2} + \sum_{i\in J_2} s_i\frac{(1+\sqrt{y})^2}{y} - \sum_{i\in J_2} s_i\frac{\lambda_i}{y\sigma^2} + (m-d)\ln\left(\frac{m-d}{m}\right)$$

and $\ln(D^*_{m,p,d})$ by

$$\sum_{i=1}^{h} s_i \ln\left\{1+\frac{\lambda_i}{\sigma^2(y-1)}\right\} - \sum_{i\in J_1} s_i \ln\left\{\lambda_i\left(1+y\frac{\sigma^2}{\lambda_i-\sigma^2}\right)\right\} - \sum_{i\in J_2} s_i\sigma^2(1+\sqrt{y})^2$$

with J_1 and J_2 as before. The proofs of these observations are completely analogous to the previous case. From this, and since the other terms do not involve λ_i,

$$\hat{\mu}_{m,p,d} - \mu_{m,p,d} \;\rightarrow\; \sum_{i\in J_2} s_i \left\{\frac{\lambda_i}{y\sigma^2} - \frac{1+\sqrt{y}}{y} - \ln\frac{\sigma^2(y-1)+\lambda_i}{\sigma^2\sqrt{y}(1+\sqrt{y})}\right\}$$

and $\sigma_{m,p,d}$ goes to $\sqrt{-2\left\{1/y+\ln(1-1/y)\right\}} < \infty$ when both p and m tend to infinity.

C. Proof of Proposition 2

Calling $r = \min(m,p)$, we can write

$$LRT_d = LRT_{d_1}\hat{\lambda}_{d+1}\times\cdots\times\hat{\lambda}_{d_1}\frac{\left(\frac{1}{r-d_1}\sum_{i=d_1+1}^{r}\hat{\lambda}_i\right)^{r-d_1}}{\left(\frac{1}{r-d}\sum_{i=d+1}^{r}\hat{\lambda}_i\right)^{r-d}} \text{ when } d<d_1 \quad (27)$$

$$LRT_d = LRT_{d_1}\frac{1}{\hat{\lambda}_{d_1+1}\times\cdots\times\hat{\lambda}_d}\frac{\left(\frac{1}{r-d_1}\sum_{i=d_1+1}^{r}\hat{\lambda}_i\right)^{r-d_1}}{\left(\frac{1}{r-d}\sum_{i=d+1}^{r}\hat{\lambda}_i\right)^{r-d}} \text{ when } d>d_1 \quad (28)$$

Using Proposition 1

$$\frac{\ln(LRT_{d_1}) - \mu_{m,p,d_1}}{\sigma_{m,p,d_1}} \rightsquigarrow \mathcal{N}(0,1), \quad (29)$$

and since for $i\in J_{1,k}$, $\widehat{\lambda}_i \rightarrow \lambda_i\{1+y\sigma^2/(\lambda_i-\sigma^2)\}$ and for $i\in J_{2,k}$, $\widehat{\lambda}_i \rightarrow \sigma^2(1+\sqrt{y})^2$ when $p,m\rightarrow\infty$ and $p/m\rightarrow y$,

$$\ln(\hat{\lambda}_{d+1}\cdots\hat{\lambda}_{d_1}) - \textstyle\sum_{i=h_0+1}^{h_1} s_i\ln(\sigma^2 k_i) \rightarrow 0, \text{ when } d<d_1, \quad (30)$$

$$\ln(\hat{\lambda}_{d_1+1}\cdots\hat{\lambda}_d) - (d-d_1)\ln\{\sigma^2(1+\sqrt{y})^2\} \rightarrow 0, \text{ when } d_1<d\le q_0, \quad (31)$$

where (31) follows from the fact that since $d_1 < d$ all the population eigenvalues $\lambda_{d_1+1}, \ldots, \lambda_d$ are equal to σ^2. Consequently, since $d \leq q_0$ fixed independent on m and p, $\hat{\lambda}_j \to \sigma^2(1+\sqrt{y})^2$ as p, m grow to infinity and $p/m \to y$ for $j = d_1 + 1, \ldots, d \leq q_0$. On the other hand we have

$$\frac{\left(\frac{1}{r-d_1}\sum_{i=d_1+1}^{r}\hat{\lambda}_i\right)^{r-d_1}}{\left(\frac{1}{r-d}\sum_{i=d+1}^{r}\hat{\lambda}_i\right)^{r-d}} = ABC, \tag{32}$$

With

$$A = \left(\frac{\sum_{i=d_1+1}^{r}\hat{\lambda}_i}{\sum_{i=d+1}^{r}\hat{\lambda}_i}\right)^{r-d} \to \begin{cases} \exp(-\frac{1}{\tilde{y}}\sum_{i=h_0+1}^{h_1} s_i k_i) \text{ if } d < d_1, & (33) \\ \exp\{\frac{1}{\tilde{y}}(d-d_1)(1+\sqrt{y})^2\} \text{ if } d < d_1, & (34) \end{cases}$$

$$B = \left(\frac{1}{r-d_1}\sum_{i=d_1+1}^{r}\hat{\lambda}_i\right)^{d-d_1} \to \left(\tilde{y}\sigma^2\right)^{d-d_1}, \tag{35}$$

$$C = \left(\frac{r-d}{r-d_1}\right)^{r-d} \to \exp(d_1 - d), \tag{36}$$

where $\tilde{y} = \max(1, y)$, and to prove (33) and (34) we use the limit result $(1 + a_t/t)^t \to \exp\{a\}$ when a_t is such that $a_t \to a$ as t grows to infinity, taking

$$a_t = -\left(\sum_{i=d+1}^{d_1}\hat{\lambda}_i\right) / \left\{\sum_{i=d+1}^{r}\hat{\lambda}_i/(r-d)\right\}$$

and

$$a_t = \left(\sum_{i=d_1+1}^{d}\hat{\lambda}_i\right) / \left\{\sum_{i=d+1}^{r}\hat{\lambda}_i/(r-d)\right\},$$

respectively. The result follows from (29)–(36) applied to (27) for the case $d < d_1$ and applied to (28) for the case $d > d_1$.

D. Proof of Proposition 3

Since all the eigenvalues are greater that the threshold, is equivalent (asymptotically) to use $\hat{\mu}_{m,p,d}$ or $\mu_{m,p,d}$. Using Proposition 2 for $d_1 > d$, we get for $Z \sim \mathcal{N}(0,1)$,

- Case $p < m$:

$$\begin{aligned}\Psi(d_1) &= \Pr\{\ln(LRT_d) < \mu_{m,p,d} + z_\alpha \sigma_{m,p,d} | d = d_1\} \\ &= \Pr\left(Z < \frac{\mu_{m,p,d} - \mu_{m,p,d,d_1} + z_\alpha \sigma_{m,p,d}}{\sigma_{m,p,d_1}}\right) \\ &= \Pr\left\{Z < \frac{\mu_{m,p,d} - \mu_{m,p,d_1} + \sum_{i=h_0+1}^{h_1} s_i(k_i - \ln k_i - 1) + z_\alpha \sigma_{m,p,d}}{\sigma_{m,p,d_1}}\right\}.\end{aligned}$$

- Case $p > m + d_1$:

$$\begin{aligned}\Psi(d_1) &= \Pr\{\ln(LRT_d) < \mu_{m,p,d} + z_\alpha \sigma_{m,p,d} | d = d_1\} \\ &= \Pr\left(Z < \frac{\mu_{m,p,d} - \mu_{m,p,d,d_1} + z_\alpha \sigma_{m,p,d}}{\sigma_{m,p,d_1}}\right) \\ &= \Pr\left\{Z < \frac{\mu_{m,p,d} - \mu_{m,p,d_1} + \sum_{i=h_0+1}^{h_1} s_i(\frac{k_i}{y} - \ln \frac{k_i}{y} - 1) + z_\alpha \sigma_{m,p,d}}{\sigma_{m,p,d_1}}\right\}.\end{aligned}$$

The result follows since $\mu_{m,p,d} - \mu_{m,p,d_1}$ converges to

$$-\sum_{i=h_0+1}^{h_1} s_i \left\{\ln\left(\frac{\lambda_i}{\sigma^2}\right) - \frac{\lambda_i}{\sigma^2}\right\} - \sum_{i=h_0+1}^{h_1} s_i(k_i - \ln k_i) \tag{37}$$

when $p < m$ and to

$$-\sum_{i=h_0+1}^{h_1} s_i \left\{\ln\left(1 - \frac{1}{y} + \frac{\lambda_i}{y\sigma^2}\right) - 1 + \frac{1}{y} - \frac{\lambda_i}{y\sigma^2}\right\} - \sum_{i=h_0+1}^{h_1} s_i \left(\frac{k_i}{y} - \ln \frac{k_i}{y}\right) \tag{38}$$

when $p > m + d_1$.

Now, from the proof of Lemma 1, for $p < m$, $\ln(B_{m,p,d})$ is asymptotically equivalent to $\sum_{i=1}^{h_k} s_i(k_i - \lambda_i/\sigma^2)$ from where (37) follows directly. Now for

$p > m + d_1$ (38) follows from

$$\begin{aligned}
\tilde{\mu}^*_{m,p,d} - \tilde{\mu}^*_{m,p,d_1} &= (p-m)\left\{\ln\left(1-\frac{m}{p-d_1}\right) - \ln\left(1-\frac{m}{p-d}\right)\right\} \\
&\quad -(d_1+\frac{1}{2})\ln\left(1-\frac{m}{p-d_1}\right) \\
&\quad +(d+\frac{1}{2})\ln\left(1-\frac{m}{p-d}\right) \\
&\sim \frac{d-d_1}{y} + (d-d_1)\ln(1-m/p), \\
\ln(B^*_{m,p,d}) - \ln(B^*_{m,p,d_1}) &\sim -\sum_{i=h_0+1}^{h_1} s_i\left(\frac{k_i}{y} - \frac{\lambda_i}{y\sigma^2}\right) - (d-d_1), \\
\ln(C^*_{m,p,d}) - \ln(C^*_{m,p,d_1}) &\sim (d-d_1)\ln(y\sigma^2), \\
\ln(D^*_{m,p,d}) - \ln(D^*_{m,p,d_1}) &\sim -\sum_{i=h_0+1}^{h_1} s_i\left\{\ln\left(1+\frac{\lambda_i}{\sigma^2}\frac{1}{p/m-1}\right) - \ln\sigma^2 k_i\right\}.
\end{aligned}$$

where $\sim$ means asymptotic equivalent.

E. Proof of Proposition 4.

Since $\lambda_i > \sigma^2(1+\sqrt{y})$ we get $J_{2,1} = J_{2,0} = \emptyset$. Using (9) and (10) we have that for $d \geq d_1$, $g(d) = \mu_{m,p,d_1} - (d-d_1)\epsilon$. Therefore g is a decreasing function for $d \geq d_1$. Now, for $d < d_1$, using again $\tilde{y} = \max(1, y)$, the function $g(d)$ is given by

$$\begin{aligned}
g(d) &= \mu_{m,p,d_1} + \sum_{i=h_0+1}^{h_1} s_i\left(\ln\frac{k_i}{\tilde{y}} - \frac{k_i}{\tilde{y}} + 1\right) \\
&\quad -(d-d_1)\left\{\frac{(1+\sqrt{y})^2}{\tilde{y}} - \ln\frac{(1+\sqrt{y})^2}{\tilde{y}} - 1 + \epsilon\right\} \\
&= \mu_{m,p,d_1} - \sum_{i=h_0+1}^{h_1} s_i[h(\lambda_i) - h\{\sigma^2(1+\sqrt{y})\} - \epsilon].
\end{aligned}$$

Since $\lambda_i > \lambda^* > \sigma^2(1+\sqrt{y})$ and h is a strictly increasing function, using the definition of ϵ, $h(\lambda_i) - h\{\sigma^2(1+\sqrt{y})\} - \epsilon > 0$. Therefore for $d \leq d_1$ as d increase we are adding less negative terms and therefore g increases for $d < d_1$. This allows us to conclude.

F. Proof of Proposition 5.

The proposition follows if we prove that for all $\delta > 0$ there exist m_0, p_{m_0} such that for $m > m_0$, $p > p_{m_0}$ and p/m close to y

$$\Pr(\cap_{d=0}^{q_0} A_d) \geq 1 - \delta$$

where

$$A_d = A_d(p, m, \epsilon) = \begin{cases} \{F(d, p, m, \epsilon) - F(d-1, p, m, \epsilon) > 0\} & \text{if } 0 \leq d \leq d_1, \\ \\ \{F(d, p, m, \epsilon) - F(d-1, p, m, \epsilon) < 0\} & \text{for } d_1 < d \leq q_0, \end{cases}$$

with $F(d, p, m, \epsilon) = \ln(LTR_d) - (d - d_1)\left[h\{\sigma^2(1 + \sqrt{y})\} + \epsilon\right]$. Let us call $r = \min(m, p)$. We consider first $d \leq d_1$. From the definition of $F(d, p, m, \epsilon)$,

$$F(d, p, m, \epsilon) - F(d-1, p, m, \epsilon) = \ln\left(\frac{LRT_d}{LRT_{d-1}}\right) - h\left\{\sigma^2(1 + \sqrt{y})\right\} - \epsilon.$$

We can now write

$$\frac{LRT_d}{LRT_{d-1}} = \frac{1}{\hat{\lambda}_d}\left(\frac{1}{r - d + 1}\sum_{i=d+1}^{r}\hat{\lambda}_i\right)\frac{1}{(1 + \frac{1}{r-d})^{r-d}}\left(1 + \frac{\hat{\lambda}_d}{\sum_{i=d+1}^{r}\hat{\lambda}_i}\right)^{r-d}.$$

Using again $\tilde{y} = \min(1, y)$, we then deduce that

$$\frac{LRT_d}{LRT_{d-1}} \to \frac{\tilde{y}\sigma^2}{\lambda_d\left(1 + \frac{y\sigma^2}{\lambda_d - \sigma^2}\right)}\exp(-1)\exp\left\{\frac{\lambda_d}{\tilde{y}\sigma^2}\left(1 + \frac{y\sigma^2}{\lambda_d - \sigma^2}\right)\right\}.$$

Therefore $\ln(LRT_d/LRT_{d-1}) \to h(\lambda_d)$ and, since $\lambda_d > \lambda^*$, the monotonicity of the function h implies that for large enough p, m (if $d \leq d_1$),

$$\Pr\left\{F(d, p, m) - F(d-1, p, m) > 0\right\} > 1 - \frac{\delta}{q_0}.$$

The case $d > d_1$ is very similar. We only have to notice that, in this case, $\ln(LRT_d/LRT_{d-1}) \to h\{\sigma^2(1 + \sqrt{y})\}$. So, for large enough p, m, we get

$$\Pr\{F(d, p, m) - F(d-1, p, m) < 0\} > 1 - \frac{\delta}{q_0}.$$

As a consequence $\Pr(\cap_{d=1}^{q_0} A_d) \geq 1 - \delta$ and the result follows by noticing that for $\hat{g}(y) = \ln(LRT_d) - (d - d_1)\{h(y) + \epsilon\}$,

$$\cap_{d=1}^{q_0} A_d \subseteq \left\{\hat{g}(y) \text{ is increasing for } d \leq d_1 \text{ and decreasing for } d \geq d_1\right\}.$$

References

[1] Z. Bai and J. Yao. Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. H. Poincaré Probab. Statist.*, 44(3):447–474, 06 2008. doi:10.1214/07-AIHP118. URL http://dx.doi.org/10.1214/07-AIHP118.

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

[3] J. Baik, G. Ben Arous, and S. Péché. Phase transition for the largest eigenvalue of large sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.

[4] M. S. Bartlett. A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):296–298, 1954. ISSN 00359246. URL http://www.jstor.org/stable/2984057.

[5] F. Benaych-Georges, A. Guionnet, and M. Maida. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.*, 16:no. 60, 1621–1662, 2011. ISSN 1083-6489. doi:10.1214/EJP.v16-929. URL http://ejp.ejpecp.org/article/view/929.

[6] J. Eaton. *Multivariate statistics: a vector space approach*. Institute of Mathematical Statistics, 1983.

[7] L. J. Gleser. A note on the sphericity test. *Ann. Math. Statist.*, 37 (2):464–467, 04 1966. doi:10.1214/aoms/1177699529. URL http://dx.doi.org/10.1214/aoms/1177699529.

[8] M. C. Harding, D. Jorgenson, G. King, O. Linton, G. Lorenzoni, B. Mazur, S. Panageas, and K. Patel. Structural estimation of high-dimensional factor models. Technical report, 2007.

[9] A. T. James. Test of equality of the latent roots of the covariance matrix. In P. R. Krishnaiah, editor, *Multivariate Analysis*, pages 205–218. Academic Press, New York, 1969.

[10] T. Jiang and F. Yang. Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Statist.*, 41 (4):2029–2074, 08 2013. doi:10.1214/13-AOS1134. URL `http://dx.doi.org/10.1214/13-AOS1134`.

[11] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 04 2001. doi:10.1214/aos/1009210544. URL `http://dx.doi.org/10.1214/aos/1009210544`.

[12] S. Kritchman and B. Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009.

[13] D. N. Lawley. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43(1-2):128–136, 1956. doi:10.1093/biomet/43.1-2.128. URL `http://biomet.oxfordjournals.org/content/43/1-2/128.short`.

[14] J. W. Mauchly. Significance test for sphericity of a normal n-variate distribution. *Ann. Math. Statist.*, 11(2):204–209, 06 1940. doi:10.1214/aoms/1177731915. URL `http://dx.doi.org/10.1214/aoms/1177731915`.

[15] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics, 1982, 2005.

[16] A. Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009. ISSN 1468-0262. doi:10.3982/ECTA6964. URL `http://dx.doi.org/10.3982/ECTA6964`.

[17] D. Passemier and J. Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 01(01):1150002, 2012. doi:10.1142/S201032631150002X. URL `http://www.worldscientific.com/doi/abs/10.1142/S201032631150002X`.

[18] D. Passemier and J. Yao. Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, 127:173–183, 2014.

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

[20] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65:066126, Jun 2002. doi:10.1103/PhysRevE.65.066126. URL `http://link.aps.org/doi/10.1103/PhysRevE.65.066126`.

[21] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14 (5):465–471, 09 1978.

[22] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2): 461–464, 03 1978. doi:10.1214/aos/1176344136. URL `http://dx.doi.org/10.1214/aos/1176344136`.

[23] M. S. Srivastava. Some tests criteria for the covariance matrix with fewer observations than the dimension. *Acta Comment. Univ. Tartu. Math.*, 10:77–93, 2006.

[24] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001. ISSN 1467-9868. doi:10.1111/1467-9868.00293. URL `http://dx.doi.org/10.1111/1467-9868.00293`.

[25] D. von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, 15(2):97–109, 1988. ISSN 03036898, 14679469. URL `http://www.jstor.org/stable/4616090`.

[26] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):387–392, 1985.