

CENTRAL LIMIT THEOREMS FOR EIGENVALUES IN A SPIKED POPULATION MODEL*

BY ZHIDONG BAI AND JIAN-FENG YAO

Northern-East Normal University of China, National University of Singapore and IRMAR/Université de Rennes 1

Abstract In a spiked population model, the population covariance matrix has all its eigenvalues equal to unit except for a few fixed eigenvalues (spikes). This model is proposed by Johnstone to cope with empirical findings on various data sets. The question is to quantify the effect of the perturbation caused by the spike eigenvalues. A recent work by Baik and Silverstein establishes the almost sure limits of the extreme sample eigenvalues associated to the spike eigenvalues when the population and the sample sizes become large. This paper establishes the limiting distributions of these extreme sample eigenvalues. As another important result of the paper, we provide a central limit theorem on random sesquilinear forms.

1. Introduction. It is well-known that the empirical spectral distribution (E.S.D) of a large sample covariance matrix converges to the family of Marčenko-Pastur laws under fairly general condition on the sample variables [8, 1]. On the other hand, the study of the largest or smallest eigenvalues is more complex. In a variety of situations, the almost sure limits of these extreme eigenvalues are proved to coincide with the boundaries of the support of the limiting distribution. As an example, when the sample vectors have independent coordinates and unit variances and assuming that the ratio p/n of the population size p over the sample size n tends to a positive limit $y \in (0, 1)$, then the limiting distribution is the classical Marčenko-Pastur law $F_y(dx)$

$$F_y(dx) = \frac{1}{2\pi xy} \sqrt{(x - a_y)(b_y - x)} dx, \quad a_y \leq x \leq b_y, \quad (1.1)$$

where $a_y = (1 - \sqrt{y})^2$, and $b_y = (1 + \sqrt{y})^2$. Moreover, the smallest and the largest eigenvalue converge almost surely to the boundary a_y and b_y , respectively.

Recent empirical data analysis from fields like wireless communication engineering, speech recognition or gene expression experiments suggest that frequently, some extreme eigenvalues of sample covariance matrices are well-separated from the rest. For instance, see Figures 1 and 2 in Johnstone [7] which display the sample eigenvalues of the functional data consisting of a speech dataset of 162 instances of a phoneme “dcl” spoken by males calculated at 256 points. As a way for possible explanation of this phenomenon, this author proposes a *spiked population model* where all eigenvalues of the population covariance matrix are equal to one except a fixed and relatively small number among them (*spikes*). Clearly, a spiked population model can be considered as a small perturbation of the so-called *null case* where all the eigenvalues of the population covariance matrix are unit. It then raises the question how such

*Research was (partially) completed while J.-F. Yao was visiting the Institute for Mathematical Sciences, National University of Singapore in 2006

AMS 2000 subject classifications: Primary 62H25, 62E20; secondary 60F05, 15A52

Keywords and phrases: Sample covariance matrices, Spiked population model, Central limit theorems, Largest eigenvalue, Extreme eigenvalues, Random sesquilinear forms, Random quadratic forms

a small perturbation affects the limits of the extreme eigenvalues of the sample covariance matrix as compared to the *null case*.

The behavior of the largest eigenvalue in case of complex Gaussian variables has been recently studied in Baik et al. [5]. These authors prove a transition phenomenon: the weak limit as well as the scaling of the largest eigenvalue is different according to the largest spike eigenvalue is larger, equal or less than the critical value $1 + \sqrt{y}$. In Baik and Silverstein [4], the authors consider the spiked population model with general random variables: complex or real and not necessarily Gaussian. For the almost sure limits of the extreme sample eigenvalues, they also find that these limits depend on the critical values $1 + \sqrt{y}$ and $1 - \sqrt{y}$ from above and below, respectively. For example, if there are M eigenvalues in the population covariance matrix larger than $1 + \sqrt{y}$, then the M largest eigenvalues from the sample covariance matrix will have their (almost surely) limits above the right edge b_y of of the limiting Marčenko-Pastur law. Analogous results are also proposed for the case $y > 1$ and $y = 1$.

An important question here is to find the limiting distributions of these extreme eigenvalues. As mentioned above, the results are proposed in [5] for the largest eigenvalue and the Gaussian complex case. In this perspective, assuming that the population vector is real Gaussian with a diagonal covariance matrix and that the M spike eigenvalues are all simple, Paul [10] found that each of the M largest sample eigenvalues has a Gaussian limiting distribution.

In this paper, we follow the general set-up of [4]. Assuming $y \in (0, 1)$ and general population variables, we will establish central limit theorems for the largest as well as for the smallest sample eigenvalues associated to spike eigenvalues outside the interval $[1 - \sqrt{y}, 1 + \sqrt{y}]$. Furthermore, we prove that the limiting distribution of such sample extreme eigenvalues is Gaussian only if the corresponding spike population eigenvalue is simple. Otherwise, if a spiked eigenvalue is multiple, say of index k , then there will be k packed-consecutive sample eigenvalues $\lambda_{n,1}, \dots, \lambda_{n,k}$ which converge jointly to the distribution of a $k \times k$ symmetric (or Hermitian) Gaussian random matrix. Consequently in this case, the limiting distribution of a single $\lambda_{n,j}$ is generally non Gaussian.

The main tools of our analysis are borrowed from the random matrix theory on one hand. For general background of this theory, we refer to the book Mehta [9] and a modern review by Bai [1]. On the other hand, we introduce in this paper another important tool, namely a CLT for random sesquilinear forms which should have its own interests.

The remaining sections of the paper are organized as follows. First in Section 2, we introduce the spiked population model and recall known results on the almost sure limits of extreme sample eigenvalues. A determinant equation involving a random sesquilinear form $K_n(\lambda)$ is also introduced. This equation serves as the starting block of our analysis. Section 3 is devoted to a review of some preliminary results on related random matrices. Next in Section 4, using general CLT's on random sesquilinear forms, we provide a CLT for the key random sesquilinear form $K_n(\lambda)$. Then, in Section 5, we establish the main result of the paper, namely a general CLT for extreme sample eigenvalues. In particular, we recover a CLT given in [10] as a special instance.

2. Spiked population model. We consider a zero-mean, complex-valued random vector $x = (\xi^T, \eta^T)^T$ where $\xi = (\xi(1), \dots, \xi(M))^T$, $\eta = (\eta(1), \dots, \eta(p))^T$ are independent, of dimension M and p respectively. Moreover, we assume that $\mathbb{E}[\|x\|^4] < \infty$ and the coordinates of η are independent and identically distributed with unit variance. The *population*

covariance matrix of the vector x is therefore

$$V = \text{cov}(x) = \begin{pmatrix} \Sigma & 0 \\ 0 & I_p \end{pmatrix}.$$

We consider the following spiked population model by assuming that Σ has K non null and non unit eigenvalues $\alpha_1, \dots, \alpha_K$ with respective multiplicity n_1, \dots, n_K ($n_1 + \dots + n_K = M$). Therefore, the eigenvalues of the population covariance matrix V are unit except the (α_j) , called *spike eigenvalues*.

Let $x_i = (\xi_i^T, \eta_i^T)^T$ be n copies i.i.d. of x . The *sample covariance matrix* is

$$S_n = \frac{1}{n} \sum_{i=1}^n x_i x_i^*$$

which can be rewritten as

$$S_n = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} X_1 X_1^* & X_1 X_2^* \\ X_2 X_1^* & X_2 X_2^* \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum \xi_i \xi_i^* & \sum \xi_i \eta_i^* \\ \sum \eta_i \xi_i^* & \sum \eta_i \eta_i^* \end{pmatrix}, \quad (2.2)$$

with

$$X_1 = \frac{1}{\sqrt{n}} (\xi_1, \dots, \xi_n)_{M \times n} = \frac{1}{\sqrt{n}} \xi_{1:n}, \quad X_2 = \frac{1}{\sqrt{n}} (\eta_1, \dots, \eta_n)_{p \times n} = \frac{1}{\sqrt{n}} \eta_{1:n}.$$

It is assumed in the sequel that M is fixed, and p and n are related so that when $n \rightarrow \infty$, $p/n \rightarrow y \in (0, 1)$. The E.S.D of S_n , as well as the one of S_{22} , converges to the Marčenko-Pastur distribution $F_y(dx)$ given in (1.1). As explained in Introduction, a central question is to quantify the effect caused by the small number of spiked eigenvalues on the asymptotic of the extreme sample eigenvalues.

2.1. Almost sure convergence of the extreme eigenvalues. Assume that among the M eigenvalues of Σ , there are exactly M_b greater than $1 + \sqrt{y}$ and M_a smaller than $1 - \sqrt{y}$:

$$\alpha_1 > \dots > \alpha_{M_b} > 1 + \sqrt{y}, \quad \alpha_M < \dots < \alpha_{M-M_b+1} < 1 - \sqrt{y}, \quad (2.3)$$

and $1 - \sqrt{y} \leq \alpha_k \leq 1 + \sqrt{y}$ for the other α_k 's. Moreover, for $\alpha \neq 1$, we define the function

$$\lambda = \phi(\alpha) = \alpha + \frac{y\alpha}{\alpha - 1}. \quad (2.4)$$

As $y < 1$, we have $p \leq n$ for large n . Let

$$\lambda_{n,1} \geq \lambda_{n,2} \geq \dots \geq \lambda_{n,p}$$

be the eigenvalues of the sample covariance matrix S_n . Let $s_i = n_1 + \dots + n_i$ for $1 \leq i \leq M_b$ and $t_j = n_M + \dots + n_j$ for $1 \leq j \leq M_a$ (by convention $s_0 = t_0 = 0$).

As a first general answer on the effect of spike eigenvalues, Baik and Silverstein [4] completely determines the almost sure limits of the s_{M_b} largest and the t_{M-M_a+1} smallest sample eigenvalues. Namely, for each $k \in \{1, \dots, M_b\}$ and $s_{k-1} < j \leq s_k$ (largest eigenvalues) or $k \in \{1, \dots, M_a\}$ and $p - t_k < j \leq p - t_{k-1}$ (smallest eigenvalues),

$$\lambda_{n,j} \rightarrow \alpha_k + \frac{y\alpha_k}{\alpha_k - 1}, \quad \text{almost surely.} \quad (2.5)$$

In other words, if a spike eigenvalue α_k lies outside the interval $[1 - \sqrt{y}, 1 + \sqrt{y}]$ and has multiplicity n_k , then $\phi(\alpha_k)$ is the limit of n_k packed sample eigenvalue $\{\lambda_{n,j}, j \in J_k\}$. Here we have denoted by J_k the corresponding set of indexes: $J_k = \{s_{k-1}+1, \dots, s_k\}$ for $\alpha_k > 1 + \sqrt{y}$ and $J_k = \{p - t_k + 1, \dots, p - t_{k-1}\}$ for $\alpha_k < 1 - \sqrt{y}$.

We remark that ϕ is symmetrical about the point $(1, 1 + y)$: $\phi(1 + u) + \phi(1 - u) = 2(1 + y)$ for all real $u \neq 0$. Moreover on $(1, \infty)$, ϕ is convex, minimum at $1 + \sqrt{y}$ with $\phi(1 + \sqrt{y}) = b_y$. Furthermore, ϕ is strictly increasing on both intervals $[0, 1 - \sqrt{y}]$ and $[1 + \sqrt{y}, \infty)$, varying from 0 to a_y on one hand, and from b_y to infinity on the other hand.

2.2. A determinant equation. Let us fix a spike eigenvalue $\alpha_k \notin [1 - \sqrt{y}, 1 + \sqrt{y}]$ with multiplicity n_k . The aim of the paper is to derive a CLT for the n_k -packed sample eigenvalues

$$\sqrt{n}[\lambda_{n,j} - \phi(\alpha_k)], \quad j \in J_k.$$

By definition, $\lambda_{n,j}$ solves the equation

$$0 = |\lambda I - S_n| = |\lambda I - S_{22}| |\lambda I - K_n(\lambda)|, \quad (2.6)$$

where

$$K_n(\lambda) = S_{11} + S_{12}(\lambda I - S_{22})^{-1}S_{21}. \quad (2.7)$$

As when $n \rightarrow \infty$, with probability 1, the limit $\lambda_{n,j} \rightarrow \phi(\alpha_k) \notin [a_y, b_y]$ and the eigenvalues of S_n go inside the interval $[a_y, b_y]$, the probability of the event Q_n

$$Q_n = \{\lambda_{n,j} \notin [a_y, b_y]\} \cap \{\text{spectrum of } S_{22} \subset [a_y, b_y]\}$$

tends to 1. Conditional on this event, the $(\lambda_{n,j})$'s then solve the determinant equation

$$|\lambda I - K_n(\lambda)| = 0. \quad (2.8)$$

Therefore without loss of generality, we can assume that $\lambda_{n,j} \notin [a_y, b_y]$ and they are solutions of this equation.

3. Preliminary results and useful lemmas. This section is devoted to review some preliminary results for later use. For $\lambda \notin [a_y, b_y]$, we define

$$m_1(\lambda) = \int \frac{x}{\lambda - x} F_y(dx), \quad (3.1)$$

$$m_2(\lambda) = \int \frac{x^2}{(\lambda - x)^2} F_y(dx), \quad (3.2)$$

$$m_3(\lambda) = \int \frac{x}{(\lambda - x)^2} F_y(dx). \quad (3.3)$$

It is easily seen that

$$\int \frac{\lambda}{\lambda - x} F_y(dx) = 1 + m_1(\lambda), \quad \int \frac{\lambda^2}{(\lambda - x)^2} = 1 + 2m_1(\lambda) + m_2(\lambda).$$

If a real constant $\alpha \notin [1 - \sqrt{y}, 1 + \sqrt{y}]$, then $\phi(\alpha) \notin [a_y, b_y]$ and we have

$$m_1 \circ \phi(\alpha) = \frac{1}{\alpha - 1}, \quad (3.4)$$

$$m_2 \circ \phi(\alpha) = \frac{(\alpha - 1) + y(\alpha + 1)}{(\alpha - 1)[(\alpha - 1)^2 - y]}, \quad (3.5)$$

$$m_3 \circ \phi(\alpha) = \frac{1}{(\alpha - 1)^2 - y}. \quad (3.6)$$

Let us mention that all these formula can be obtained by derivation of the Stieltjes transform of the Marčenko-Pastur law $F_y(dx)$

$$m(z) = \int \frac{1}{x - z} F_y(dx) = \frac{1}{2yz} \{1 - y - z + \sqrt{(y + 1 - z)^2 - 4y}\}, \quad z \notin [a_y, b_y].$$

Here, \sqrt{u} denotes the square root with positive imaginary part for $u \in \mathbb{C}$.

Another important quantity is the random matrix

$$A_n = (a_{ij}) = A_n(\lambda) = X_2^*(\lambda I - X_2 X_2^*)^{-1} X_2, \quad \lambda \notin [a_y, b_y]. \quad (3.7)$$

The following lemma gives the law of large numbers for some useful statistics related to A_n .

Lemma 3.1 *We have*

$$\frac{1}{n} \text{tr} A_n \xrightarrow{P} y m_1(\lambda), \quad (3.8)$$

$$\frac{1}{n} \text{tr} A_n A_n^* \xrightarrow{P} y m_2(\lambda), \quad (3.9)$$

$$\frac{1}{n} \sum_{i=1}^n a_{ii}^2 \xrightarrow{P} \left(\frac{y[1 + m_1(\lambda)]}{\lambda - y[1 + m_1(\lambda)]} \right)^2. \quad (3.10)$$

Proof. Let $\beta_{n,j}$, $j = 1, \dots, p$ be the eigenvalues of $S_{22} = X_2 X_2^*$. The first equality is easy. For the second one, We have

$$\begin{aligned} \frac{1}{n} \text{tr} A_n A_n^* &= \frac{1}{n} \text{tr} (\lambda I - X_2 X_2^*)^{-1} X_2 X_2^* (\lambda I - X_2 X_2^*)^{-1} X_2 X_2^* \\ &= \frac{p}{n} \sum_{j=1}^p \frac{\beta_{n,j}^2}{(\lambda - \beta_{n,j})^2} \\ &\xrightarrow{P} y \int \frac{x^2}{(\lambda - x)^2} F_y(dx). \end{aligned}$$

For (3.10), let $e_i \in \mathbb{C}^n$ be the column vector whose i -th element is 1 and others are 0 and X_{2i} denote the matrix obtained from X_2 by deleting the i -th column of X_2 . We have $X_2 = X_{2i} + \frac{1}{n} \eta_i \eta_i^*$. Therefore,

$$a_{ii} = e_i^* X_2^* (\lambda I - X_2 X_2^*)^{-1} X_2 e_i = \frac{1}{n} \eta_i^* (\lambda I - X_2 X_2^*)^{-1} \eta_i = - \frac{\frac{1}{n} \eta_i^* (X_{2i} X_{2i}^* - \lambda I)^{-1} \eta_i}{1 + \frac{1}{n} \eta_i^* (X_{2i} X_{2i}^* - \lambda I)^{-1} \eta_i}.$$

Using Lemma 2.7 of Bai and Silverstein [3],

$$\mathbb{E} \left| \frac{1}{n} \eta_i^* (X_{2i} X_{2i}^* - \lambda I)^{-1} \eta_i - \frac{1}{n} \text{tr} (X_{2i} X_{2i}^* - \lambda I)^{-1} \right|^2 \leq \frac{K}{n^2} \mathbb{E} |\eta(1)|^4 \mathbb{E} \text{tr} (X_{2i} X_{2i}^* - \lambda I)^{-2}$$

which gives that

$$a_{ii} \xrightarrow{P} -\frac{y \int \frac{1}{x-\lambda} F_y(dx)}{1 + y \int \frac{1}{x-\lambda} F_y(dx)} = \frac{y[1 + m_1(\lambda)]}{\lambda - y[1 + m_1(\lambda)]}. \quad (3.11)$$

Further, it is easy to verify that

$$\lim_{n \rightarrow \infty} \mathbb{E} \frac{\text{tr}(X_2^*(\lambda I - X_2 X_2^*)^{-1} X_2)^4}{n} < \infty$$

which implies, together with inequality 3.3.41 of [6] that

$$\sup_n \mathbb{E} a_{11}^4 = \sup_n \frac{1}{n} \sum_{i=1}^n \mathbb{E} a_{ii}^4 \leq \sup_n \mathbb{E} \frac{\text{tr}(X_2^*(\lambda I - X_2 X_2^*)^{-1} X_2)^4}{n} < \infty.$$

Therefore, the family of the random variables $\{a_{11}^2\}$ indexed by n is uniformly integrable. Combining with (3.11), we get

$$\mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n a_{ii}^2 - \left(\frac{y[1 + m_1(\lambda)]}{\lambda - y[1 + m_1(\lambda)]} \right)^2 \right| \leq \mathbb{E} |a_{11}^2 - \left(\frac{y[1 + m_1(\lambda)]}{\lambda - y[1 + m_1(\lambda)]} \right)^2| \rightarrow 0.$$

Thus (3.10) follows. \blacksquare

4. The random form $K_n(\lambda)$. We now consider in details the random form K_n introduced in (2.7). With $A_n = X_2^*(\lambda I - X_2 X_2^*)^{-1} X_2$, we have

$$\begin{aligned} K_n(\lambda) &= S_{11} + X_1 A_n X_1^* = \frac{1}{n} \xi_{1:n}(I + A_n) \xi_{1:n}^* \\ &= \frac{1}{n} \{ \xi_{1:n}(I + A_n) \xi_{1:n}^* - \Sigma \text{tr}(I + A_n) \} + \frac{1}{n} \Sigma \text{tr}(I + A_n) \\ &= \frac{1}{\sqrt{n}} R_n + [1 + y m_1(\lambda)] \Sigma + o_P\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \quad (4.1)$$

with

$$R_n = R_n(\lambda) = \frac{1}{\sqrt{n}} \{ \xi_{1:n}(I + A_n) \xi_{1:n}^* - \Sigma \text{tr}(I + A_n) \}. \quad (4.2)$$

In the last derivation, we have used the fact

$$\frac{1}{n} \text{tr}(I + A_n) = 1 + y m_1(\lambda) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

which follows from a CLT for $\text{tr}(A_n)$ [see 2].

Our next step is to find the limit distribution of the sequence of random matrices $\{R_n(\lambda)\}$.

4.1. Limiting distribution $R(\lambda)$ of $R_n(\lambda)$: real variables case. In this section, we assume that the variables ξ and η are **real-valued**. We apply CLT's to the $K = \frac{1}{2}M(M+1)$ bilinear forms

$$u(i)(I + A_n)u(j)^T, \quad 1 \leq i \leq j \leq M$$

with

$$u(i) = (\xi_1(i), \dots, \xi_n(i)).$$

More precisely, with $\ell = (i, j)$, we are substituting $u(i)^T$ for $X(\ell)$, and $u(j)^T$ for $Y(\ell)$, respectively.

We have, by Lemma 3.1,

$$\begin{aligned}\theta = \tau &= \lim_n \frac{1}{n} \operatorname{tr}(I + A_n)^2 = 1 + 2ym_1(\lambda) + ym_2(\lambda) , \\ \omega &= \lim_n \frac{1}{n} \sum_{i=1}^n [(I + A_n)_{ii}]^2 = 1 + 2ym_1(\lambda) + \left(\frac{y[1 + m_1(\lambda)]}{\lambda - y[1 + m_1(\lambda)]} \right)^2 .\end{aligned}$$

Therefore, R_n converges weakly to a symmetric random matrix with zero-mean Gaussian variables $R = (R_{ij})$ with the following covariance function, assuming $1 \leq i \leq j \leq M$,

$$\begin{aligned}\operatorname{cov}(R_{ij}, R_{i'j'}) &= \omega \{ \mathbb{E} [\xi(i)\xi(j)\xi(i')\xi(j')] - \Sigma_{ij}\Sigma_{i'j'} \} \\ &\quad + (\theta - \omega) \{ \mathbb{E}[\xi(i)\xi(j')] \mathbb{E}[\xi(i')\xi(j)] \} \\ &\quad + (\theta - \omega) \{ \mathbb{E}[\xi(i)\xi(i')] \mathbb{E}[\xi(j)\xi(j')] \} .\end{aligned}\tag{4.3}$$

In particular, we have the following formula for the variances

$$\operatorname{var}(R_{ij}) = \theta(\Sigma_{ii}\Sigma_{jj} + \Sigma_{ij}^2) + \omega \{ \mathbb{E}[\xi^2(i)\xi^2(j)] - 2\Sigma_{ij}^2 - \Sigma_{ii}\Sigma_{jj} \} .\tag{4.4}$$

In case of a diagonal element R_{ii} , this expression simplifies to

$$\operatorname{var}(R_{ii}) = [2\theta + \beta_i\omega]\Sigma_{ii}^2 , \quad \text{with } \beta_i = \frac{\mathbb{E}[\xi(i)^4]}{\Sigma_{ii}^2} - 3.\tag{4.5}$$

Note that if $\xi(i)$ is Gaussian, $\beta_i = 0$.

Remark. If the coordinates $\{\xi(i)\}$ of ξ are independent, then the limiting covariance matrix in (4.3) is diagonal: the limiting Gaussian matrix is made with independent entries. Their variances simplify to (4.5) and

$$\operatorname{var}(R_{ij}) = \theta\Sigma_{ii}\Sigma_{jj} , \quad i < j.\tag{4.6}$$

4.2. *Limiting distribution $R(\lambda)$ of $R_n(\lambda)$: complex variables case.* Now we assume the general case with complex-valued variables ξ and η . By applying CLT's to the $K = \frac{1}{2}M(M+1)$ sesquilinear forms

$$u(i)(I + A_n)u(j)^* , \quad 1 \leq i \leq j \leq M$$

with

$$u(i) = (\xi_1(i), \dots, \xi_n(i)).$$

More precisely, with $\ell = (i, j)$, we are substituting $u(i)^*$ for $X(\ell)$, and $u(j)^*$ for $Y(\ell)$, respectively.

Again by Lemma 3.1,

$$\begin{aligned}\theta &= \lim_n \frac{1}{n} \operatorname{tr}(I + A_n)^2 = 1 + 2ym_1(\lambda) + ym_2(\lambda) , \\ \omega &= \lim_n \frac{1}{n} \sum_{i=1}^n [(I + A_n)_{ii}]^2 = 1 + 2ym_1(\lambda) + \left(\frac{y[1 + m_1(\lambda)]}{\lambda - y[1 + m_1(\lambda)]} \right)^2 .\end{aligned}$$

Here we need an additional condition which is specific to the complex case. Assume that the following limit exists

$$m_4(\lambda) = \lim_n \frac{1}{n} \operatorname{tr} A_n A_n^T = 0, \quad \lambda \notin [a_y, b_y]. \quad (4.7)$$

Therefore,

$$\tau = \lim_n \frac{1}{n} \operatorname{tr} (I + A_n)(I + A_n)^T = 1 + 2ym_1(\lambda) + m_4(\lambda).$$

Consequently, R_n converges weakly to a zero-mean Hermitian random matrix $R = (R_{ij})$. Moreover, the joint distribution of the real and imaginary parts of the upper-triangular bloc $\{R_{ij}, 1 \leq i \leq j \leq M\}$ is a $2K$ -dimensional Gaussian vector with covariance matrix

$$\Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix} \quad (4.8)$$

where

$$\begin{aligned} \Gamma_{11} &= \frac{1}{4} \sum_{j=1}^3 \{2\Re(B_j) + B_{ja} + B_{jb}\}, \\ \Gamma_{22} &= \frac{1}{4} \sum_{j=1}^3 \{-2\Re(B_j) + B_{ja} + B_{jb}\}, \\ \Gamma_{12} &= \frac{1}{2} \sum_{j=1}^3 \Im(B_j), \end{aligned}$$

and for $1 \leq i \leq j \leq M$ and $1 \leq i' \leq j' \leq M$,

$$\begin{aligned} B_1(ij, i'j') &= \omega (\mathbb{E}[\xi_i \bar{\xi}_j \xi_{i'} \bar{\xi}_{j'}] - \Sigma_{ij} \Sigma_{i'j'}), \\ B_2(ij, i'j') &= (\theta - \omega) \Sigma_{ij'} \Sigma_{i'j}, \\ B_3(ij, i'j') &= (\tau - \omega) (\mathbb{E}[\xi_i \xi_{i'}] \mathbb{E}[\bar{\xi}_j \bar{\xi}_{j'}]), \\ B_{1a}(ij, i'j') &= \omega (\mathbb{E}[|\xi_i \xi_{i'}|^2] - \Sigma_{ii} \Sigma_{i'i'}), \\ B_{1b}(ij, i'j') &= \omega (\mathbb{E}[|\xi_j \xi_{j'}|^2] - \Sigma_{jj} \Sigma_{j'j'}), \\ B_{2a}(ij, i'j') &= (\theta - \omega) |\Sigma_{ii'}|^2, \\ B_{2b}(ij, i'j') &= (\theta - \omega) |\Sigma_{jj'}|^2, \\ B_{3a}(ij, i'j') &= (\tau - \omega) |\mathbb{E}[\xi_i \xi_{i'}]|^2, \\ B_{3b}(ij, i'j') &= (\tau - \omega) |\mathbb{E}[\xi_j \xi_{j'}]|^2. \end{aligned}$$

This covariance matrix Γ has a complicated expression. However, the variance of a diagonal element R_{ii} has a much simpler expression if moreover, $\mathbb{E}[\xi^2(i)] = 0$ for all $1 \leq i \leq M$,

$$\operatorname{var}(R_{ii}) = [\theta + \beta'_i \omega] \Sigma_{ii}^2, \quad \text{with } \beta'_i = \frac{\mathbb{E}[\xi(i)^4]}{\Sigma_{ii}^2} - 2. \quad (4.9)$$

In particular, if $\xi(i)$ is Gaussian, $\beta'_i = 0$.

5. CLT for extreme eigenvalues. We are in order to introduce the main result of the paper. Let the spectral decomposition of Σ ,

$$\Sigma = U \begin{pmatrix} \alpha_1 I_{n_1} & \cdots & 0 \\ 0 & \ddots & 0 \\ \cdots & 0 & \alpha_K I_{n_K} \end{pmatrix} U^* , \quad (5.1)$$

where U is an unitary matrix. Following Section 2.1, for each spiked eigenvalue $\alpha_k \notin [1 - \sqrt{y}, 1 + \sqrt{y}]$, let $\{\lambda_{n,j}, j \in J_k\}$ be the n_k packed eigenvalues of the sample covariance matrix which all tend almost surely to $\lambda_k = \phi(\alpha_k)$. Let $R(\lambda_k)$ be the Gaussian matrix limit of the sequence of matrices of random forms $[R_n(\lambda_k)]_n$ given in Section 4.1 (real variables case) and 4.2 (complex variables case), respectively. Let

$$\tilde{R}(\lambda_k) = U^* R(\lambda_k) U . \quad (5.2)$$

Theorem 5.1 *For each spike eigenvalue $\alpha_k \notin [1 - \sqrt{y}, 1 + \sqrt{y}]$, the n_k -dimensional real vector*

$$\sqrt{n}\{\lambda_{n,j} - \lambda_k, j \in J_k\} ,$$

converges weakly to the distribution of the n_k eigenvalues of the Gaussian random matrix

$$\frac{1}{1 + ym_3(\lambda_k)\alpha_k} \tilde{R}_{kk}(\lambda_k).$$

where $\tilde{R}_{kk}(\lambda_k)$ is the k -th diagonal bloc of $\tilde{R}(\lambda_k)$ corresponding to the indexes $\{u, v \in J_k\}$.

References.

- [1] Z.D. Bai. Methodologies in spectral analysis of large dimensional random matrices. a review. *Statistica Sinica*, 9:611–677, 1999.
- [2] Z.D. Bai and J.W. Silverstein. Clt for linear spectral statistics of large-dimensional sample covariance matrices. *Annals of Probability*, 32(1A):553–605, 2004.
- [3] Z.D. Bai and J.W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large dimensional sample covariance matrices. *Ann. Probab.*, 26:316–345, 1998.
- [4] J. Baik and J.W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. Technical report, North Carolina State University, 2005.
- [5] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.
- [6] R. A. Horn and Ch. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- [7] I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statistics*, 29(2):295–327, 2001.
- [8] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, 1:457–483, 1967.

- [9] M.L. Mehta. *Random Matrices*. Academic Press, New York, 1991.
- [10] Debashis Paul. Asymptotics of the leading sample eigenvalues for a spiked covariance model. Technical report, Stanford University, 2004.

ZHIDONG BAI
DEPARTMENT OF MATHEMATICS
NORTHERN-EAST NORMAL UNIVERSITY
5268 PEOPLE'S ROAD
130024 CHANGCHUN, CHINA

AND
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
10, KENT RIDGE CRESCENT
SINGAPORE 119260
E-MAIL: stabaizd@nus.edu.sg

JIAN-FENG YAO
IRMAR/UNIVERSITÉ DE RENNES 1
CAMPUS DE BEAULIEU
35042 RENNES CEDEX, FRANCE
E-MAIL: jian-feng.yao@univ-rennes1.fr