Statistica Sinica Preprint No: SS-2021-0346	
Title	A Universal Test on Spikes in a High-Dimensional Generalized Spiked Model and Its Applications
Manuscript ID	SS-2021-0346
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0346
Complete List of Authors	Dandan Jiang
Corresponding Authors	Dandan Jiang
E-mails	jiangdd@xjtu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

A UNIVERSAL TEST ON SPIKES IN A HIGH-DIMENSIONAL GENERALIZED SPIKED MODEL AND ITS APPLICATIONS

Dandan Jiang

School of Mathematics and Statistics

Xi'an Jiaotong University

Abstract: We test the number of spikes in a generalized spiked covariance matrix, the spiked eigenvalues of which may be much larger or smaller than the nonspiked ones. For high-dimensional problems, we first propose a general test statistic and derive its central limit theorem using random matrix theory without a Gaussian population constraint. We then apply the result to estimate the noise variance and test the equality of the smallest roots in generalized spiked models. The results of our simulation studies show that the proposed test method is sized correctly, and the power outcomes demonstrate the robustness of our statistic to deviations from a Gaussian population. Moreover, our estimator of the noise variance results in much smaller mean absolute errors and mean squared errors than those of existing methods. In contrast to other methods, we eliminate the strict conditions of a diagonal or a block-wise diagonal form of the population covariance matrix, and extend the work to a wider range, without the assumption of normality. Thus, the proposed method is highly suitable for real problems.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Key words and phrases: Generalized spiked model; High-dimensional covariance matrix; Testing the spikes; Central limit theorem.

1. Introduction

In this study, we consider a generalized spiked model from a population covariance matrix Σ, without a diagonal or a block-wise diagonal assumption, and with a Gaussian population constraint. We let T_p be a $p \times p$ deterministic matrix, and let $\Sigma = T_p T_p^*$ be a general population covariance matrix with spikes $\alpha_1, \ldots, \alpha_K$ and multiplicity m_k, for $k = 1, \ldots, K$, arranged arbitrarily in groups among all the eigenvalues. The condition $m_1 + \cdots + m_K = M$ is satisfied, where M is a fixed integer compared with the large dimension p. Furthermore, a few fixed eigenvalues (spikes) are allowed to be much larger or smaller than the majority of the eigenvalues. This model is the so-called generalized spiked model, and is closely related to a principal component analysis (PCA) and a factor analysis (FA), which are important and powerful tools for dimensionality reduction, data visualization, and feature extraction. Spiked models are also used in other scientific fields, for example, as factor models in economics and as signal-plus-noise models in wireless communication. In such cases, we need to determine whether the spikes affect the identification of the key factors in

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

a data set. Thus, the limiting behaviors of sample spiked eigenvalues and eigenvectors have attracted significant interest from researchers. In a pioneering work, Johnstone (2001) assumes the population covariance is a high-dimensional identity matrix with fixed spikes. Under this simplified framework, Baik et al. (2005), Baik and Silverstein (2006), Paul (2007), and Bai and Yao (2008) examined the limiting results of sample spiked eigenvalues. Bai and Yao (2012), Fan and Wang (2017), Cai et al. (2020), and Jiang and Bai (2021a,b) extended the structure of the population covariance to a more general form and investigated the asymptotic distributions of the sample spiked eigenvalues in high-dimensional settings.

However, relatively fewer studies test the number of spikes for large dimensionality p, which is offen a fundamental step in reconstructing the structure of the population covariance. Most related works determine the number of spikes in a high-dimensional setting using random matrix theory, such as those of Kritchman and Nadler (2008) and Passemier and Yao (2012). In contrast, Johnstone and Onatski (2020) test for the existence of spikes, Passemier et al. (2017) derive a goodness-of-fit test for a high-dimensional principal component model and determine the number of principal components, and Onatski (2009) test the number of factors in large factor models with a white noise assumption. These previous approaches

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

are limited in various ways, such as requiring a diagonal or a block-wise diagonal form of the population covariance matrix or a Gaussian population assumption, or only including extremely large spikes, but not extremely small ones.

To relax these restrictions, we recall the generalized spiked model described earlier and provide a corrected pseudo-likelihood ratio test on the number of spikes for this model. The proposed test is universal for all population assumptions and the general form of the spiked covariance matrix. We apply the test to estimate the noise variance and to test the equality of the smallest roots in the generalized spiked model. The proposed test method has several advantages over existing methods. First, we extend the population covariance matrix to a general nonnegative definite matrix and remove the diagonal and block-wise diagonal assumptions. Second, we establish the asymptotic distribution of the proposed test statistic in a high-dimensional setting, without assuming a Gaussian population. Moreover, under this setting, the spikes are allowed to be significantly larger or smaller than the nonspiked eigenvalues, occurring in several groups. Overall, our assumptions are more practical than those imposed in previous studies.

The remainder of this paper proceeds as follows. In Section 2, we describe the problem in a generalized setting, and propose a universal test on

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

the spikes, without imposing Gaussian and diagonal or block-wise diagonal assumptions. In Sections 3 and 4, we estimate the noise variance and test the equality of the smallest roots in generalized spiked models, respectively. We also conduct simulations for each result to compare our work with existing works. Then, we analyze two sets of real data and give the corresponding statistical inferences in Section 5. Finally, Section 6 concludes the paper. Detailed proofs are provided in the Appendix.

2. Test on spikes in a high-dimensional generalized spiked model

We consider the generalized spiked model first proposed by Jiang and Bai (2021a), and define the singular value decomposition of T_p as

$$T_p = V \begin{pmatrix} D_1^{\frac{1}{2}} & 0 \\ 0 & D_2^{\frac{1}{2}} \end{pmatrix} U^*, \tag{2.1}$$

where U and V are unitary matrices, D_1 is a diagonal matrix of the M spiked eigenvalues, and D_2 is a diagonal matrix of the nonspiked eigenvalues with bounded components. We define U_1 and U_2 as the first M and the last $p - M$ columns, respectively, of the matrix U defined in Eq. (2.1).

We assume that the double array $\{x_{ij}, i, j = 1, 2, ...\}$ consists of independent and identically distributed (i.i.d.) random variables with mean zero and variance one. Furthermore, $\mathrm{E}(x_{ij}) = 0$ and $\mathrm{E}(x_{ij}^2) = 0$ for the complex

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

case. Thus,

$$T_pX = (T_px_1, \cdots, T_px_n) \tag{2.2}$$

can be viewed as a random sample from a population with general population covariance matrix Σ, where $x_j = (x_{1j}, \cdots, x_{pj})'$, for $1 \leq j \leq n$. The corresponding sample covariance matrix of observations T_pX is

$$S = T_p \left(\frac{1}{n} XX^* \right) T_p^*, \tag{2.3}$$

which is the generalized spiked sample covariance matrix.

To test the number of the spikes, we use the following hypothesis:

$$\mathcal{H}_0 : M = M_0 \quad \text{vs.} \quad M \neq M_0, \tag{2.4}$$

where M_0 is a given nonnegative integer, such that (2.4) tests whether the true number of spikes is M_0. We consider the hypothesis test in (2.4) under a high-dimensional setting in which $p/n = c_n \to c > 0$ and both n and p tend to infinity simultaneously.

We define the empirical spectral distribution (ESD) of Σ as $H_n(t)$, which tends to a proper probability measure $H(t)$ as $p \to \infty$. We let

$$\mathcal{J}_k = \{j_k + 1, \ldots, j_k + m_k\} \tag{2.5}$$

denote the set of ranks of the m_k-ple eigenvalue α_k in the descending population eigenvalues, where α_k is out of the support of $H(t)$. Moreover,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

$\{l_j(S), j \in \mathcal{J}_k\}$, for $k = 1, \ldots, K$, are the associated sample eigenvalues of the matrix S, denoted as l_j, henceforth. By Proposition 2.1 of Jiang and Bai (2021a), for each population spiked eigenvalue α_k with multiplicity m_k satisfying the separation condition $\min_{i \neq k} |\alpha_k/\alpha_i - 1| > d$, where d is some positive constant, we have $l_j(S)/\phi_k - 1 \to 0, a.s.$, for all $j \in \mathcal{J}_k$ and the function $\phi(x) = x\{1 + c \int t/(x-t) \mathrm{d}H(t)\}$. This conclusion holds under the bounded fourth-moment assumption. However, based on the truncation procedures of Jiang and Bai (2021a), the convergence in Proposition 2.1 still holds in probability without the bounded fourth-moment assumption if one of the tail probabilities is satisfied, that is,

$$\lim_{\tau \to \infty} \tau^4 \mathrm{P}\left(|x_{ij}| > \tau\right) = 0. \tag{2.6}$$

Inspired by this result, we propose a test statistic for (2.4) and derive its asymptotic distribution. Recall that the likelihood ratio test statistic for (2.4) in the probabilistic principal component analysis model $\Sigma = \mathrm{diag}(a_1, \cdots, a_M, 0 \cdots, 0) + \sigma^2 I$ is expressed by

$$L = \left[\frac{1}{p-M} \sum_{i=M+1}^{p} l_i \left(\prod_{i=M+1}^{p} l_i \right)^{-\frac{1}{p-M}} \right]^{-\frac{(p-M)n}{2}} \tag{2.7}$$

in classical statistical theory. The test statistic $-2 \log L$ relies mainly on the partial linear spectral statistic involved with the nonspiked eigenvalues, such as $\sum_{i=M+1}^{p} l_i$ and $\sum_{i=M+1}^{p} \log l_i$. Following Anderson and Rubin (1956), we

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

also use this type of statistic in the goodness-of-fit test for the probabilistic principal component analysis model. Therefore, for the generalized spiked model, we propose the statistic

$$\sum_{j=1}^{p} f(l_j) - \sum_{j\in\mathcal{J}_k,k=1}^{K} f(l_j), \tag{2.8}$$

where $\mathcal{J}_k$ is defined as (2.5), $f \in \mathcal{A}$, and $\mathcal{A}$ is a set of analytic functions defined on an open set of the complex plane, including the whole supporting set of the limiting spectral distribution (LSD), $H(t)$. We define $F^{c,H}$ as the LSD of the sample matrix S, and F^{c_n,H_n} is the analogue of $F^{c,H}$, with c and H replaced by c_n and H_n, respectively. Furthermore, $\underline{m}(z) \equiv m_{\underline{F}^{c,H}}(z)$ is defined as the Stieltjes transform of $\underline{F}^{c,H} \equiv (1-c)I_{[0,\infty)} + cF^{c,H}$. To obtain the asymptotic distribution of the test statistic (2.8), we require the following assumptions:

Assumption (a): The tail probability (2.6) is satisfied and $p/n = c_n \to c > 0$ as both $n, p \to \infty$;

Assumption (b): Assume that $\lim \sum_{j=1}^{p} |u_{ji}|^4 \mathrm{E}\{|x_{11}|^4 I(|x_{11}| \le \sqrt{n}) - q - 2\} < \infty$, where $q = 1$ for the real case, $q = 0$ for the complex case, $I(\cdot)$ is the indicator function, and $u_i = (u_{1i}, \ldots, u_{pi})'$ is the ith column of the matrix U_1.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Assumption (b*): Suppose that

$$\max_{1\le i\le M,1\le j\le p}|u_{ji}|^2\mathrm{E}\big\{|x_{11}|^4I(|x_{11}|<\sqrt{n})-q-2\big\}\to 0. \tag{2.9}$$

Thus, the central limit theorem (CLT) for the test statistic (2.8) is established as follows; the proof is provided in the Appendix.

Theorem 1. *For the testing problem (2.4), suppose that Assumptions (a) and (b) [or (b*)] hold simultaneously. Then, the asymptotic distribution of the test statistic (2.8) is as follows:*

$$T_{f,H}=\nu_{f,H}^{-\frac{1}{2}}\left\{\sum_{j=1}^{p}f(l_j)-\sum_{j\in\mathcal{J}_k,k=1}^{K}f(l_j)-b_{f,H_n}-\mu_{f,H}\right\}\Rightarrow\mathcal{N}(0,1),\tag{2.10}$$

where

$$\begin{aligned}
b_{f,H_n}&=p\int f(t)\mathrm{d}F^{c_n,H_n}(t)-\sum_{k=1}^{K}m_kf\left(\alpha_k+c\alpha_k\int\frac{t}{\alpha_k-t}\mathrm{d}H(t)\right),\\
\mu_{f,H}&=-\frac{q}{2\pi i}\oint f(z)\frac{c\int\underline{m}^3(z)t^2\{1+t\underline{m}(z)\}^{-3}\mathrm{d}H(t)}{\big[1-c\int\underline{m}^2(z)t^2\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)\big]^2}\mathrm{d}z\\
&\quad-\frac{\beta c}{2\pi i}\oint f(z)\frac{\underline{m}^3(z)\cdot\int t\{1+t\underline{m}(z)\}^{-1}\mathrm{d}H(t)\cdot\int\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)}{1-c\int\underline{m}^2(z)t^2\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)}\mathrm{d}z,\\
\nu_{f,H}&=-\frac{q+1}{4\pi^2}\oint\oint\frac{f(z_1)f(z_2)}{\{\underline{m}(z_1)-\underline{m}(z_2)\}^2}\mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2)\\
&\quad-\frac{\beta c}{4\pi^2}\oint\oint f(z_1)f(z_2)\int\frac{t\mathrm{d}H(t)}{\{1+t\underline{m}(z_1)\}^2}\int\frac{t\mathrm{d}H(t)}{\{1+t\underline{m}(z_2)\}^2}\mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2).
\end{aligned}\tag{2.11}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Here, $\beta = \lim \sum_{j=1}^{p} |u_{ji}|^4 \mathrm{E}\{|x_{11}|^4 I(|x_{11}| \leq \sqrt{n}) - q - 2\}$ *if Assumption (b) is met, and* $\beta = 0$ *if Assumption (b*) holds instead. The contours all contain the support of* $F^{c,H}$ *and are non-overlapping in (2.11).*

Remark 1. Note that the CLT given by Theorem 1 depends on the number m_k, the values of the population spikes α_k, and the LSD $H(t)$. However, these parameters are usually unknown in data analyses, and need to be estimated; see (Li et al. (2013), Jiang and Bai (2021a), Bao et al. (2019), and Zheng et al. (2021)).

For $H(t) = \delta_{\{1,+\infty\}}$, we select $f(x) = x$ and $f(x) = \log x$ as two examples, and present the details of the computations in the Supplementary Material.

Example 1. If $f(x) = x$ and $H(t) = \delta_{\{1,+\infty\}}$, the statistic in (2.10) simplifies to

$$T_{x,1} = \nu_{x,1}^{-\frac{1}{2}} \left\{ \sum_{j=1}^{p} l_j - \sum_{j \in \mathcal{J}_k, k=1}^{K} l_j - b_{x,1} - \mu_{x,1} \right\} \Rightarrow \mathcal{N}(0,1),$$

where $b_{x,1} = (p-M) - \sum_{i=1}^{K} m_k c \alpha_k / (\alpha_k - 1)$, $\mu_{x,1} = 0$, and $\nu_{x,1} = (q+1+\beta)c$.

Example 2. If $f(x) = \log x$ and $H(t) = \delta_{\{1,+\infty\}}$, the statistic in (2.10) is given as

$$T_{\log,1} = \nu_{\log,1}^{-\frac{1}{2}} \left\{ \sum_{j=1}^{p} \log(l_j) - \sum_{j \in \mathcal{J}_k, k=1}^{K} \log(l_j) - b_{\log,1} - \mu_{\log,1} \right\} \Rightarrow \mathcal{N}(0,1),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

where

$$b_{\log,1} = p\left\{\frac{(c-1)}{c}\log(1-c) - 1\right\} - \sum_{i=1}^{K} m_k \log(1 + \frac{c}{\alpha_k - 1}),$$

$$\mu_{\log,1} = \frac{q}{2}\log(1-c) - \frac{1}{2}\beta c, \quad \nu_{\log,1} = -(q+1)\log(1-c) + \beta c.$$

2.1 Monte Carlo experiments

To demonstrate the effectiveness of the proposed CLT using simulations, we first provide the following two models:

Model 1: Assume that $\Sigma = \Lambda$, where Λ is assumed to be an identity matrix with a finite-rank perturbation. In descending order, its spikes are $(25, 16, 16, 0.2, 0.2, 0.1)$, with multiplicities $(1, 2, 2, 1)$.

Model 2: Assume that $\Sigma = U_0 \Lambda U_0^*$, where Λ is defined in Model 1, and U_0 is a matrix of the eigenvectors of a $p \times p$ matrix, with entries sampled independently from $N(0, 1)$. Thus, we relax the diagonal assumption of Σ.

Furthermore, to show that the conclusion is widely applicable and free of the population assumption, we consider the following Gaussian and gamma populations for each model:

Gaussian Assumption: $\{x_{ij}\}$ are i.i.d. samples from a standard Gaussian population.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Gamma Assumption: $\{x_{ij}\}$ are i.i.d. from the population distribution $Gamma(4, 0.5) - 2$.

For the sake of simplicity, we select the function $f(x) = x$ and the LSD $H(t) = \delta_{\{1,+\infty\}}$. For each case, the sample size is set to $n = 100$, 200, and 400, and $p/n = 0.5$, 1, and 1.5, respectively. We report the empirical probability of rejecting the null hypothesis (2.4), $\mathcal{H}_0 : M = M_0$, with 1000 replicates in Table 1 and Table 2. We list only the results for Model 2, because those for Model 1 are similar, and thus given in the Supplementary Material. Moreover, we plot the empirical distributions of the proposed test statistic when M_0 is equal to the true value of M. Figure 1 shows the performance of our proposed method under Model 2 with a gamma population assumption; the figures for the other cases are provided in the Supplementary Material.

To further demonstrate that the proposed CLT is valid for a distribution with infinite fourth moments, we generate i.i.d. samples x_{ij} from a $2^{-1/2}t(4)$ population distribution under Model 2, where the fourth moments of x_{ij} are infinite. The simulated results are presented in the Supplementary Material.

The results of the simulation study show that the proposed test on the number of spikes provides good sizes for both the Gaussian and the non-Gaussian, diagonal and off-diagonal populations when the null hypothesis

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Table 1: Empirical probability of rejecting the null hypothesis (2.4) for Model 2 under a Gaussian assumption when the true value of M is six.

Values of M_0	1	2	3	4	5	6	7
$p = 50\ ;\ n = 100$	1	1	0.649	0.319	0.096	**0.048**	0.068
$p = 100;\ n = 200$	1	1	0.691	0.366	0.116	**0.049**	0.115
$p = 200;\ n = 400$	1	1	0.698	0.389	0.129	**0.055**	0.150
$p = 100;\ n = 100$	1	1	0.383	0.195	0.085	**0.044**	0.125
$p = 200;\ n = 200$	1	1	0.411	0.223	0.099	**0.042**	0.174
$p = 400;\ n = 400$	1	1	0.403	0.231	0.121	**0.052**	0.213
$p = 150;\ n = 100$	1	0.998	0.263	0.151	0.078	**0.039**	0.174
$p = 300;\ n = 200$	1	1	0.286	0.178	0.102	**0.040**	0.244
$p = 600;\ n = 400$	1	1	0.301	0.199	0.124	**0.054**	0.263

is true. Moreover, the power increases as the alternative hypothesis moves further from the null hypothesis. Based on the above results, we infer that M_0 tested in (2.4) usually matches the true value of M at the inflection points of the empirical sizes. This corresponds to the first local minimum value of the empirical sizes.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Table 2: Empirical probability of rejecting the null hypothesis (2.4) for Model 2 under a gamma assumption when the true value of M is six.

Values of M_0	1	2	3	4	5	6	7
$p = 50$; $n = 100$	1	1	0.403	0.190	0.071	**0.038**	0.051
$p = 100$; $n = 200$	1	1	0.446	0.235	0.085	**0.046**	0.086
$p = 200$; $n = 400$	1	1	0.456	0.236	0.101	**0.043**	0.112
$p = 100$; $n = 100$	1	0.996	0.242	0.115	0.060	**0.040**	0.086
$p = 200$; $n = 200$	1	1	0.282	0.155	0.082	**0.044**	0.127
$p = 400$; $n = 400$	1	1	0.292	0.164	0.084	**0.049**	0.144
$p = 150$; $n = 100$	1	0.980	0.137	0.078	0.055	**0.043**	0.094
$p = 300$; $n = 200$	1	0.991	0.177	0.122	0.077	**0.044**	0.148
$p = 600$; $n = 400$	1	0.998	0.214	0.125	0.091	**0.052**	0.181

3. Estimating the noise variance in a generalized spiked model

We suppose the spiked model described in Section 2 has the following structure:

$$\Sigma = AA' + \Psi, \tag{3.1}$$

where A is a $p \times M$ matrix, the eigenvalues of $A'A$ are M distinct elements, and the eigenvalues of Ψ are groups of the general population eigenvalues.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

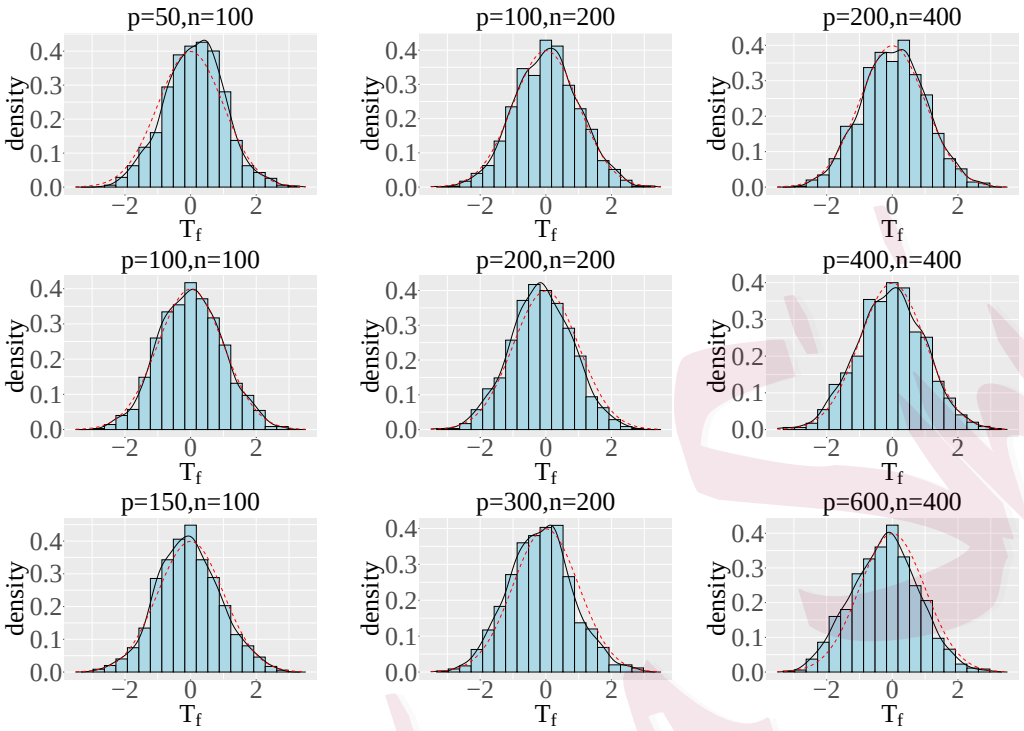

Figure 1: Model 2 under a gamma assumption.

Thus, the spectrum of Σ is denoted as

$$\sigma^2(\underbrace{\tilde{\alpha}_1,\cdots,\tilde{\alpha}_1}_{m_1},\cdots,\underbrace{\tilde{\alpha}_k,\cdots,\tilde{\alpha}_k}_{m_k},\cdots,\underbrace{r_1,\cdots,r_1,\cdots,r_s,\cdots,r_s}_{p-M}), \tag{3.2}$$

where $m_1+\cdots+m_k=M$, and both M and s are fixed small numbers.

It is relatively common for general population eigenvalues to be divided into several groups with fixed distinct spikes. Thus, we assume that the LSD of Σ excluding the spikes, denoted as $H(t)$, follows a probability distribution that takes the value $r_i\sigma^2$ with probability ω_i, for $i=1,\cdots,s$, where ω_1+

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

$\cdots + \omega_s = 1$.

In traditional statistical theory, the statistic

$$\hat{\sigma}^2 = \frac{1}{(p-M)(\omega_1 r_1 + \cdots + \omega_s r_s)} \left(\sum_{j=1}^{p} l_j - \sum_{j \in \mathcal{J}_k, k=1}^{K} l_j \right) \tag{3.3}$$

is a reasonable estimate of the noise variance σ^2, where l_j is a sample eigenvalue of Σ and $\mathcal{J}_k$ is a set of ranks of the spikes.

As is well known, when the dimensionality p is large relative to the sample size, the sample spiked eigenvalues do not converge to the population eigenvalues. Thus, the estimator (3.3) has a negative bias. Using the CLT proposed in Theorem 1, we establish the following CLT for the estimator $\hat{\sigma}^2$ in the high-dimensional setting, which can be used to identify the bias of estimator (3.3).

Theorem 2. *For the spiked model (3.1) with the spectrum in (3.2), we assume that the assumptions in Theorem 1 hold and $c_n = p/n \to c > 0$ when both the dimensionality p and the sample size n go to infinity. Then, we have that*

$$\nu_x^{-\frac{1}{2}} \left\{ (\hat{\sigma}^2 - \sigma^2)(p-M) \sum_{i=1}^{s} \omega_i r_i + b_(\tilde{\alpha}_k, \sigma^2) - \mu_x \right\} \Rightarrow \mathcal{N}(0,1),$$

where

$$b_(\tilde{\alpha}_k, \sigma^2) = \sum_{k=1}^{K} \sum_{i=1}^{s} \frac{m_k c \tilde{\alpha}_k \sigma^2 r_i \omega_i}{\tilde{\alpha}_k - r_i},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

$$
\begin{aligned}
\mu_x =& -\frac{q}{2\pi i}\oint \frac{c\underline{m}^2(z)\big[c\underline{m}(z)\int t\{1+t\underline{m}(z)\}^{-1}\mathrm{d}H(t)-1\big]}{\big[1-c\int \underline{m}^2(z)t^2\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)\big]^2}\\
&\times \int t^2\{1+t\underline{m}(z)\}^{-3}\mathrm{d}H(t)\mathrm{d}z\\
&-\frac{\beta c}{2\pi i}\oint \underline{m}^2(z)\big[-1+c\underline{m}(z)\int t\{1+t\underline{m}(z)\}^{-1}\mathrm{d}H(t)\big]\\
&\times \frac{\int t\{1+t\underline{m}(z)\}^{-1}\mathrm{d}H(t)\int\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)}{1-c\int \underline{m}^2(z)t^2\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)}\mathrm{d}z,\\
\nu_x =& -\frac{q+1}{4\pi^2}\oint\oint \frac{\big[-\underline{m}^{-1}(z_1)+c\int t\{1+t\underline{m}(z_1)\}^{-1}\mathrm{d}H(t)\big]}{\{\underline{m}(z_1)-\underline{m}(z_2)\}^2}\\
&\times\big[-\underline{m}^{-1}(z_2)+c\int t\{1+t\underline{m}(z_2)\}^{-1}\mathrm{d}H(t)\big]\mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2)\\
&-\frac{\beta c}{4\pi^2}\oint\oint\big[-\underline{m}^{-1}(z_1)+c\int t\{1+t\underline{m}(z_1)\}^{-1}\mathrm{d}H(t)\big]\\
&\times\big[-\underline{m}^{-1}(z_2)+c\int t\{1+t\underline{m}(z_2)\}^{-1}\mathrm{d}H(t)\big]\\
&\times\int\frac{t\mathrm{d}H(t)}{\{1+t\underline{m}(z_1)\}^2}\cdot\int\frac{t\mathrm{d}H(t)}{\{1+t\underline{m}(z_2)\}^2}\,\mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2).
\end{aligned}
$$

Note that $b(\tilde{\alpha}_k,\sigma^2)$ depends on the number m_k, the LSD $H(t)$, and the values of the population spikes, which are most likely unknown in practice; refer to the literature discussed in Section 2 on how to estimate these values.

We now use the above theorem to correct the bias of $\hat{\sigma}^2$. As shown in Theorem 2, the bias of the estimator is also related to the unknown parameter σ^2, which we estimate; a plug-in estimator is given in the following corollary.

Corollary 1. *For the spiked model (3.1), from Theorem 2, a bias-corrected*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

plug-in estimator is given by

$$\hat{\sigma}_c^2 = \hat{\sigma}^2 + \frac{b_(\tilde{\alpha}_k, \hat{\sigma}^2) - \mu_x}{(p-M)\sum_{i=1}^{s} \omega_i r_i}. \tag{3.4}$$

Therefore, the asymptotic distribution of $\hat{\sigma}_c^2$ is a natural consequence of Theorem 2.

Theorem 3. *If the conditions in Theorem 2 all hold, then the following holds:*

$$\nu_x^{-\frac{1}{2}}(p-M)\sum_{i=1}^{s} \omega_i r_i(\hat{\sigma}_c^2 - \sigma^2) \Rightarrow \mathcal{N}(0,1).$$

3.1 Simulation study for the estimation of the noise variance

For the estimation of the noise variance in the spiked model (3.1), we establish the following models:

Model 3 Assume that $\Sigma = \sigma^2\Lambda$, where Λ is defined in Model 1 and $\sigma^2 = 4$.

Model 4 Assume that $\Sigma = \sigma^2 U_0 \Lambda U_0^*$, where U_0 and Λ are defined in Model 2 and $\sigma^2 = 4$.

We use simulations to compare our proposed estimator $\hat{\sigma}_c^2$ in (3.4) with other estimation methods, such as the maximum likelihood estimation (MLE) $\hat{\sigma}^2$ defined in (3.3), the estimator $\hat{\sigma}_*^2$ presented by Passemier et al. (2017), the estimator $\hat{\sigma}_{us}^2$ presented by Ulfarsson and Solo (2008), and the estimator $\hat{\sigma}_m^2$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

presented by Johnstone and Lu (2009). The Gaussian and gamma population assumptions in Section 2 remain. The mean absolute error (MAE) and mean squared error (MSE) of these estimators for Model 4 are reported in Tables 3 and 4 for 1000 replicates. Similar results for Model 3 are presented in the Supplementary Material.

As shown by the simulated results, our proposed method yields the lowest MAEs and MSEs for the various populations and models. Furthermore, the advantage of our estimation becomes increasingly evident as the dimensionality increases. The estimator $\hat{\sigma}_*^2$ presented by Passemier et al. (2017) performs well for the diagonal population covariance matrix with only extremely large spikes. Their method involves a consistent estimate of $\tilde{\alpha}_k$, obtained by solving the equation $l_j \to \sigma^2\{\tilde{\alpha}_k + c\tilde{\alpha}_k/(\tilde{\alpha}_k - 1)\}$. However, for the general spiked matrix with both extremely large and extremely small spikes, their method always yields estimates close to one for the small spikes 0.1 and 0.2. Thus, the estimates of σ^2 are ineffective.

4. Testing the equality of the smallest roots in a probabilistic principal component analysis model

Suppose that the observable covariance matrix

$$\Sigma = AA' + \sigma^2 I \tag{4.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Table 3: MAEs and MSEs of $\hat{\sigma}_c^2$ compared with those of existing estimators of the noise variance for Model 4 under a Gaussian assumption.

Estimators		$\hat{\sigma}_c^2$	$\hat{\sigma}^2$	$\hat{\sigma}_*^2$	$\hat{\sigma}_{US}^2$	$\hat{\sigma}_m^2$
$p = 50;\ n = 100$	MAE	**0.0672**	0.1091	6.87×10^2	0.1195	3.3 049
	MSE	**0.0071**	0.0166	6.51×10^5	0.0220	11.103
$p = 100;\ n = 200$	MAE	**0.0335**	0.0569	2.77×10^2	0.0647	1.7115
	MSE	**0.0018**	0.0045	1.17×10^5	0.0065	2.9538
$p = 200;\ n = 400$	MAE	**0.0159**	0.0277	8.51×10^1	0.0335	0.7907
	MSE	**0.0004**	0.0011	1.59×10^4	0.0017	0.6279
$p = 100;\ n = 100$	MAE	**0.0549**	0.1154	1.09×10^3	0.2239	1.6165
	MSE	**0.0062**	0.0165	1.66×10^6	0.0628	2.6569
$p = 200;\ n = 200$	MAE	**0.0257**	0.0551	5.13×10^2	0.1102	0.7743
	MSE	**0.0010**	0.0038	3.65×10^5	0.0156	0.6047
$p = 400;\ n = 400$	MAE	**0.0127**	0.0265	1.98×10^2	0.0558	0.4261
	MSE	**0.0003**	0.0009	5.58×10^4	0.0041	0.1824
$p = 150;\ n = 100$	MAE	**0.0398**	0.1137	2.02×10^3	3.2501	1.0087
	MSE	**0.0025**	0.0150	4.07×10^6	10.564	1.0320
$p = 300;\ n = 200$	MAE	**0.0195**	0.0573	7.73×10^2	3.2277	0.5286
	MSE	**0.0006**	0.0038	5.98×10^5	10.418	0.2816
$p = 600;\ n = 400$	MAE	**0.0097**	0.0289	3.77×10^2	3.2167	0.2778
	MSE	**0.0001**	0.0009	1.42×10^5	10.347	0.0776

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Table 4: MAEs and MSEs of $\hat{\sigma}_c^2$ compared with those of existing estimators of the noise variance for Model 4 under a gamma assumption.

Estimators		$\hat{\sigma}_c^2$	$\hat{\sigma}^2$	$\hat{\sigma}_*^2$	$\hat{\sigma}_{US}^2$	$\hat{\sigma}_m^2$
$p = 50;\ n = 100$	MAE	**0.0834**	0.1226	6.93×10^2	0.1407	3.0216
	MSE	**0.0108**	0.0219	6.54×10^5	0.0308	9.3263
$p = 100;\ n = 200$	MAE	**0.0426**	0.0625	2.44×10^2	0.0727	1.5157
	MSE	**0.0028**	0.0057	9.48×10^4	0.0083	2.3219
$p = 200;\ n = 400$	MAE	**0.0223**	0.0314	8.53×10^1	0.0386	0.8012
	MSE	**0.0008**	0.0014	1.59×10^4	0.0023	0.6454
$p = 100;\ n = 100$	MAE	**0.0706**	0.1149	1.12×10^3	0.2546	1.5130
	MSE	**0.0102**	0.0178	1.63×10^6	0.0796	2.3255
$p = 200;\ n = 200$	MAE	**0.0322**	0.0588	4.85×10^2	0.1301	0.7568
	MSE	**0.0019**	0.0046	3.28×10^5	0.0212	0.5788
$p = 400;\ n = 400$	MAE	**0.0167**	0.0281	1.99×10^2	0.0647	0.4352
	MSE	**0.0008**	0.0011	5.63×10^4	0.0054	0.1905
$p = 150;\ n = 100$	MAE	**0.0511**	0.1183	2.02×10^3	3.2568	1.0796
	MSE	**0.0041**	0.0175	4.06×10^6	10.608	1.1852
$p = 300;\ n = 200$	MAE	**0.0263**	0.0576	7.73×10^2	3.2311	0.5576
	MSE	**0.0011**	0.0042	5.98×10^5	10.439	0.3144
$p = 600;\ n = 400$	MAE	**0.0126**	0.0293	3.77×10^2	3.2184	0.2839
	MSE	**0.0003**	0.0011	1.42×10^5	10.358	0.0813

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

has a characteristic root of σ^2 with multiplicity $p-M$, where $A'A$ is the positive semidefinite matrix of rank M. We denote the population eigenvalues of Σ as $\lambda_1, \cdots, \lambda_p$, in descending order. We then test the null hypothesis that

$$\mathcal{H}_\prime : \lambda_{M+1} = \cdots = \lambda_p. \tag{4.2}$$

This is equivalent to the null hypothesis that $\Sigma = AA' + \sigma^2 I$ when $A'A$ is positive semidefinite of rank M.

As shown in Section 11.7 in Anderson (2003), the pseudo-likelihood ratio criterion is the statistic L defined in (2.7), where l_i denotes a sample eigenvalue. Moreover, $-2\log L$ has a limiting χ^2-distribution with $(p-M+2)(p-M-1)/2$ degrees of freedom. However, this conclusion no longer holds when the dimension p goes to infinity. Passemier et al. (2017) propose a goodness-of-fit test for a probabilistic principal component analysis model of the form of (4.1). However, their result applies only to a Gaussian population. Therefore, we propose a corrected test statistic $-2\log L/\{n(p-M)\}$ and derive its limiting distribution using Theorem 1, which is widely used without a Gaussian assumption constraint.

Theorem 4. *For the test problem (4.2), we suppose that the standardized entries for the model (4.1) satisfy condition (2.6) and* $p/n = c_n \to c > 0$ *when both* n *and* p *go to infinity simultaneously. For the test statistic*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

$-2\log L/\{n(p-M)\}$, *we have that*

$$T_L = \nu_L^{-\frac{1}{2}} \left\{ -\frac{2\log L}{n(p-M)} - \log\left(\frac{b_{x,\sigma^2}}{p-M}\right) + \frac{b_{\log,\sigma^2} + \mu_{\log,\sigma^2}}{p-M} \right\} \Rightarrow \mathcal{N}(0,1),$$

where ν_L *is expressed by (4.7) and* b_{x,σ^2}, $b_{\log,\sigma^2}$, $\mu_{\log,\sigma^2}$, ν_{x,σ^2}, *and* $\nu_{\log,\sigma^2}$ *are defined in (4.4).*

Proof. We set α_k^*, for $k = 1, \ldots, K$, as the M nonzero eigenvalues of AA' with m_k multiplicity. The spikes of Σ in model (4.1) are $\sigma^2\tilde{\alpha}_k$, for $k = 1, \ldots, K$, where $\tilde{\alpha}_k = \alpha_k^*/\sigma^2 + 1$ and also has multiplicity m_k. We define $\Delta_1 = \sum_{i=M+1}^{p} l_i - b_{x,\sigma^2} - \mu_{x,\sigma^2}$ and $\Delta_2 = \sum_{i=M+1}^{p} \log l_i - b_{\log,\sigma^2} - \mu_{\log,\sigma^2}$. By Theorem 1, we have that

$$T_{x,\sigma^2} = \nu_{x,\sigma^2}^{-\frac{1}{2}} \Delta_1 \to \mathcal{N}(0,1) \quad \text{and} \quad T_{\log,\sigma^2} = \nu_{\log,\sigma^2}^{-\frac{1}{2}} \Delta_2 \to \mathcal{N}(0,1), \tag{4.3}$$

where

$$\begin{aligned}
b_{x,\sigma^2} &= (p-M)\sigma^2 - \sum_{i=1}^{K} \frac{m_k c\sigma^2\tilde{\alpha}_k}{\tilde{\alpha}_k - 1}, \ \mu_{x,\sigma^2} = 0, \ \nu_{x,\sigma^2} = (q+1+\beta)c\sigma^4, \\
b_{\log,\sigma^2} &= p\left\{\frac{(c-1)}{c}\log(1-c) - 1\right\} - \sum_{i=1}^{K} m_k \log(1 + \frac{c}{\alpha_k - 1}), \\
\mu_{\log,\sigma^2} &= \frac{q}{2}\log(1-c) - \frac{1}{2}\beta c, \quad \nu_{\log,\sigma^2} = -(q+1)\log(1-c) + \beta c.
\end{aligned} \tag{4.4}$$

The calculations are similar to those in Examples 1 and 2. Furthermore, by the expression of L given in (2.7) and the Taylor expansion, it follows that

$$-\frac{2\log L}{n(p-M)} = \log(\sum_{i=M+1}^{p} l_i) - \frac{1}{p-M}\sum_{i=M+1}^{p} \log l_i - \log(p-M)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

$$= \log(\Delta_1 + b_{x,\sigma^2} + \mu_x) - \frac{1}{p-M}(\Delta_2 + b_{\log} + \mu_{\log}) - \log(p-M)$$
$$= \log\left(\frac{b_{x,\sigma^2} + \mu_{x,\sigma^2}}{p-M}\right) + \frac{\Delta_1}{b_{x,\sigma^2} + \mu_{x,\sigma^2}} - \frac{\Delta_2}{p-M} - \frac{b_{\log,\sigma^2} + \mu_{\log,\sigma^2}}{p-M}. \quad (4.5)$$

Based on equation (4.3), we have

$$\frac{\Delta_1}{b_{x,\sigma^2} + \mu_x} - \frac{\Delta_2}{p-M} \to \mathcal{N}(0, \nu_L), \quad (4.6)$$

where

$$\nu_L = \frac{\nu_{x,\sigma^2}(p - M - 2b_{x,\sigma^2})}{(p-M)b_{x,\sigma^2}^2} + \frac{\nu_{\log,\sigma^2}}{(p-M)^2} \quad (4.7)$$

is calculated from (4.4) and (4.6).

Thus, by (4.5) and (4.6), the proof is complete.

4.1 Simulation study for testing the equality of the smallest roots

We use simulations to compare our proposed test statistic T_L with the classical pseudo-likelihood ratio test statistic (T_{PLR}) and the corrected likelihood ratio test (CLRT) presented by Passemier et al. (2017). These test methods all rely on the pseudo-likelihood function, and therefore have good statistical properties, but can be used only for the case $p < n$, owing to their correlation with the log function. Thus, to expand the application of our method, we also include the test statistic T_{x,σ^2} in (4.3). The value of the test statistic of the CLRT presented by Passemier et al. (2017) cannot be calcu-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

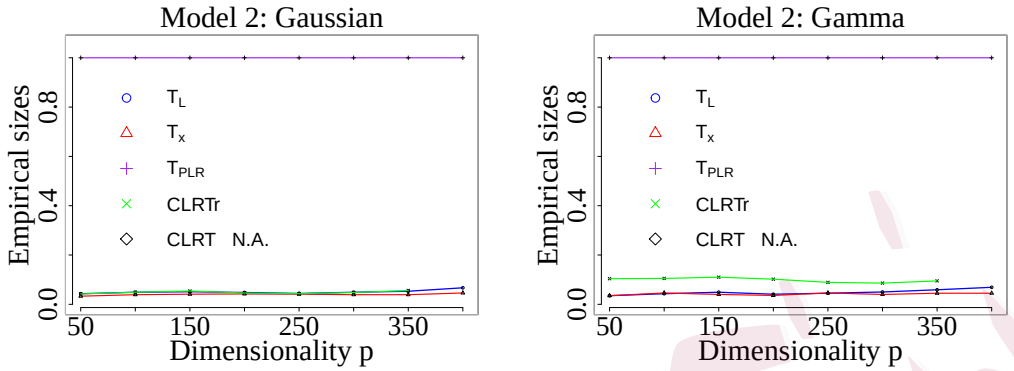

Figure 2: Empirical sizes of the competitive tests for hypothesis (4.2) when $n = 500$ and $p/n < 1$.

lated under our general model assumptions, as mentioned in Section 3.1. We use CLRT$_r$ to represent their test method, with their estimated $\tilde{\alpha}_k$ replaced by the real values of $\tilde{\alpha}_k$. Models 1 and 2 and the population assumptions in Section 2 are used again here. The empirical sizes of the competitive tests for hypothesis (4.2) are calculated for 1000 replicates. The simulated results for Model 2 are presented in Figures 2 and 3. Similar figures for Model 1 are included in the Supplementary Material.

Our proposed test statistics T_L and T_x provide the empirical sizes around the selected test level of 5%. Moreover, T_x can be applied more broadly to the case with $p > n$. Furthermore, the classical test statistic T_{PLR} rejects the null hypothesis when the dimensionality increases. As ex-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

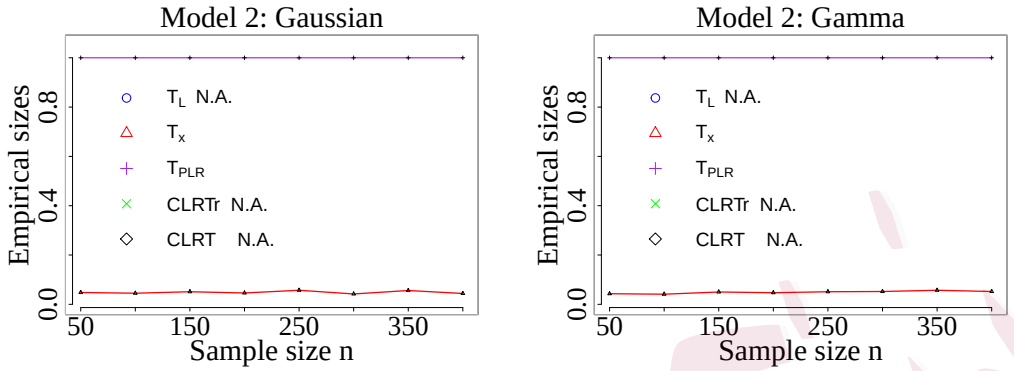

Figure 3: Empirical sizes of the competitive tests for hypothesis (4.2) when $p = 500$ and $p/n > 1$.

plained in Section 3.1, the CLRT is not available (N.A.) under our model assumptions, which has both extremely large and small spikes. Even if we used the true values of the spikes instead of their estimates in the CLRT, there are still problems. First, in the case of the gamma population, the empirical size of CLRT_r is significantly higher than the given test level. Second, for the high-dimension case $p = 400$ and $n = 500$, we still cannot calculate the value of the CLRT_r test statistic.

5. Real-Data Analysis

To demonstrate the feasibility of our proposed test method, we examine two real data sets. The first is an environmental data set for countries, freely

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

available from the website https://www.kaggle.com/zanderventer/environmental-variables-for-world-countries. Because country-level social and economic statistics are often limited to socio-economic data, this data set enables us to use environmental statistics to predict social and economic data. Determining how many environmental variables have a significant impact on socioeconomic status is an important problem. The data set consists of 243 countries and 27 environmental variables. We apply our testing method to determine the number of spikes in the covariance matrix generated from the standardized data, based on 188 observations without missing values. The p-values of the sequential tests are listed in Table 5. We infer that the true value of the number of spikes appears at the first inflection point of the p-value, where the first local maximum value occurs, enabling us to determine the number of spikes and provide appropriate estimates. The estimates of the values of the spikes are also included.

The second data set is the Wisconsin Breast Cancer Diagnosis data set, downloaded from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. This data set contains 569 instances, with 357 benign (62.7%) and 212 malignant (37.3%) cases of breast cancer. We compute 30 real-valued input features for each cell nucleus, as well as two nominal features: ID number and diagnosis. This database is

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

a standardized version of the original Wisconsin Breast Cancer Diagnosis data set, and we use our proposed testing method to determine the number of spikes for all instances. The test results are listed in Table 6.

Table 5: Estimates of the number and values of the population spikes for the environmental data set by country.

Values of M_0:	1	2	3	4	5	6	7
p-values:	0	0.0211	0.3498	0.6976	**0.9488**	0.4211	02205
Estimated number:					**5**		
Estimated values of spikes:		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	
		9.7848	7.0903	2.1086	1.7522	1.3634	

Table 6: Estimates of the number and values of the population spikes for the Wisconsin Breast Cancer Diagnosis data set.

Values of M_0:	1	2	3	4	5	6
p-values:	0	0	8.11×10^{-10}	**0.1026**	1.18×10^{-14}	0
Estimated number:				**4**		
Estimated values of spikes:		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	
		13.1817	5.6174	2.7219	1.9264	

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

6. Conclusion

We have established a universal test for the number of spikes in a high-dimensional generalized spiked model, with assumptions that are more relaxed than those of previous tests. We applied our method to two typical statistical problems, and proved its effectiveness using simulation results. In this study, we focus on the one-sample spiked model related to the covariance matrix. In future work, we will examine the two-sample spiked model with the Fisher matrix.

Supplementary Material

The supplementary material for "A universal test on spikes in a high-dimensional generalized spiked model and its applications" is available online and includes some simulation results as well as detailed proofs for Examples 1 and 2.

Acknowledgments

We thank LetPub (www.letpub.com) for the linguistic assistance and pre-submission expert review. The author was supported by NSFC Grant No.11971371 and the Fundamental Research Funds for the Central Universities and Natural Science Foundation of Shaanxi Province Grant No.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

2020JM-049.

Appendix

A.1 Proof of Theorem 1

To prove Theorem 1, we first need to generalize the CLT for the linear spectral statistic (LSS) of a sample covariance matrix in Bai and Silverstein (2004) and Jiang (2016). Similar to their works, the LSS for our generalized sample covariance matrix is also expressed as

$$\begin{aligned}\sum_{j=1}^{p} f(l_j) &= p\int f(x)\mathrm{d}F_n(x)\\ &= p\int f(x)\mathrm{d}(F_n - F^{c_n,H_n})(x) + p\int f(x)\mathrm{d}F^{c_n,H_n}(x)\\ &= G_n(x) + p\int f(x)\mathrm{d}F^{c_n,H_n}(x),\end{aligned}$$

where $F_n(x)$ is the empirical spectral distribution of the sample covariance matrix S, and $F^{c,H}(x)$ is the corresponding limiting spectral distribution. Denote $F^{c_n,H_n}(x)$ as the analogue of $F^{c,H}(x)$, except that the parameter c and H are replaced by c_n and H_n.

Under different finite 4th moment assumptions, both of Bai and Silverstein (2004) and Jiang (2016) proved that the process

$$G_n(x) = \sum_{j=1}^{p} f(l_j) - p\int f(x)\mathrm{d}F^{c_n,H_n}(x)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

would converge to a Gaussian random variable with some specific mean and variance. For example, when the condition of $\mathrm{E}|x_{11}|^4 = q + 2$ was assumed with $q = 1$ for the real case and $q = 0$ for the complex case, Bai and Silverstein (2004) showed that

$$G_n(x) \Rightarrow \mathcal{N}\left(\mu_{f,H}^{(1)}, \quad \nu_{f,H}^{(1)}\right), \tag{A.1}$$

where

$$\begin{aligned}
\mu_{f,H}^{(1)} =& -\frac{q}{2\pi i}\oint f(z)\frac{c\int \underline{m}^3(z)t^2\{1+t\underline{m}(z)\}^{-3}\mathrm{d}H(t)}{\left[1-c\int \underline{m}^2(z)t^2\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)\right]^2}\mathrm{d}z \\
\nu_{f,H}^{(1)} =& -\frac{q+1}{4\pi^2}\oint\oint \frac{f(z_1)f(z_2)}{\{\underline{m}(z_1)-\underline{m}(z_2)\}^2}\mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2).
\end{aligned} \tag{A.2}$$

Furthermore, Jiang (2016) extended their work to a wider range of application to the non-Gaussian populations. Under the assumption of $\mathrm{E}|x_{11}|^4 < \infty$ and $\beta = \mathrm{E}|x_{11}|^4 - q - 2$, Jiang (2016) claimed that

$$G_n(x) \Rightarrow \mathcal{N}\left(\mu_{f,H}^{(1)} + \mu_{f,H}^{(2)}, \quad \nu_{f,H}^{(1)} + \nu_{f,H}^{(2)}\right), \tag{A.3}$$

where $\mu_{f,H}^{(2)}$ and $\nu_{f,H}^{(2)}$ are compensations for the difference between Gaussian and non-Gaussian populations, and expressed as below,

$$\begin{aligned}
\mu_{f,H}^{(2)} =& -\frac{\beta c}{2\pi i}\oint f(z)\frac{\underline{m}^3(z)\int t\{1+t\underline{m}(z)\}^{-1}\mathrm{d}H(t)\cdot\int\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)}{1-c\int \underline{m}^2(z)t^2\{1+t\underline{m}(z)\}^{-2}\mathrm{d}H(t)}\mathrm{d}z, \\
\nu_{f,H}^{(2)} =& -\frac{\beta c}{4\pi^2}\oint\oint f(z_1)f(z_2)\int \frac{t\mathrm{d}H(t)}{\{1+t\underline{m}(z_1)\}^2}\int \frac{t\mathrm{d}H(t)}{\{1+t\underline{m}(z_2)\}^2}\ \mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2).
\end{aligned} \tag{A.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

We find that both of the proofs in Bai and Silverstein (2004) and Jiang (2016) depend on following formula, i.e

$$(x_t^* A x_t - \mathrm{tr}(A))^2 = \sum_{i=1}^{p} (\mathrm{E}|x_{it}|^4 - |\mathrm{E}x_{it}^2|^2 - 2)a_{ii} + \mathrm{tr}(A_x A_x^\top) + \mathrm{tr}(A_x^2), \quad \text{(A.5)}$$

where $A = (a_{ij})$ is a $p \times p$ matrix, and $A_x = (\mathrm{E}x_{it}^2 a_{ij})$. All of the compensation for the non-Gaussian populations come from the first term of (A.5). Bai and Silverstein (2004) supposed that $\mathrm{E}|x_{it}|^4 = 3$ for real case and $\mathrm{E}|x_{it}|^4 = 2$ for complex case, then the first item is 0. In Jiang (2016), they assumed that $\mathrm{E}|x_{it}|^4$'s are the same and bounded, then the coefficient of the non-Gaussian compensation is $\beta = \mathrm{E}|x_{11}|^4 - |\mathrm{E}x_{11}^2|^2 - 2$. For our generalized model, the CLT for $G_n(x)$ also relies heavily on the equations with the same form in (A.5). We take the following item as an example to illustrate our result,

$$\zeta = \left[\frac{1}{n}\{x_t^* T_p^* A_n^{-1} T_p x_t - \mathrm{tr}(T_p^* A_n^{-1} T_p)\}\right]^2,$$

where $A_n = S - \lambda I - \frac{1}{n} T_p x_t x_t^* T_p^*$ and λ is an eigenvalue of S defined in (2.3). By the decomposition of T_p in (2.1), we let $\xi_t = 1/\sqrt{n} U^* x_t$, then

$$\zeta = \left\{\xi_t^* B_n^{-1} \xi_t - \frac{1}{n}\mathrm{tr}\left(B_n^{-1}\right)\right\}^2,$$

where

$$B_n^{-1} = \begin{pmatrix} D_1^{\frac{1}{2}} & 0 \\ 0 & D_2^{\frac{1}{2}} \end{pmatrix} V^* A_n^{-1} V \begin{pmatrix} D_1^{\frac{1}{2}} & 0 \\ 0 & D_2^{\frac{1}{2}} \end{pmatrix}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

It is obvious that the matrix B_n^{-1} is bounded even if the spiked eigenvalues of S converge to infinity. Thus, if the bounded 4th moment condition is assumed, our result is identical to the one in Jiang (2016). When the 4th moment of x_{it} may not exist, by the equation (A.5), we have

$$\begin{aligned}\zeta &= \sum_{i=1}^{p}(\mathrm{E}|\xi_{it}|^4 - |\mathrm{E}\xi_{it}^2|^2 - 2)\check{b}_{ii} + (|\mathrm{E}\xi_{it}^2|^2 + 1)\mathrm{tr}(B_n^2) \\ &= \sum_{i=1}^{p}(\sum_{j=1}^{p}|u_{ji}|^4\mathrm{E}|x_{11}|^4 - \sum_{j=1}^{p}|u_{ji}|^2|\mathrm{E}x_{11}^2|^2 - 2)\check{b}_{ii} + (\sum_{j=1}^{p}|u_{ji}|^2|\mathrm{E}x_{11}^2|^2 + 1)\mathrm{tr}(B_n^2) \\ &= \sum_{i=1}^{p}(\sum_{j=1}^{p}|u_{ji}|^4\mathrm{E}|x_{11}|^4 - |\mathrm{E}x_{11}^2|^2 - 2)\check{b}_{ii} + (|\mathrm{E}x_{11}^2|^2 + 1)\mathrm{tr}(B_n^2)\end{aligned}$$

due to the properties of Hermitian matrix U, i.e. $\sum_{j=1}^{p} u_{ji}^2 = 1$ and $\sum_{j_1 \neq j_2} u_{j_1 i}u_{j_2 i} = 0$, and $\check{b}_{ii}$ is the ith diagonal element of B_n^{-1}.

Let $\hat{x}_{ij} = x_{ij}I(|x_{ij}| < \eta_n\sqrt{n})$ and $\tilde{x}_{ij} = (\hat{x}_{ij} - \mathrm{E}\hat{x}_{ij})/\sigma_n$ with $\sigma_n^2 = \mathrm{E}\,|\hat{x}_{ij} - \mathrm{E}\hat{x}_{ij}|^2$, where $\eta_n \to 0$ with a slow rate. It was demonstrated in Jiang and Bai (2021a) that it is equivalent to replace the entries of X with the truncated and renormalized ones under the Assumption (a). So when Assumption (b) holds for one side, and we let

$$\beta = \lim \sum_{j=1}^{p} |u_{ji}|^4\mathrm{E}\{|x_{11}|^4 I(|x_{11}| \leq \sqrt{n}) - q - 2\},$$

then $\zeta = \beta\sum_{i=1}^{p}\check{b}_{ii} + (q+1)\mathrm{tr}(B_n^2)$, which is identical to the result in Jiang (2016), except that the coefficient β is replaced by the limit. For the other

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

side, if Assumption (b) is not met, but the Assumption (b*) is valid, then

$$\begin{aligned}\beta &= \sum_{j=1}^{p} |u_{ji}|^4 \mathrm{E}\{|x_{11}|^4 I(|x_{11}| \le \sqrt{n}) - q - 2 \\ &\le \max_{1\le i\le M, 1\le j\le p} |u_{ji}|^2 \mathrm{E}\big\{|x_{11}|^4 I(|x_{11}| < \sqrt{n}) - q - 2\big\} \to 0,\end{aligned}$$

because only the eigenvectors $u_i, i = 1, \cdots, M$ corresponding to the extreme eigenvalues. Thus the proof is completed.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis. 3rd ed.* Wiley New York.

Anderson, T.W. and Rubin, H. (1956). Statistical Inference in Factor Analysis. *In: Neyman, J., Ed., Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, **5**, Berkeley, 111–150.

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

Bao, Z. G., Hu, J., Pan, G. M. and Zhou, W. (2019). Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case. *The Annals of Statistics*, **47(1)**, 612–640.

Bai, Z. D. and Silverstein, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32(1)**, 553–605.

Bai, Z. D. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matri-*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

ces. Springer Series in Statistics, Springer-Verlag, New York, ISSN: 0172-7397.

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

Bai, Z. D. and Yao, J. F. (2008). Central limit theorems for eigenvalues in a spiked population model. *Annales de l'Institut Henri Poincaré—Probabilités et Statistiques*, **44(3)**, 447–474.

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

Cai, T. T., Han, X. and Pan, G. M. (2020). Limiting laws for divergent spiked eigenvalues and largest non-spiked eigenvalue of sample covariance matrices. *The Annals of Statistics*, **48(3)**, 1255–1280.

Fan, J. and Wang, W. (2017). Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *The Annals of Statistics*, **45(3)**, 1342–1374.

Jiang, D. D. (2016). Tests for large dimensional covariance structure based on Rao's Score test.*Journal of Multivariate Analysis*, **152**, 28–39.

Jiang, D. D. and Bai, Z. D. (2021). Generalized four moment theorem and an application to CLT for spiked eigenvalues of large-dimensional covariance matrices.*Bernoulli*, **27(1)**, 274–294.

Jiang, D. D. and Bai, Z. D. (2021). Partial generalized four moment theorem revisited.*Bernoulli*, Forthcoming. https://doi.org/10.3150/20-BEJ1310.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

analysis. *The Annals of Statistics*, **29**, 295–327.

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

Johnstone, I. M. and Onatski, A. (2020). Testing in high-dimensional spiked models. *The Annals of Statistics*, **48(3)**, 1231–1254.

Kritchman, S. and Nadler, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chem. Int. Lab. Syst.*, **94**, 19–32.

Li, W. M., Chen, J. Q., Qin, Y. L., Bai, Z. D. and Yao, J. F. (2013). Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *Journal of Statistical Planning and Inference*, **143(11)**,1887–1897.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, **77**, 1447–1479.

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

Passemier, D. and Yao, J. F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Rand. Matr. Theor. Appl.,* **1**, article 1150002.

Passemier, D., Li, Z. and Yao, J. F. (2017). On estimation of the noise variance in high dimensional probabilistic principal component analysis. *J. R. Statist. Soc. B,* **79(1)**, 51–67.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0346

Ulfarsson, M. O. and Solo, V. (2008). Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Process.*, **56**, 5804–5816.

Zheng, T. K., Ma, Y. C. and Lin, X. H. (2021). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. *arXiv:2006.00436v1.*

School of Mathematics and Statistics, Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an, Shannxi, 710049, P.R. China.

E-mail: (jiangdd@xjtu.edu.cn)