Bernoulli **27**(1), 2021, 274–294
https://doi.org/10.3150/20-BEJ1237

Generalized four moment theorem and an application to CLT for spiked eigenvalues of high-dimensional covariance matrices

DANDAN JIANG[1] and ZHIDONG BAI[2]

[1]*School of Mathematics and Statistics, Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an 710049, China. E-mail: jiangdd@xjtu.edu.cn*
[2]*KLASMOE and School of Mathematics and Statistics, Northeast Normal University, No. 5268 People's Street, Changchun 130024, China. E-mail: baizd@nenu.edu.cn*

We consider a more generalized spiked covariance matrix, which is a general non-negative definite matrix with the spiked eigenvalues scattered into spaces of a few bulks and the largest ones allowed to tend to infinity. The study is split into two cases by whether the maximum absolute value of the eigenvector of the corresponding spikes tends to zero or not. On one hand, if it is zero, a Generalized Four Moment Theorem (G4MT) is proposed by relaxing the matching of the 3rd and the 4th moment to the tail probability decaying with certain rate, which shows the universality of the asymptotic law for the spiked eigenvalues of the generalized spiked covariance model. On the other hand, if it is not zero, the matches of the third and fourth moments in usual four moment theorem are weakened to only requiring the match of the 4th moment. Moreover, by applying the results to the Central Limit Theorem (CLT) for the spiked eigenvalues of the generalized spiked covariance model, we successively remove the restrictive condition of block wise diagonal assumption on the population covariance matrix in the previous works. This condition implies an unrealistic fact that the spiked eigenvalues and bulked eigenvalues are generated by independent variables, respectively. Thus, the new CLT will have much better application domain.

Keywords: central limit theorem; generalized four moment theorem; high-dimensional covariance matrix; random matrix theory; spiked model

1. Introduction

The study on the universality conjecture for the spectral statistics of random matrices, which is motivated by similar phenomena in physics, has been one of the key topics in random matrix theory. It not only plays an important role in the local field of statistics, but has also been widely used in many other fields, such as mathematical physics, combinatorics and computing science. In this paper, we are going to propose a Generalized Four Moment Theorem (G4MT) to prove the universality of the asymptotic law for the spiked eigenvalues of generalized spiked covariance matrices, and then apply it to the Central Limit Theorem (CLT) for the spiked eigenvalues of the generalized spiked model in a general case.

1.1. Background of universality

As well known, universality has been conjectured by many statisticians since the 1960s, including Wigner [29], Dyson [13], and Mehta [21]; it states that local statistics are universal, implying that the conclusions hold not only for the Gaussian Unitary Ensemble (GUE) but also the general Wigner random matrix. It provides new ideas and techniques for the research of random matrix theory, which

implies that to prove one result suitable for Non-Gaussian case, it is sufficient to show the same result under the Gaussian assumption if the universality is true.

The similar universality phenomena of the bulk of the spectrum has been also investigated in many studies, such as Soshnikov [26], Johansson [19], Ben Arous and Péché [11], Erdős *et al.* [14], Erdős *et al.* [15]. More recently, Tao and Vu [27] showed the universality of the asymptotic law for the local spectral statistics of the Wigner matrix by the so called Four Moment Theorem, which assumes that the moments of the entries match that of the complex standardized Gaussian ensemble up to the 4th order and requires the C_0 condition satisfying the uniform exponential decay to hold. Although they asserted that the fine spacing statistics of a random Hermitian matrix in the bulk of the spectrum are only sensitive to the first four moments of the entries, they also conjectured that it may be possible to reduce the number of matching moments in their theorem.

Inspired by these previous works, the G4MT is proposed by replacing the condition of matching the 3rd and 4th moments by a tail probability as detailed in Assumption (b). Then the universality of the asymptotic law for the bulks of spiked eigenvalues of generalized covariance matrices is automatically proved by the proposed G4MT. As an application, we also apply the proposed G4MT to the CLT for the spiked eigenvalues in the generalized spiked covariance model.

1.2. Related works of spiked model

The spiked model in high dimensional settings is originated from the common phenomenon of large or even huge dimensionality p compared to the sample size n, occurring in many modern scientific fields, such as wireless communication, gene expression and climate studies. It was first proposed by Johnstone [20] under the assumptions of high dimensionality and an identity population covariance matrix with fixed and relatively small spikes. Then some impressive works are devoted to investigating on the limiting properties of the spiked eigenvalues under this simplified assumption, including Baik, Ben Arous and Péché [9], Baik and Silverstein [10], Paul [24], Bai and Yao [4], etc.

To improve the simplified assumptions, Bai and Yao [5] contributed to deal with a more general spiked model, in which a condition of the diagonal block independence and finite 4th moments are assumed. Efforts have also been devoted to Principal Component Analysis (PCA) or Factor Analysis (FA) as a different way to improve the work on the spiked population model. For example, Bai and Ng [1], Hoyle and Rattray [16], Onatski [22] and so on. The more general works are the recent contributions from Wang and Fan [28] and Cai, Han and Pan [12], which both study the asymptotic distributions of the spiked eigenvalues and eigenvectors of a general covariance matrix. However, the result of Wang and Fan [28] only has one threshold for the spiked eigenvalues. More importantly, their main theorems are involved with an unspecified "$O_p(\cdot)$" term, because they study the difference between the ratio λ_i/α_i and 1, where λ_i is the corresponding sample eigenvalue. Furthermore, both of the works in Wang and Fan [28] and Cai, Han and Pan [12] require the bounded 4th moments and the condition $p/(n\alpha_i) \to 0$, with $\alpha_i, i = 1, \ldots, K$ being the spikes, so that it limits the relationship between the dimensionality and the spikes. On the basis of these works, we further consider a general spiked covariance matrix, by applying the proposed G4MT to the CLT for its spiked eigenvalues, we give the explicit CLT for the spiked eigenvalues of high-dimensional generalized covariance matrices.

1.3. Highlights of the paper

Highlights are mainly in two aspects: The universality and the CLT of spiked model. In the first aspect, it takes several advantages as follows: First, when proving the universality of the asymptotic law for the bulk of the spiked eigenvalues, it only requires the condition of matching moments up to the 2nd order

and a rate $o(x^{-4})$ of the tail probability $P(|X| \geq x)$ as $x \to \infty$, which is a necessary and sufficient condition in the weak convergence of the largest eigenvalue. Second, we reduce the study of universality of an asymptotic law of the normalized spiked eigenvalues to the asymptotic law of a low-dimensional matrix, unlike Tao and Vu [27], which considers a general function of a finite number of eigenvalues of a large dimensional Wigner matrix under the strong C_0 condition as well as assumptions on the partial derivatives. As a by-product, we get rid of the restrictive C_0 condition with uniform exponential decay.

In the other aspect, the proposed CLT demonstrates several advantages as below: First, the spiked covariance matrix we considered is a general non-negative definite matrix with the spectrum formulated in (2.1). Automatically, the diagonal block independent assumption given in Bai and Yao [4,5] is removed. In fact, it is an unrealistic condition in practice, which is added to avoid the difficulty caused by the dependence of spiked eigenvalues with respect to non-spiked ones. Many people have raised the issue that whether the diagonalizing assumption is necessary, but it is open due to technical difficulties until solved in this paper. Second, our method permits the spiked eigenvalues to be scattered into a few bulks, any of which are larger than their related left-threshold or smaller than their related right-threshold. So our focused work is extended to a generalized case with a few pairs of thresholds. Finally, the spiked eigenvalues and the population 4th moments are not necessarily required to be bounded in our work, thus meeting the actual cases better.

1.4. Outline of the paper

The rest of the paper is arranged as follows: In Section 2, the problem is described in a generalized setting, and the phase transition for the spiked eigenvalues of generalized covariance matrix is also presented. Section 3 gives the main results of the G4MT and applies it to the CLT for the spiked eigenvalues of the generalized spiked covariance matrix in high-dimensional setting. In Section 4, simulations are conducted to evaluate our work comparing with the existing works. Then, an application to determining the number of the spikes and real data analysis are also discussed in Section 5. Finally, we draw a conclusion in the Section 6. Detailed proofs are all provided in the Supplementary file [18].

2. Phase transition for the spiked eigenvalues of generalized covariance matrix

Consider the random samples $\mathbf{T}_p\mathbf{X}$, where

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = (x_{ij}), \quad 1 \leq i \leq p, 1 \leq j \leq n,$$

and $\mathbf{T}_p$ is a $p \times p$ deterministic matrix. Thus $\mathbf{T}_p\mathbf{T}_p^* = \mathbf{\Sigma}$ is the population covariance matrix, which can be seen as a general non-negative definite matrix with the spectrum formed as

$$\rho_{p,1}, \ldots, \rho_{p,j}, \ldots, \rho_{p,p} \tag{2.1}$$

in descending order. Let $\rho_{p,j_k+1}, \ldots \rho_{p,j_k+m_k}$ be equal to $\alpha_k, k = 1, \ldots, K$, respectively, where $\mathcal{J}_k = \{j_k + 1, \ldots, j_k + m_k\}$ is the set of ranks of the m_k-ple eigenvalue α_k in the array (2.1). Then $\alpha_1, \ldots, \alpha_K$ with multiplicity $m_k, k = 1, \ldots, K$, respectively, satisfying $m_1 + \cdots + m_K = M$, a fixed integer, are the population spiked eigenvalues of $\mathbf{\Sigma}$ lined arbitrarily in groups among all the eigenvalues.

Define the corresponding sample covariance matrix of the observations $\mathbf{T}_p\mathbf{X}$ as

$$\mathbf{S} = \frac{1}{n}\mathbf{T}_p\mathbf{X}\mathbf{X}^*\mathbf{T}_p^*, \tag{2.2}$$

then $\mathbf{S}$ is the so-called generalized spiked sample covariance matrix.

Define the singular value decomposition of $\mathbf{T}_p$ as

$$\mathbf{T}_p = \mathbf{V}\begin{pmatrix} \mathbf{D}_1^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2^{\frac{1}{2}} \end{pmatrix}\mathbf{U}^*, \tag{2.3}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices, $\mathbf{D}_1$ is a diagonal matrix of the M spiked eigenvalues and $\mathbf{D}_2$ is the diagonal matrix of the non-spiked eigenvalues with bounded components. Since the investigation on the limiting distribution of the spiked eigenvalues of the sample covariance matrix depends on the characteristic equation $|\lambda\mathbf{I} - \mathbf{S}| = 0$, it is obvious that it only involves the right unitary matrix $\mathbf{U}$ but not the left one.

Denote the eigenvalues of a $p \times p$ matrix $\mathbf{A}$ by $\{l_j(\mathbf{A})\}$. The sample eigenvalues of the generalized spiked sample covariance matrix $\mathbf{S}$ are sorted in descending order as

$$l_1(\mathbf{S}), \ldots, l_j(\mathbf{S}), \ldots, l_p(\mathbf{S}).$$

To consider the limiting distribution of the spiked eigenvalues of a generalized sample covariance matrix $\mathbf{S}$, it is necessary to determine the following Assumptions (a)–(e):

Assumption (a). The double array $\{x_{ij}, i, j = 1, 2, \ldots\}$ consist of i.i.d. random variables with mean 0 and variance 1. Furthermore, $\mathrm{E}x_{ij}^2 = 0$ for the complex case, where both x's and $\mathbf{T}_p$ are complex.

Assumption (b). Suppose that

$$\lim_{\tau\to\infty} \tau^4 \mathrm{P}\big(|x_{ij}| > \tau\big) = 0$$

for the i.i.d. sample $(x_{i1}, \ldots, x_{in})$, $i = 1, \ldots, p$, where the 4th moments may unnecessarily exist.

Assumption (c). The $p \times p$ matrix $\mathbf{\Sigma} = \mathbf{T}_p\mathbf{T}_p^*$ forms a sequence of population covariance matrices $\{\mathbf{\Sigma}_p\}$ and $\mathbf{T}_p$ admits the singular decomposition (2.3). The matrix $\mathbf{D}_2$ is bounded in the spectral norm. Moreover, denote the empirical spectral distribution (ESD) of $\mathbf{\Sigma}$ as H_n, which tends to a proper probability measure H as $p \to \infty$.

Assumption (d). Suppose that

$$\max_{t,s} |u_{ts}|^2 \mathrm{E}\big\{|x_{11}|^4 I\big(|x_{11}| < \sqrt{n}\big) - \mu\big\} \to 0, \tag{2.4}$$

where for some constant μ, $I(\cdot)$ is the indicator function and $\mathbf{U}_1 = (u_{ts})_{t=1,\ldots,p;s=1,\ldots,M}$ is the first M columns of matrix $\mathbf{U}$ defined in (2.3). The detailed explanation of Assumption (d) can be found in the Supplement A.

Assumption (e). Assuming that $p/n = c_n \to c > 0$ and both n and p go to infinity simultaneously, the spiked eigenvalues of the matrix $\mathbf{\Sigma}$, $\alpha_1, \ldots, \alpha_K$ with multiplicities $m_1, \ldots, m_K$ laying outside the

support of H, satisfy $\phi'(\alpha_k) > 0$ for $1 \le k \le K$, where

$$\phi(x) = x\left(1 + c\int \frac{t}{x-t}\,\mathrm{d}H(t)\right)$$

is detailed in the following Proposition 2.1.

The phase transition for each spiked eigenvalue of a generalized sample covariance matrix is detailed in the following Proposition. To avoid the sample spikes tending to a common limit, we regulate the spikes by the separation condition

$$\min_{j\neq k}\left|\frac{\alpha_k}{\alpha_j} - 1\right| > d, \tag{2.5}$$

for some constant $d > 0$, when the phase transition and the CLT for the spiked eigenvalues of a generalized spiked covariance matrix are studied.

For each population spiked eigenvalue α_k with multiplicity m_k and the associated sample eigenvalues $\{l_j(\mathbf{S}),\, j \in \mathcal{J}_k\}$, $k = 1, \ldots, K$, we have

Proposition 2.1. *For the spiked sample covariance matrix* $\mathbf{S}$ *given in* (2.2), *assume that* $p/n = c_n \to c > 0$ *and both the dimensionality* p *and the sample size* n *grow to infinity simultaneously. For any population spiked eigenvalue* α_k, $(k = 1, \ldots, K)$, *let*

$$\psi_k = \begin{cases} \phi(\alpha_k), & \text{if } \phi'(\alpha_k) > 0, \\ \phi(\underline{\alpha}_k), & \text{if there exists } \underline{\alpha}_k \text{ such that } \phi'(\underline{\alpha}_k) = 0 \\ & \text{and } \phi'(t) < 0, \text{ for all } \alpha_k \le t < \underline{\alpha}_k, \\ \phi(\overline{\alpha}_k), & \text{if there exists } \overline{\alpha}_k \text{ such that } \phi'(\overline{\alpha}_k) = 0 \\ & \text{and } \phi'(s) < 0, \text{ for all } \overline{\alpha}_k < s \le \alpha_k, \end{cases}$$

where

$$\phi_k := \phi(\alpha_k) = \alpha_k\left(1 + c\int \frac{t}{\alpha_k - t}\,\mathrm{d}H(t)\right). \tag{2.6}$$

Then, it holds that for all $j \in \mathcal{J}_k$, $\{l_j/\psi_k - 1\}$ *almost surely converges to* 0 *under the bounded* 4*th-moment assumption. The conclusion also holds in probability under the Assumption* (d).

The Proposition 2.1 theoretically shows that the diagonal block independent assumption

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_M & 0 \\ 0 & \mathbf{V}_{p-M} \end{pmatrix}$$

in Bai and Yao [5] can be removed. The proof of Proposition 2.1 can be easily obtained following the truncation procedure and the G4MT, which are presented in the next section. By the truncation procedure, the limiting behavior of the sample spiked eigenvalues are the same in probability for both the cases of the bounded 4th-moment assumption and Assumption (d). By the G4MT, it is reasonable to assume the Gaussian entries from $\mathbf{X}$; then, Proposition 2.1 is proved by the almost sure convergence and the exact separation of eigenvalues in Bai and Silverstein [7].

Remark 2.1. Since the convergence of $c_n \to c$ and $H_n \to H$ may be very slow, the difference $\sqrt{n}(l_j - \psi_k)$ may not have a limiting distribution. So, we usually use

$$\phi_{n,k} := \phi_n(\alpha_k) = \alpha_k\left(1 + c_n \int \frac{t}{\alpha_k - t}\,\mathrm{d}H_n(t)\right), \tag{2.7}$$

instead of ϕ_k in ψ_k, in particular during the process of CLT. Then, we only require $c_n = p/n$, and both the dimensionality p and the sample size n grow to infinity simultaneously, but not necessarily in proportion.

3. Main results

The main results are in two key points: First, it is the G4MT, which shows that the samples satisfying the Assumptions (a)–(e) lead to the same asymptotic distributions of the spiked eigenvalues of a generalized spiked covariance matrix. Second, it is the CLT for the spiked eigenvalues of a high-dimensional generalized covariance matrix under our relaxed assumptions. For ease of reading and understanding, the G4MT is introduced during its application to the CLT for the spiked eigenvalues of a generalized covariance matrix. The proof of G4MT will be postponed to Section D in the Supplement for the consistency of reading. Before that, we also give some explanations of the truncation procedure as below.

3.1. Truncation

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. We can demonstrate that it is equivalent to replace the entries of $\mathbf{X}$ with the truncated and renormalized ones by Assumption (b). Details of the proof are presented in Supplement B and the convergence rates of arbitrary moments of $\tilde{x}_{ij}$ are depicted.

Therefore, we only need to consider the limiting distribution of the spiked eigenvalues of $\tilde{\mathbf{S}}$, which is generated from the entries truncated at $\eta_n\sqrt{n}$, centralized and renormalized. For simplicity, it is equivalent to assume that $|x_{ij}| < \eta_n\sqrt{n}$, $\mathrm{E}x_{ij} = 0$, $\mathrm{E}|x_{ij}^2| = 1$, and Assumption (b) is satisfied for the real case. But it cannot meet the requirement of $\mathrm{E}x_{ij}^2 = 0$ for the complex case; instead, one can show that $\mathrm{E}x_{ij}^2 = o(n^{-1})$.

3.2. CLT for the spiked eigenvalues of generalized covariance matrix

As seen from the Proposition 2.1, there is a packet of m_k consecutive sample eigenvalues $\{l_j(\mathbf{S}), j \in \mathcal{J}_k\}$ converging to a limit ψ_k laying outside the support of the limiting spectral distribution (LSD), $F^{c,H}$, of $\mathbf{S}$. Since the spiked eigenvalues may be allowed to tend to infinity in our work, and the difference between $l_j(\mathbf{S})$ and ψ_k make convergence very slow as mentioned in Remark 2.1, we consider the renormalized random vector

$$\boldsymbol{\gamma}_k = (\gamma_{kj})' = \left(\sqrt{n}\left(\frac{l_j(\mathbf{S})}{\phi_{n,k}} - 1\right), j \in \mathcal{J}_k\right)' \tag{3.1}$$

which can be seen as an improved version of the one in Bai and Yao [5]. Then, we are going to propose a CLT for $(\gamma_{kj}, j \in \mathcal{J}_k)'$ for a general case. Before that, we introduce some of the characteristics of the sample spikes first.

For the generalized spiked covariance matrix $\mathbf{\Sigma} = \mathbf{T}_p\mathbf{T}_p^*$, consider the corresponding sample covariance matrix $\mathbf{S} = \mathbf{T}_p\mathbf{S}_x\mathbf{T}_p^*$, where $\mathbf{S}_x = n^{-1}\mathbf{X}\mathbf{X}^*$ is the standard sample covariance with sample size n. By singular value decomposition of $\mathbf{T}_p$ in (2.3), the eigen-equation is

$$0 = |\lambda\mathbf{I} - \mathbf{S}| = \left|\lambda\mathbf{I} - \mathbf{V}\begin{pmatrix}\mathbf{D}_1^{\frac{1}{2}} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_2^{\frac{1}{2}}\end{pmatrix}\mathbf{U}^*\mathbf{S}_x\mathbf{U}\begin{pmatrix}\mathbf{D}_1^{\frac{1}{2}} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_2^{\frac{1}{2}}\end{pmatrix}\mathbf{V}^*\right|,$$

where $\mathbf{I}$ denotes the identity matrix with suitable dimension. If no confusion, we omit the subscript of the identity matrix. Set $\mathbf{Q} = \mathbf{U}^*\mathbf{S}_x\mathbf{U}$, and partition it in the same way as the form

$$\begin{pmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{pmatrix} \triangleq \begin{pmatrix}\mathbf{U}_1^*\mathbf{S}_x\mathbf{U}_1 & \mathbf{U}_1^*\mathbf{S}_x\mathbf{U}_2\\ \mathbf{U}_2^*\mathbf{S}_x\mathbf{U}_1 & \mathbf{U}_2^*\mathbf{S}_x\mathbf{U}_2\end{pmatrix},$$

then we have

$$\begin{aligned}0 &= \left|\lambda\mathbf{I} - \begin{pmatrix}\mathbf{D}_1^{\frac{1}{2}}\mathbf{Q}_{11}\mathbf{D}_1^{\frac{1}{2}} & \mathbf{D}_1^{\frac{1}{2}}\mathbf{Q}_{12}\mathbf{D}_2^{\frac{1}{2}}\\ \mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_{21}\mathbf{D}_1^{\frac{1}{2}} & \mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_{22}\mathbf{D}_2^{\frac{1}{2}}\end{pmatrix}\right|\\ &= \left|\lambda\mathbf{I} - \mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_{22}\mathbf{D}_2^{\frac{1}{2}}\right|\left|\lambda\mathbf{I} - \mathbf{D}_1^{\frac{1}{2}}\mathbf{Q}_{11}\mathbf{D}_1^{\frac{1}{2}} - \mathbf{D}_1^{\frac{1}{2}}\mathbf{Q}_{12}\mathbf{D}_2^{\frac{1}{2}}\big(\lambda\mathbf{I} - \mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_{22}\mathbf{D}_2^{\frac{1}{2}}\big)^{-1}\mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_{21}\mathbf{D}_1^{\frac{1}{2}}\right|.\end{aligned}$$

If we only consider the sample spiked eigenvalues of $\mathbf{S}$, l_j, $j \in \mathcal{J}_k$, $k = 1, \ldots, K$, then $|l_j\mathbf{I} - \mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_{22}\mathbf{D}_2^{\frac{1}{2}}| \neq 0$, hence

$$\begin{aligned}0 &= \left|l_j\mathbf{D}_1^{-1} - \frac{1}{n}\mathbf{U}_1^*\mathbf{X}\left\{\mathbf{I} + \frac{1}{n}\mathbf{X}^*\mathbf{U}_2\mathbf{D}_2^{\frac{1}{2}}\left(l_j\mathbf{I} - \frac{1}{n}\mathbf{D}_2^{\frac{1}{2}}\mathbf{U}_2^*\mathbf{X}\mathbf{X}^*\mathbf{U}_2\mathbf{D}_2^{\frac{1}{2}}\right)^{-1}\mathbf{D}_2^{\frac{1}{2}}\mathbf{U}_2^*\mathbf{X}\right\}\mathbf{X}^*\mathbf{U}_1\right|\\ &= \left|l_j\mathbf{D}_1^{-1} - \frac{l_j}{n}\mathbf{U}_1^*\mathbf{X}\left(l_j\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{U}_2\mathbf{D}_2\mathbf{U}_2^*\mathbf{X}\right)^{-1}\mathbf{X}^*\mathbf{U}_1\right|,\end{aligned} \tag{3.2}$$

where the last equality above follows from the identity $\mathbf{Z}(\mathbf{Z}^*\mathbf{Z} - \lambda\mathbf{I})^{-1}\mathbf{Z}^* = \mathbf{I} + \lambda(\mathbf{Z}\mathbf{Z}^* - \lambda\mathbf{I})^{-1}$. Define

$$\mathbf{\Omega}_M(\lambda, \mathbf{X}) = \frac{\lambda}{\sqrt{n}}\left[\mathrm{tr}\left\{\left(\lambda\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-1}\right\}\mathbf{I} - \mathbf{U}_1^*\mathbf{X}\left(\lambda\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-1}\mathbf{X}^*\mathbf{U}_1\right], \tag{3.3}$$

where $\mathbf{\Gamma} = \mathbf{U}_2\mathbf{D}_2\mathbf{U}_2^*$. Let

$$\theta_k = \phi_k^2\underline{m}_2(\phi_k), \tag{3.4}$$

where

$$\underline{m}_2(\lambda) = \int \frac{1}{(\lambda - x)^2}\,\mathrm{d}\underline{F}(x) \tag{3.5}$$

with $\underline{F}(x)$ being the LSD of the matrix $n^{-1}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}$. Set $\mathbf{\Omega}_{\phi_k}$ as an $M \times M$ Hermitian matrix, which follows the limiting distribution of $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$ and is detailed in Corollary 3.1. Then, the CLT for $(\gamma_{kj}, j \in \mathcal{J}_k)'$ for a general case is proposed in the following theorem.

Theorem 3.1. *Suppose that the Assumptions* (a)–(e) *hold with the constant* $\mu = 2 + q$ *in Assumption* (*d*), *where* $q = 1$ *for real case and* 0 *for complex. The random vector* $\boldsymbol{\gamma}_k$ *defined in* (3.1) *converges*

weakly to the joint distribution of the m_k eigenvalues of Gaussian random matrix

$$-\frac{1}{\theta_k}[\mathbf{\Omega}_{\phi_k}]_{kk}$$

where ϕ_k, $\phi_{n,k}$, θ_k are defined in (2.6), (2.7) and (3.4), respectively. Moreover, $[\mathbf{\Omega}_{\phi_k}]_{kk}$ is the kth diagonal block of $\mathbf{\Omega}_{\phi_k}$ corresponding to the indices $\{i, j \in \mathcal{J}_k\}$.

Proof. By the definition of (3.3), it follows from (3.2) that

$$0 = \left| l_j \mathbf{D}_1^{-1} - \frac{l_j}{n} \operatorname{tr}\left\{ \left(l_j \mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} \right\} \mathbf{I} + \frac{1}{\sqrt{n}} \mathbf{\Omega}_M(l_j, \mathbf{X}) \right|. \tag{3.6}$$

It seems natural to get that

$$\mathbf{\Omega}_M(l_j, \mathbf{X}) = \mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X}) + o_p(1)$$

since $(l_j - \phi_{n,k})/\phi_{n,k} \to 0$.

Thus, it follows from (3.6) that

$$\begin{aligned} &\left| \phi_{n,k} \mathbf{D}_1^{-1} - \frac{\phi_{n,k}}{n} \operatorname{tr}\left\{ \left(\phi_{n,k} \mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} \right\} \mathbf{I} \right. \\ &\qquad \left. + \mathbf{B}_1(l_j) + \mathbf{B}_2(l_j) + \frac{1}{\sqrt{n}} \mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X}) + o_p\left(\frac{1}{\sqrt{n}} \right) \right| \\ &\quad = 0, \end{aligned} \tag{3.7}$$

where

$$\mathbf{B}_1(l_j) = (l_j - \phi_{n,k})\mathbf{D}_1^{-1} = \frac{1}{\sqrt{n}} \phi_{n,k} \gamma_{kj} \mathbf{D}_1^{-1}, \tag{3.8}$$

$$\begin{aligned} \mathbf{B}_2(l_j) &= \frac{\phi_{n,k}}{n} \operatorname{tr}\left(\phi_{n,k}\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} \mathbf{I} - \frac{l_j}{n} \operatorname{tr}\left(l_j \mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} \mathbf{I} \\ &= \frac{\phi_{n,k}}{n} \left[\operatorname{tr}\left(\phi_{n,k}\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} \right. \\ &\quad \left. - \left(1 + n^{-\frac{1}{2}}\gamma_{kj}\right) \operatorname{tr}\left\{ \phi_{n,k}\left(1 + n^{-\frac{1}{2}}\gamma_{kj}\right)\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right\}^{-1} \right] \mathbf{I} \\ &= \frac{\phi_{n,k}}{n} \left[\operatorname{tr}\left(\phi_{n,k}\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} - \operatorname{tr}\left\{ \phi_{n,k}\left(1 + n^{-\frac{1}{2}}\gamma_{kj}\right)\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right\}^{-1} \right] \mathbf{I} \\ &\quad - n^{-\frac{1}{2}}\gamma_{kj}\phi_{n,k} \frac{1}{n} \operatorname{tr}\left[\left\{ \phi_{n,k}\left(1 + n^{-\frac{1}{2}}\gamma_{kj}\right)\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right\}^{-1} \right] \mathbf{I} \\ &= \frac{\phi_{n,k}^2 \gamma_{kj}}{n^{3/2}} \operatorname{tr}\left[\left(\phi_{n,k}\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} \left\{ \phi_{n,k}\left(1 + n^{-\frac{1}{2}}\gamma_{kj}\right)\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right\}^{-1} \right] \mathbf{I} \\ &\quad - \frac{\gamma_{kj}\phi_{n,k}}{n^{1/2}} \frac{1}{n} \operatorname{tr}\left\{ \left(\phi_{n,k}\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X} \right)^{-1} \right\} \mathbf{I} + o\left(n^{-\frac{1}{2}}\right) \end{aligned}$$

$$= \frac{\gamma_{kj}}{n^{1/2}}\left[\frac{\phi_{n,k}^2}{n}\operatorname{tr}\left\{\left(\phi_{n,k}\mathbf{I}-\frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-2}\right\}-\phi_{n,k}\frac{1}{n}\operatorname{tr}\left\{\left(\phi_{n,k}\mathbf{I}-\frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-1}\right\}\right]\mathbf{I}$$
$$+ o\big(n^{-\frac{1}{2}}\big)$$
$$= \frac{\gamma_{kj}}{n^{1/2}}\{\phi_k^2\underline{m}_2(\phi_k)+\phi_k\underline{m}(\phi_k)\}\mathbf{I}+o_p\big(n^{-\frac{1}{2}}\big), \tag{3.9}$$

with $\underline{m}_2$ defined in (3.5) and $\underline{m}(\lambda)=\int 1/(x-\lambda)\,\mathrm{d}\underline{F}(x)$. The calculations are based on the formula $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$ with A, B being two arbitrary $n\times n$ invertible matrices, and the facts that

$$\frac{1}{n}\operatorname{tr}\left(\phi_{n,k}\mathbf{I}-\frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-1}\to-\underline{m}(\phi_k);\qquad \frac{1}{n}\operatorname{tr}\left(\phi_{n,k}\mathbf{I}-\frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-2}\to\underline{m}_2(\phi_k).$$

Furthermore, if consider the kth diagonal block of the item

$$\phi_{n,k}\mathbf{D}_1^{-1}-\frac{\phi_{n,k}}{n}\operatorname{tr}\left\{\left(\phi_{n,k}\mathbf{I}-\frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-1}\right\}\mathbf{I}$$

in (3.7), denote the analogue of $\underline{m}$ as $\underline{m}_n$, with H substituted by the ESD H_n and c by c_n, which satisfies the equation

$$\phi_{n,k}=-\frac{1}{\underline{m}_n(\phi_{n,k})}+c_n\int\frac{t}{1+t\underline{m}_n(\phi_{n,k})}\,\mathrm{d}H_n(t)$$

by the definition of the Stieltjes transform. By the proof of Theorem 1.1 in Bai and Silverstein [8], it is found that

$$\frac{1}{n}\phi_{n,k}\operatorname{tr}\left(\phi_{n,k}\mathbf{I}-\frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-1}+\phi_{n,k}\underline{m}_n(\phi_{n,k})=o\left(\frac{1}{\sqrt{n}}\right). \tag{3.10}$$

Note that $\phi_{n,k}$ is the inverse of the Stieltjes transform $\underline{m}_n$ at $-1/\alpha_k$, we have $\underline{m}_n(\phi_{n,k})=-1/\alpha_k$, hence

$$\phi_{n,k}+\phi_{n,k}\underline{m}_n(\phi_{n,k})\alpha_k=0. \tag{3.11}$$

Therefore, to complete the proof of Theorem 3.1, it is needed to derive the limiting distributions of $\Omega_M(\phi_{n,k},\mathbf{X})$. So the theoretical tool named G4MT is established in the following theorem, which is used to prove the limiting distributions of $\Omega_M(\phi_{n,k},\mathbf{X})$. For the consistence of reading, we only introduce the theorem here, but postpone the proof to the Supplement D.

3.3. Generalized four moment theorem

The G4MT is established in the following theorem, which shows that the limiting distributions of the spiked eigenvalues of a generalized spiked covariance matrix is independent of the actual population distributions provided the samples to satisfy the Assumptions (a)–(e).

Theorem 3.2 (G4MT). *Assuming that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are two sets of double arrays satisfying Assumptions* (a)–(e), $\mathbf{X}$ *and* $\mathbf{Y}$ *should share same* μ *in condition* (d), *then it holds that* $\mathbf{\Omega}_M(\phi_{n,k},\mathbf{X})$ *and* $\mathbf{\Omega}_M(\phi_{n,k},\mathbf{Y})$ *have the same limiting distribution, provided one of them has.*

By Theorem 3.2, we may assume that $\mathbf{X}$ consists of entries of i.i.d. standard normal variables in deriving the limiting distribution of $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$. Namely, we have the following corollary, which is proved in the Supplement E.

Corollary 3.1. *If* $\mathbf{X}$ *satisfies the Assumptions* (a)–(e) *with* $\mu = 2 + q$ *in Assumption* (d), *let* θ_k *be defined as* (3.4), *then* $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$ *tends to a limiting distribution of an* $M \times M$ *Hermitian matrix* $\mathbf{\Omega}_{\phi_k}$, *where* $\mathbf{\Omega}_{\phi_k}/\sqrt{\theta_k}$ *is Gaussian Orthogonal Ensemble* (*GOE*) *for the real case*, *with the entries above the diagonal being* i.i.d.$\mathcal{N}(0, 1)$ *and the entries on the diagonal being* i.i.d.$\mathcal{N}(0, 2)$. *For the complex case*, *the* $\mathbf{\Omega}_{\phi_k}/\sqrt{\theta_k}$ *is GUE*, *whose diagonal entries are* i.i.d. *real* $\mathcal{N}(0, 1)$, *and the off diagonal entries are* i.i.d. *complex* $\mathcal{CN}(0, 1)$.

Remark 3.1. If the Assumption (d) is not met, it is weaken to the Assumption (d'), that is, all $\pi_{x,i_1j_1i_2j_2} = \lim \sum_{t=1}^{p} \bar{u}_{ti_1} u_{tj_1} u_{ti_2} \bar{u}_{tj_2} \mathrm{E}\{|x_{11}|^4 I(|x_{11}| \le \sqrt{n}) - 2 - q\}$ are finite, and $\mathbf{u}_i = (u_{1i}, \ldots, u_{pi})'$ are the ith column of the matrix $\mathbf{U}_1$. Then, the conclusion of this corollary remains hold, with the variances and covariances of the element ω_{ij} of $\mathbf{\Omega}_{\phi_k}$ given by

$$\mathrm{Cov}(\omega_{i_1,j_1}, \omega_{i_2,j_2}) = \begin{cases} (q+1)\theta_k + \pi_{x,iiii}\nu_k, & i_1 = j_1 = i_2 = j_2 = i; \\ \theta_k + \pi_{x,ijij}\nu_k, & i_1 = i_2 = i \neq j_1 = j_2 = j; \\ \pi_{x,i_1j_1i_2j_2}\nu_k, & \text{other cases} \end{cases}$$

where θ_k is define in (3.4), $\nu_k = \phi_k^2 \underline{m}^2(\phi_k)$.

For such cases, one may derive a partial G4MT by replacing the matrix $\mathbf{X}$ in $\mathbf{U}_2^*\mathbf{X}$ with $\mathbf{U}_2^*\mathbf{Y}$ as column to column and keeping $\mathbf{U}_1^*\mathbf{X}$ unchanged. The readers are reminded that in the definition of π_x function, the factor $\mathrm{E}|x_{11}|^4 I(|x_{11}| \le \sqrt{n})$ seems ought to be $\mathrm{E}|x_{11}|^4 I(|x_{11}| \le \eta_n\sqrt{n})$. However, it can be shown that the limit of π_x functions remain unchanged by using either one of the two. The derivation is detailed in the Supplement E.

3.4. Completing the proof of Theorem 3.1

Now, we continue to the previous proof of Theorem 3.1. For every sample spiked eigenvalue, l_j, $j \in \mathcal{J}_k$, $k = 1, \ldots, K$, it follows from (3.7) that

$$\begin{aligned} 0 = & \Big| \phi_{n,k}\mathbf{D}_1^{-1} - \frac{\phi_{n,k}}{n}\mathrm{tr}\left\{\left(\phi_{n,k}\mathbf{I} - \frac{1}{n}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}\right)^{-1}\right\}\mathbf{I} \\ & + \mathbf{B}_1(l_j) + \mathbf{B}_2(l_j) + \frac{1}{n^{1/2}}\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X}) + o_p\big(n^{-\frac{1}{2}}\big)\Big| \\ = & \Big| \phi_{n,k}\mathbf{D}^{-1} + \phi_{n,k}\underline{m}_n(\phi_{n,k})\mathbf{I}_M + \frac{1}{n^{1/2}}\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X}) \\ & + \frac{1}{n^{1/2}}\gamma_{kj}\big[\phi_{n,k}\mathbf{D}^{-1} + \big\{\phi_{n,k}^2\underline{m}_2(\phi_{n,k}) + \phi_{n,k}\underline{m}(\phi_{n,k})\big\}\mathbf{I}_M\big] + o_p\big(n^{-\frac{1}{2}}\big)\Big| \end{aligned} \tag{3.12}$$

by the equations (3.8), (3.9) and (3.10).

By the G4MT, we can derive the limiting distribution of $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$ under the assumption of Gaussian entries. Details of the proof for the limiting distribution of $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$ is provided in Supplement E. Therefore, applying Skorokhod strong representation theorem (see Skorohod [25], Hu and Bai [17]),

we may assume that the convergence of $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$ and (3.12) are in the sense of "almost surely" by choosing an appropriate probability space.

To be specific, by (3.12) and noting $\underline{m}_n(\phi_{n,k}) = -1/\alpha_k$, it yields

$$
\begin{aligned}
0 = &\left| \begin{pmatrix}
\frac{\phi_{n,k}}{\alpha_k}\left(\frac{\alpha_k}{\alpha_1} - 1\right)\mathbf{I}_{m_1} & 0 & \cdots & 0 \\
0 & \ddots & & \\
 & & \frac{\phi_{n,k}}{\alpha_k}\{1 + \alpha_k \underline{m}_n(\phi_{n,k})\}\mathbf{I}_{m_k} & \vdots \\
\vdots & & & \\
 & & \ddots & 0 \\
0 & \cdots & 0 & \frac{\phi_{n,k}}{\alpha_k}\left(\frac{\alpha_k}{\alpha_K} - 1\right)\mathbf{I}_{m_K}
\end{pmatrix} \right. \\
&+ \frac{\gamma_{kj}}{n^{1/2}} \begin{pmatrix}
\left(\frac{\phi_{n,k}}{\alpha_1} + \phi_{n,k}^2 \underline{m}_2 + \phi_{n,k}\underline{m}\right)\mathbf{I}_{m_1} & 0 & \cdots & 0 \\
0 & \ddots & & \\
 & & \phi_{n,k}^2 \underline{m}_2 & \vdots \\
\vdots & & & \\
 & & \ddots & 0 \\
0 & \cdots & 0 & \left(\frac{\phi_{n,k}}{\alpha_K} + \phi_{n,k}^2 \underline{m}_2 + \phi_{n,k}\underline{m}\right)\mathbf{I}_{m_K}
\end{pmatrix} \\
&\left. + \frac{1}{n^{1/2}} \mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X}) + o\left(n^{-\frac{1}{2}}\right) \right|.
\end{aligned}
$$

where $\underline{m}$, $\underline{m}_2$ are the simplified notations of $\underline{m}(\phi_{n,k})$ and $\underline{m}_2(\phi_{n,k})$, respectively.

For the population eigenvalues α_u in the uth diagonal block of $\mathbf{D}_1$, if $u \neq k$, then $\underline{m}_n(\phi_{n,k}) = -1/\alpha_k$ by the definition of $\phi_{n,k}$, hence $\phi_{n,k}\alpha_k^{-1}(\alpha_k\alpha_u^{-1} - 1)$ keeps away from 0, by the separation condition of spikes (2.5). Moreover, $\phi_{n,k}\alpha_k^{-1}\{1 + \alpha_k \underline{m}_n(\phi_{n,k})\} = 0$ by definition. Then, multiplying $n^{\frac{1}{4}}$ to the kth block row and kth block column of the above equation, by Lemma 4.1 in Bai *et al.* [6], it follows that as $n \to \infty$

$$
0 = \left| \left[\mathbf{\Omega}_M(\phi_k, \mathbf{X})\right]_{kk} + \lim \gamma_{kj}\left\{\phi_k^2 \underline{m}_2(\phi_k)\right\}\mathbf{I}_{m_k} \right|,
$$

where $[\cdot]_{kk}$ is the kth diagonal block of a matrix corresponding to the indices $\{i, j \in \mathcal{J}_k\}$. This shows that $(\gamma_{kj}\phi_k^2\underline{m}_2(\phi_k), j \in \mathcal{J}_k)'$ tends to the m_k eigenvalues of the $m_k \times m_k$ matrix $-[\mathbf{\Omega}_{\phi_k}]_{kk}$, or equivalently $(\gamma_{kj}, j \in \mathcal{J}_k)'$ tends to the eigenvalues of the $m_k \times m_k$ matrix $-[\mathbf{\Omega}_{\phi_k}]_{kk}/\theta_k$, where $\theta_k = \phi_k^2\underline{m}_2(\phi_k)$, and $\mathbf{\Omega}_{\phi_k}$ is the limit of $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$ defined in Corollary 3.1. Because limiting behavior keeps orders of the variables, we claim that the m_k ordered variables $(\gamma_{kj}, j \in \mathcal{J}_k)'$ tend to the m_k ordered eigenvalues of the matrix $-[\mathbf{\Omega}_{\phi_k}]_{kk}/\theta_k$.

By the strong representation theorem, we conclude that the m_k-dimensional real vector $(\gamma_{kj}, j \in \mathcal{J}_k)'$ converges weakly to the joint distribution of the m_k eigenvalues of the Gaussian random matrix

$$-\frac{1}{\theta_k}[\mathbf{\Omega}_{\phi_k}]_{kk}$$

for each distant generalized spiked eigenvalue. Then, the CLT for each distant spiked eigenvalue of a generalized covariance matrix is obtained. □

Remark 3.2. Suppose that $\mathbf{X}$ satisfies the Assumptions (a),(b),(c) and (e), with the Assumption (d) weakened as the existence of various limit π_x functions in Assumption (d'). Then all the conclusions of Theorem 3.1 still holds, but the limiting distribution of $\mathbf{\Omega}_M(\phi_{n,k}, \mathbf{X})$ turns to an $M \times M$ Hermitian Gaussian matrix $\mathbf{\Omega}_{\phi_k} = (\omega_{st})$ whose variances and covariances are defined in Remark 3.1.

This remark is used for the case of non-Gaussian assumptions when the population covariance matrix has a diagonal or diagonal block structure in the following simulations.

4. Simulation study

Simulations are conducted in this section to evaluate the performance of our proposed method. Four scenarios are considered including two cases of the population covariance matrix structure under two different population assumptions: On one hand, the *Case I* assumes that $\mathbf{\Sigma}$ is a diagonal matrix, where the Assumption (d) is not satisfied, but weakened as the Assumption (d'). On the other hand, the *Case II* is provided as a general form of $\mathbf{\Sigma}$, where the Assumption (d) holds. They are detailed as below:

Case I: The matrix $\mathbf{\Sigma} = \text{diag}(4, 3, 3, 0.2, 0.2, 0.1, 1, \ldots, 1)$ is a finite-rank perturbation of a identity matrix $\mathbf{I}_p$ with the spikes $(4, 3, 0.2, 0.1)$ of the multiplicity $(1, 2, 2, 1)$, thus $K = 4$ and $M = 6$.

Case II: The matrix $\mathbf{\Sigma} = \mathbf{U}_0\mathbf{\Lambda}\mathbf{U}_0^*$ is a general positive definite matrix, where $\mathbf{\Lambda}$ is a diagonal matrix with the spikes $(4, 3, 0.2, 0.1)$ of the multiplicity $(1, 2, 2, 1)$ as defined in *Case I* and $\mathbf{U}_0$ is the matrix composed of eigenvectors of the following matrix

$$\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \cdots & \cdots & & & \\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1 \end{pmatrix},$$

where $\rho = 0.5$.

For each case, the following population assumptions are studied:

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

Binomial Assumption. $\{x_{ij}\}$ are i.i.d. samples from the binary variables valued at $\{-1, 1\}$ with equal probability $1/2$, and $\pi_x = \mathrm{E}|x_{11}|^4 - 3 = -2$.

The simulated results are depicted as follows with 1000 replications at the values of $p = 500$, $n = 1000$. As described above, we have the spikes $\alpha_1 = 4$, $\alpha_2 = 3$, $\alpha_3 = 0.2$ and $\alpha_4 = 0.1$.

First, the Remark 3.2 is applied to the *Case I*. For the single population spikes $\alpha_1 = 4$ and $\alpha_4 = 0.1$, we obtain the limiting distributions

$$\gamma_k = \sqrt{n}\left(\frac{l_j(\mathbf{S})}{\phi_{n,k}} - 1\right) \to N\left(0, \sigma_k^2\right)$$

where $\phi_{n,1} = 4.667$, $\phi_{n,4} = 0.044$, and $\sigma_1^2 = 1.390$, $\sigma_4^2 = 3.950$ under the Gaussian Assumption, $\sigma_1^2 = 0.074$, $\sigma_4^2 = 2.414$ under the Binomial Assumption.

For the spikes $\alpha_2 = 3$ and $\alpha_3 = 0.2$ with multiplicity 2, we obtain that the two-dimensional random vector

$$\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2})' = \left(\sqrt{n}\left(\frac{l_j(\mathbf{S})}{\phi_{n,k}} - 1\right), \sqrt{n}\left(\frac{l_{j+1}(\mathbf{S})}{\phi_{n,k}} - 1\right)\right)', \quad k = 2, 3$$

converges to the eigenvalues of random matrix $-{\theta_k}^{-1}[\boldsymbol{\Omega}_{\phi_k}]_{22}$, where $\phi_{n,2} = 3.750$, $\theta_k = 1.770$ for the spike $\alpha_2 = 3$ and $\phi_{n,3} = 0.075$, $\theta_k = 0.633$ for the spike $\alpha_3 = 0.2$.

Since it is difficult to show a good fit by the contour plots of the empirical density, so it is better to choose a asymptotically normal marginal function to investigate. Because only the sum of their linear functions is normal, then it arrives at

$$\gamma_{2,s} = \frac{\gamma_{21} + \gamma_{22}}{2\sigma_{2,s}} + 0.5 \to N(0, 1);$$

where $\sigma_{2,s}^2 = 0.847$ under the Gaussian Assumption and $\sigma_{2,s}^2 = 0.088$ under the Binomial Assumption. and the constant 0.5 is an adjusted central parameter due to the effect of the multiple root with the multiplicity 2. The following are the same except for the sign. Then, it follows that

$$\gamma_{3,s} = \frac{\gamma_{31} + \gamma_{32}}{2\sigma_{3,s}} - 0.5 \to N(0, 1);$$

where $\sigma_{3,s}^2 = 2.370$ under the Gaussian Assumption and $\sigma_{3,s}^2 = 2.012$ under the Binomial Assumption.

Second, for the *Case II*, the Theorem 3.1 is used to calculate the limiting distributions. It is easily obtained that the calculated results of the both population assumptions are the same to the one of Gaussian Assumption in *Case I*, which can well fit their corresponding limiting behaviors under the *Case II* with different population assumptions.

As shown in the calculations and simulations, our approach provides the same results to the ones in Bai and Yao [5] under the Gaussian assumption. So we only show the two scenarios where our approach is superior to the method in Bai and Yao [5]. Firstly, our method performs slightly better for the non-Gaussian distribution even if the diagonal independent assumption in Bai and Yao [5] holds under the *Case I*. Because there is a missing item in their calculation of the variance. The simulated empirical distributions of $\gamma_{2,s}$, γ_4 from Binomial assumption under *Case I* are drawn in Figure 1 comparing to the ones of standardized $(l_2 + l_3)/2$, l_p in Bai and Yao [5], as well as their Gaussian limits. In addition, their corresponding limiting distributions are in dashed lines.

Moreover, our proposed results are obviously better than the ones in Bai and Yao [5] for the non-Gaussian population assumption in the *Case II*. The simulated empirical distributions of γ_1, $\gamma_{2,s}$ from

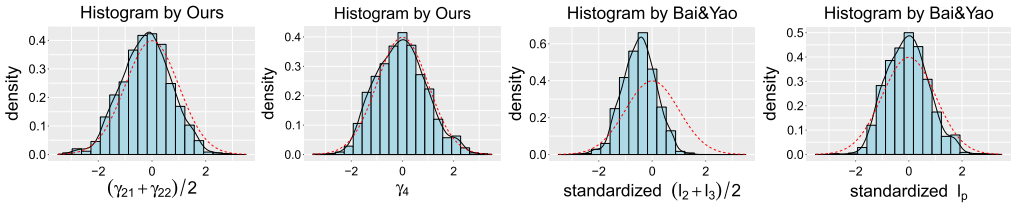

Figure 1. *Case I* under Binomial Assumption.

Binomial assumption under *Case II* are drawn in Figure 2 comparing to the ones of standardized l_1, $(l_2 + l_3)/2$ in Bai and Yao [5], as well as their Gaussian limits. As shown in the simulated results, the asymptotic distribution in Bai and Yao [5] performs not well for the non-Gaussian population assumption in the *Case II*. Because their method is involved with the 4th moment and the diagonal independent assumption. Therefore, it is reasonable to theoretically remove the diagonal independent restrictions in results of Bai and Yao [5] as illustrated in the simulations.

5. Application and real data analysis

5.1. An application to determine the number of the spikes

Since the spiked model is closely related to PCA, it has important applications to the statistical inferences in many scientific fields. For example, to reconstruct the original signals in wireless communication, to rebuild the observed assets into a low-dimensional set of unobserved variables, which are the factors in economics, and so on. One of the basic but important statistical inferences in these applications is to determine the number of principal components / signals / factors, that is, the number of spiked eigenvalues.

As formulated in (2.1), we propose to estimate the number of the spikes, M, by our result in Theorem 3.1. First, for every sample eigenvalue l_j, $j \in \mathcal{J}_k$, it follows from Theorem 3.1 that

$$\frac{\sqrt{n}}{\sigma_k}\left(\frac{l_j(\mathbf{S})}{\phi_{n,k}} - 1\right) \sim \mathcal{N}(0,1),$$

where $\sigma_k^2 = 2/\theta_k$ under our Assumption (a)–(e) and $\sigma_k^2 = (2\theta_k + \pi_{x,jjjj}\nu_k)/\theta_k^2$ under the assumptions of the diagonal or diagonal block independence with the bounded spikes and the 4th moments.

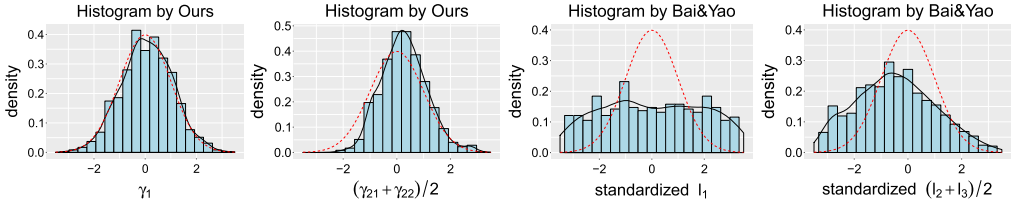

Figure 2. *Case II* under Binomial assumption.

Then, for every sample eigenvalue l_j, we can calculate an corresponding interval

$$C_j = \left[\left(\frac{z_{0.05}\sigma_k}{\sqrt{n}} + 1\right)\phi_k, \quad \left(\frac{z_{0.95}\sigma_k}{\sqrt{n}} + 1\right)\phi_k\right],$$

where $z_{0.05}$, $z_{0.95}$ are the 5% and 95% quantiles of the standard normal distribution. If $l_j \in C_j$, then it is concluded that the population eigenvalues in according to l_j is a spike; Otherwise, it is not a spike. Similarly, the same procedures are conducted for all the sample eigenvalues, and consequently a sequence of intervals $\{C_j, j = 1, \ldots, p\}$ are obtained. Therefore, we propose an estimator for the number of the spikes, M, as follows

$$\hat{M}_0 = \sum_{j=1}^{p} I_{(l_j \in C_j)}.$$

However, ϕ_k in (2.6) and σ_k^2 calculated by Theorem 3.1 or Remark 3.1 cannot be directly obtained by their expressions in practice, because they are involved with the unknown population spikes $\alpha_k, k = 1, \ldots, K$. Therefore, we refer to the work in Bai and Ding [3] and obtain the estimator of α_k. Then, we provide the estimated interval $\hat{C}_j$ and $\hat{M}_0 = \sum_{j=1}^{p} I_{(l_j \in \hat{C}_j)}$, which is feasible in practice.

In fact, by the first equation in (3.7), it asymptotically holds that $l_j + l_j \underline{m}(l_j)\alpha_k = 0$, so we get $\hat{\alpha}_k = -1/\underline{m}(l_j)$. Since the number of spikes is fixed, the LSD of $n^{-1}\mathbf{X}^*\mathbf{\Gamma}\mathbf{X}$ is approximately the same as the one of the matrix $n^{-1}\mathbf{X}^*\mathbf{U}\mathbf{D}\mathbf{U}^*\mathbf{X}$. Therefore, we further define $r_{ij} = |l_i - l_j|/\max(l_i, l_j)$ and let $\mathcal{J}_o = \{j : r_{ij} \le 0.2 \text{ for any } i = 1, \ldots, p.\}$, $\tilde{c} = (p - |\mathcal{J}_o|)/n$. Then we adopt

$$\hat{m}(l_j) = \frac{1}{p - |\mathcal{J}_o|} \sum_{i \notin \mathcal{J}_o; i=1}^{p} (l_i - l_j)^{-1},$$

which is a good estimator of $m(l_j)$. The setting $\mathcal{J}_0$ is selected to avoid the effect of multiple roots, which makes the estimations of the population spikes inaccurate. The constant 0.2 is a more suitable threshold value of the ratio based on our simulated results. Moreover, we obtain the estimator of $\underline{m}(l_j)$ as below $\hat{\underline{m}}(l_j) = -(1 - \tilde{c})/l_j + \tilde{c}\hat{m}(l_j)$. Finally, we obtain the estimator of α_k, which is expressed as

$$\hat{\alpha}_k = -1/\hat{\underline{m}}(l_j).$$

Without extra efforts, the following estimators are automatically obtained that $\hat{\phi}_k = \phi(\hat{\alpha}_k)$ and

$$\hat{m}(\phi_k) = \frac{1}{p}\sum_{i=1}^{p}(l_i - \hat{\phi}_k)^{-1}; \qquad \hat{\underline{m}}(\phi_k) = -\frac{1-c}{\hat{\phi}_k} + c\hat{m}(\hat{\phi}_k);$$

$$\hat{m}_2(\phi_k) = \frac{1}{p}\sum_{i=1}^{p}(l_i - \hat{\phi}_k)^{-2}; \qquad \hat{\underline{m}}_2(\phi_k) = \frac{1-c}{\hat{\phi}_k^2} + c\hat{m}_2(\hat{\phi}_k);$$

So the estimators of σ_k, ϕ_k for the renewal interval $\hat{C}_j$ can be expressed by the above estimations.

Through our approach, not only can we estimate the number of the spikes more accurately, but we can also give the estimations of the population spikes, as well as the limits of the sample spiked eigenvalues. More importantly, we can also provide the specific locations of these spikes.

5.2. Numerical results for Section 5.1

For the two cases of $\mathbf{\Sigma}$ designed in Section 4 with $M = 6$, we use the method provided in Section 5.1 to estimate the number of the population spikes under different population assumptions in Section 4.

To evaluate the performance of our approach, we shall compare it with some existing methods. Since the method in Onatski [22] provides a better estimator than that in Bai and Ng [1], and Cai, Han and Pan [12] shows that their approach performs better than that in both of Onatski [22] and Bai, Choi and Fujikoshi [2], so we only consider the procedure proposed in Cai, Han and Pan [12] and the method introduced by Passemier and Yao [23], which are simply denoted as CHP and PY in the tables, respectively.

The following Tables 1–4 report the estimator of the number of the spikes and its corresponding frequency by three methods. Furthermore, it provides the locations and estimates of the population spikes by our method. As shown in the tables, our method can give an accurate estimate of the number of the spikes in a large probability, while the other two methods fail to detect the very small spikes. That's because they both assume that the population spikes are the larger eigenvalues, satisfying that $\alpha_1 \geq \cdots \geq \alpha_M \geq \rho_{p,M+1} \geq \cdots \geq \rho_{p,p}$. However, it makes sense to detect all the spiked eigenvalues, including the minimal ones. For example, the original system with all the same eigenvalues has changed after the input of some signals. If we want to test which part in the system have changed, then it is equivalent to find out all the spiked eigenvalues. In addition, our method has an advantage over other methods, that is, it also presents the the estimations of the population spikes, and the specific locations of these spikes in the tables.

Table 1. Estimations about the spikes: *Case I* under Gaussian Assumption

				Frequency of $\hat{M}_0$				
	$\hat{M}_0$	1	2	3	4	5	**6**	7
$p = 200$	PY	0.358	0	0.642	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0	0.024	**0.943**	0.033
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		3.993	3.207	3.014	0.202	0.198	0.098	
				Frequency of $\hat{M}_0$				
$p = 400$	PY	0.371	0	0.629	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0	0.027	**0.928**	0.045
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		3.930	3.052	3.015	0.206	0.186	0.117	

Table 2. Estimations about the spikes: *Case I* under Binomial Assumption

				Frequency of $\hat{M}_0$				
	$\hat{M}_0$	1	2	3	4	5	**6**	7
$p = 200$	PY	0.622	0	0.378	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0	0.054	**0.943**	0.003
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		4.025	3.091	2.951	0.194	0.185	0.099	
				Frequency of $\hat{M}_0$				
$p = 400$	PY	0.640	0	0.360	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0.005	0.073	**0.910**	0.012
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		4.018	3.008	2.876	0.207	0.194	0.101	

5.3. Real data analysis

Now we apply the procedure of determining the number of the spikes proposed in Section 5.1 to the actual data titled as "Early stage of Indians Chronic Kidney Disease(CKD)"[1] The data came from records collected by a hospital in India over a period of about 2 months, which consists of 400 observations and 25 variables. The first 24 variables $X_1 \ldots, X_{24}$ are independent variables, which rerecord the various laboratory indicators and hospital records, including age, blood pressure (bp), specific gravity (sg), albumin (al), sugar (su), red blood cells (rbc), pus cell (pc), pus cell clumps (pcc), bacteria (ba), blood glucose random (bgr), blood urea (bu), serum creatinine (sc), sodium (sod), potassium (pot), hemoglobin (hemo), packed cell volume (pcv), white blood cell count (wc), red blood cell count (rc), hypertension (htn), diabetes mellitus (dm), coronary artery disease (cad), appetite (appet), pedal edema (pe), anemia (ane). The 25th variable is the dependent variable to indicate whether the patient has chronic kidney disease(ckd).

We apply our method to determine the number of the spikes of the covariance matrix $\mathbf{\Sigma}_0$ generated from the standardized data of the first 24 variables with 114 observations (For simplicity, we have only chosen 114 observations without missing values). Then, we obtain the following results in the Table 5.

As seen from the Table 5, if we define the singular value decomposition of $\mathbf{\Sigma}_0$ as $\mathbf{\Sigma}_0 = \mathbf{U}\mathbf{\Lambda}_0\mathbf{U}'$, and $\mathbf{u}_i$ is the ith column of the orthogonal matrix $\mathbf{U}$, then the factors generated from independent variables $\mathbf{X} = (X_1 \ldots, X_{24})'$ can be roughly divided into three groups: one group has a greater impact with larger

[1] The data is downloaded from https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

Table 3. Estimations about the spikes: *Case II* under Gaussian Assumption

				Frequency of $\hat{M}_0$				
	$\hat{M}_0$	1	2	3	4	5	**6**	7
$p = 200$	PY	0.375	0	0.625	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0	0.019	**0.950**	0.031
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		4.080	3.122	2.909	0.208	0.191	0.010	
	$\hat{M}_0$	1	2	3	4	5	**6**	7
$p = 400$	PY	0.356	0	0.644	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0	0.025	**0.927**	0.048
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		3.949	3.217	2.887	0.231	0.188	0.102	

spiked eigenvalues, like $\mathbf{u}_1'\mathbf{X}$, $\mathbf{u}_2'\mathbf{X}$; Another group of much weaker effects, like $\mathbf{u}_i'\mathbf{X}$, $i = 18, \ldots, 24$; The last group that may have most of the same effects, like $\mathbf{u}_i'\mathbf{X}$, $i = 3, \ldots, 17$. Furthermore, if we use the data with the missing values made up, the experimental results may be more accurate. To make up for missing values, one can use the miss Forest function in the package missForest.

6. Conclusion

In this paper, we propose a G4MT for a generalized spiked covariance matrix, which shows the universality of the asymptotic law for its spiked eigenvalues. Through the concrete example of the CLT of normalized spiked eigenvalues, we illustrate the basic idea and procedures of the G4MT to show the universality of a limiting result related to the large dimensional random matrices. Unlike Tao and Vu [27], we avoid the estimates of high-order partial derivatives of an implicit function to the entries of the random matrix, and thus, the strong condition C_0 of sub-exponential property is avoided. Moreover, the required 4th moment condition is reduced to a tail probability in Assumption (b), which is necessary for the existence of the largest eigenvalue limit. Without the constraint of the existence of the 4th moment, we only need a more regular and minor condition (2.4) on the elements of $\mathbf{U}_1$. On the one hand, our result has much wider applications than Bai and Yao [4], Bai and Yao [5]; on the other hand, the result of Bai and Yao [5] shows the necessity of the condition (2.4) for the total universality.

Table 4. Estimations about the spikes:*Case II* under Binomial Assumption

				Frequency of $\hat{M}_0$				
	$\hat{M}_0$	1	2	3	4	5	**6**	7
$p = 200$	PY	0.343	0	0.657	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0	0.018	**0.980**	0.002
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		3.838	3.275	2.922	0.216	0.192	0.098	
				Frequency of $\hat{M}_0$				
$p = 400$	PY	0.374	0.001	0.625	0	0	**0**	0
$n = 1000$	CHP	0	0	0	1	0	**0**	0
	Ours	0	0	0	0	0.041	**0.952**	0.007
				Locations				
				(1, 2, 3, 198, 199, 200)				
				Estimates of spikes				
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	
		4.123	3.211	3.001	0.216	0.195	0.096	

Acknowledgements

We are grateful to the Editor, the Associate Editors and referees for their constructive and helpful comments.

The first author was supported by Project 11971371 from NSFC.

The second author was supported by Project 11571067 from NSFC.

Supplementary Material

Supplement to "Generalized four moment theorem and an application to CLT for spiked eigenvalues of high-dimensional covariance matrices" (DOI: 10.3150/20-BEJ1237SUPP; .pdf). We provide the detailed explanation of Assumption (d), some necessary lemmas and the proofs of Theorem 3.2 and Corollary 3.1.

Table 5. Estimations of the number, sizes and locations of the spikes by the real data

Number:				9					
Location:				(1, 2, 18, 19, 20, 21, 22, 23, 24)					
Sizes:	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\alpha}_7$	$\hat{\alpha}_8$	$\hat{\alpha}_9$
	10.818	2.143	0.219	0.166	0.124	0.101	0.064	0.048	0.009

References

[1] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259 https://doi.org/10.1111/1468-0262.00273

[2] Bai, Z., Choi, K.P. and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *Ann. Statist.* **46** 1050–1076. MR3797996 https://doi.org/10.1214/17-AOS1577

[3] Bai, Z. and Ding, X. (2012). Estimation of spiked eigenvalues in spiked models. *Random Matrices Theory Appl.* **1** 1150011, 21. MR2934717 https://doi.org/10.1142/S2010326311500110

[4] Bai, Z. and Yao, J. (2008). Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** 447–474. MR2451053 https://doi.org/10.1214/07-AIHP118

[5] Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.* **106** 167–177. MR2887686 https://doi.org/10.1016/j.jmva.2011.10.009

[6] Bai, Z.D., Miao, B.Q., Rao and Radbakrisbna, C. (1991). Estimation of directions of arrival of signals: Asymptotic results. In *Advances in Spectrum Analysis and Array Processing*, *Vol. I* 327–347.

[7] Bai, Z.D. and Silverstein, J.W. (1999). Exact separation of eigenvalues of large-dimensional sample covariance matrices. *Ann. Probab.* **27** 1536–1555. MR1733159 https://doi.org/10.1214/aop/1022677458

[8] Bai, Z.D. and Silverstein, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. MR2040792 https://doi.org/10.1214/aop/1078415845

[9] Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

[10] Baik, J. and Silverstein, J.W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680 https://doi.org/10.1016/j.jmva.2005.08.003

[11] Ben Arous, G. and Péché, S. (2005). Universality of local eigenvalue statistics for some sample covariance matrices. *Comm. Pure Appl. Math.* **58** 1316–1357. MR2162782 https://doi.org/10.1002/cpa.20070

[12] Cai, T.T., Han, X. and Pan, G.M. (2019). Limiting laws for divergent spiked eigenvalues and largest non-spiked eigenvalue of sample covariance matrices. *Ann. Statist.* To appear. http://arxiv.org/abs/1711.00217v2.

[13] Dyson, F.J. (1970). Correlations between eigenvalues of a random matrix. *Comm. Math. Phys.* **19** 235–250. MR0278668

[14] Erdős, L., Péché, S., Ramírez, J.A., Schlein, B. and Yau, H.-T. (2010). Bulk universality for Wigner matrices. *Comm. Pure Appl. Math.* **63** 895–925. MR2662426 https://doi.org/10.1002/cpa.20317

[15] Erdős, L., Ramírez, J.A., Schlein, B. and Yau, H.-T. (2010). Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation. *Electron. J. Probab.* **15** 526–603. MR2639734 https://doi.org/10.1214/EJP.v15-768

[16] Hoyle, D.C. and Rattray, M. (2004). Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Phys. Rev. E* **69** 026124.

[17] Hu, J. and Bai, Z.D. (2014). Estimation of directions of arrival of signals: Asymptotic results. *Sci. China Math.* **57**. https://doi.org/10.1007/s11425-014-4855-6

[18] Jiang, D. and Bai, Z. (2020). Supplement to "Generalized four moment theorem and an application to CLT for spiked eigenvalues of high-dimensional covariance matrices." https://doi.org/10.3150/20-BEJ1237SUPP

[19] Johansson, K. (2001). Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices. *Comm. Math. Phys.* **215** 683–705. MR1810949 https://doi.org/10.1007/s002200000328

[20] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

[21] Mehta, M.L. (1967). *Random Matrices and the Statistical Theory of Energy Levels*. New York: Academic Press. MR0220494

[22] Onatski, A. (2009). Testing hypotheses about the numbers of factors in large factor models. *Econometrica* **77** 1447–1479. MR2561070 https://doi.org/10.3982/ECTA6964

[23] Passemier, D. and Yao, J.-F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices Theory Appl.* **1** 1150002, 19. MR2930380 https://doi.org/10.1142/S201032631150002X

[24] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865
[25] Skorohod, A.V. (1956). Limit theorems for stochastic processes. *Theory Probab. Appl.* **1** 261–290. MR0084897
[26] Soshnikov, A. (1999). Universality at the edge of the spectrum in Wigner random matrices. *Comm. Math. Phys.* **207** 697–733. MR1727234 https://doi.org/10.1007/s002200050743
[27] Tao, T. and Vu, V. (2015). Random matrices: Universality of local spectral statistics of non-Hermitian matrices. *Ann. Probab.* **43** 782–874. MR3306005 https://doi.org/10.1214/13-AOP876
[28] Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374. MR3662457 https://doi.org/10.1214/16-AOS1487
[29] Wigner, E.P. (1958). On the distribution of the roots of certain symmetric matrices. *Ann. of Math.* (2) **67** 325–327. MR0095527 https://doi.org/10.2307/1970008

Received October 2019 and revised May 2020