On sample eigenvalues in a generalized spiked population model

Zhidong Bai* **and Jianfeng Yao†**

Zhidong Bai

KLASMOE, School of Mathematics and Statistics

Northeast Normal University

130024 Changchun, China

e-mail: baizd@nenu.edu.cn

Jianfeng Yao

Department of Statistics and Acturaial Science

The University of Hong Kong

Pokfulam, Hong Kong

e-mail: jeffyao@hku.hk

Abstract: In the spiked population model introduced by Johnstone [11], the population covariance matrix has all its eigenvalues equal to unit except for a few fixed eigenvalues (spikes). The question is to quantify the effect of the perturbation caused by the spike eigenvalues. Baik and Silverstein [5] establishes the almost sure limits of the extreme sample eigenvalues associated to the spike eigenvalues when the population and the sample sizes become large. In a recent work [4], we have provided the limiting distributions for these extreme sample eigenvalues. In this paper, we extend this theory to a *generalized* spiked population model where the base population covariance matrix is arbitrary, instead of the identity matrix as in Johnstone's case. As the limiting spectral distribution is here arbitrary, new mathematical tools, different from those in Baik and Silverstein [5], are introduced for establishing the almost sure convergence of the sample eigenvalues generated by the spikes.

*This author'research is partly supported by a Chinese NSF grant (10871036).

†This author'research is supported by a Start-up Research Fund (2010) from The University of Hong Kong.

AMS 2000 subject classifications: Primary 62H05; secondary 15A52, 60F15.

Keywords and phrases: Sample covariance matrices, Spiked population model, Central limit theorems, Largest eigenvalue, Extreme eigenvalues.

1. Introduction

Let (T_p) be a sequence of $p \times p$ non-random and nonnegative definite Hermitian matrices and let (w_{ij}), $i, j \geq 1$ be a doubly infinite array of i.i.d. complex-valued random variables satisfying

$$\mathbb{E}(w_{11}) = 0, \quad \mathbb{E}(|w_{11}|^2) = 1, \quad \mathbb{E}(|w_{11}|^4) < \infty.$$

Write $Z_n = (w_{ij})_{1 \leq i \leq p, 1 \leq j \leq n}$, the upper-left $p \times n$ block, where $p = p(n)$ is related to n such that when $n \to \infty$, $p/n \to y > 0$. Then the matrix $S_n = \frac{1}{n} T_p^{1/2} Z_n Z_n^* T_p^{1/2}$ can be considered as the sample covariance matrix of an i.i.d. sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of p-dimensional observation vectors $\mathbf{x}_j = T_p^{1/2} \mathbf{u}_j$ where $\mathbf{u}_j = (w_{ij})_{1 \leq i \leq p}$ denotes the j-th column of Z_n. Throughout the paper, $A^{1/2}$ stands for any Hermitian square root of an nonnegative definite (n.n.d.) Hermitian matrix A.

Assume that the empirical spectral distribution (ESD) of T_p converges weakly to a nonrandom probability distribution H on $[0, \infty)$. It is then well-known that the ESD of S_n converges to a nonrandom limiting spectral distribution (LSD) G [12, 16].

Let $\lambda_{n,1} \geq \cdots \geq \lambda_{n,p}$ be the set of sample eigenvalues, i.e. the eigenvalues of the sample covariance matrix S_n. The so-called *null case* corresponds to the situation $T_p \equiv I_p$, so that, assuming $y \leq 1$, the LSD G reduces to the Marčenko-Pastur law with support $\Gamma_G = [a_y, b_y]$ where $a_y = (1 - \sqrt{y})^2$ and $b_y = (1 + \sqrt{y})^2$. Furthermore, the extreme sample eigenvalues $\lambda_{n,1}$ and $\lambda_{n,p}$ almost surely tend to b_y and a_y, respectively, and the sample

eigenvalues ($\lambda_{n,j}$) fill completely the interval $[a_y, b_y]$. However, as pointed out by Johnstone [11], many empirical data sets demonstrate a significant deviation from this null case whereby some of the extreme sample eigenvalues are well separated from an inner bulk interval. As a possible explanation for this phenomenon, Johnstone proposes a *spiked population model* where all eigenvalues of T_p are unit except a fixed small number of them (the spikes). In other words, the population eigenvalues $\{\beta_{n,j}\}$ of T_p are

$$\underbrace{\alpha_1, \ldots, \alpha_1}_{n_1}, \ldots, \underbrace{\alpha_K, \ldots, \alpha_K}_{n_K}, \underbrace{1, \ldots, 1}_{p-M},$$

where M and the multiplicity numbers (n_k) are fixed and satisfy $n_1 + \cdots + n_K = M$. Clearly, this spiked population model can be viewed as a finite-rank perturbation of the null case.

Obviously, the global LSD G of S_n is not affected by this small perturbation and still converges to the Marčenko-Pastur law. However, the asymptotic behavior of the extreme eigenvalues of S_n is significantly different from the null case. The fluctuation of the largest eigenvalue $\lambda_{n,1}$ in the case of complex Gaussian variables has been recently studied in Baik et al. [6]. These authors prove a transition phenomenon: the weak limit and the scaling of $\lambda_{n,1}$ are different according to its location with respect to a critical value $1 + \sqrt{y}$. In Baik and Silverstein [5], 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 $1 + \sqrt{y}$ for largest sample eigenvalues, and on $1 - \sqrt{y}$ for smallest ones. For example, if there are m eigenvalues in the population covariance matrix larger than $1 + \sqrt{y}$, then the m largest sample eigenvalues $\lambda_{n,1}, \ldots, \lambda_{n,m}$ will converge to a limit above the right edge b_y of the limiting Marčenko-Pastur law, see §4.1 for more details. In a recent work Bai and Yao [4], considering general

matrix entries as in [5], we have established central limit theorems for these extreme sample eigenvalues generated by spike eigenvalues which are outside the critical interval $[1-\sqrt{y}, 1+\sqrt{y}]$. Note that futher related results on these extreme sample eigenvalues are found in Paul [14] and Onatski [13].

The spiked population model has also an extension to other random matrices ensembles through the general concept of small-rank perturbations. The goal is again to examine the effect caused on the sample extreme eigenvalues by such perturbations. In a series of recent papers [15, 10, 9], these authors establish several results in this vein for ensembles of form $M_n = W_n + n^{-1/2}V$ where W_n is a standard Wigner matrix and V a small-rank matrix.

The present work is motivated by a generalization of Johnstone's spike population model defined as follows. The population covariance matrix T_p possesses two sets of eigenvalues: a small number of them, say (α_k), called *generalized spikes*, are well separated - in a sense to be defined later-, from a base set $(\beta_{n,i})$. In other words, the spectrum of T_p reads as

$$\underbrace{\alpha_1, \ldots, \alpha_1}_{n_1}, \ldots, \underbrace{\alpha_K, \ldots, \alpha_K}_{n_K}, \beta_{n,1}, \ldots, \beta_{n,p-M}.$$

Therefore, this scheme can be viewed as a finite-rank perturbation of a general population covariance matrix with eigenvalues $\{\beta_{n,j}\}$. Note that here the eigenvalues α_k's are not necessarily larger than the $\beta_{n,j}$'s and their exact relationship will be defined in Section 2.

The empirical distributions generated by the eigenvalues $(\beta_{n,i})$ will be assumed to have a limit distribution H. Note that H is also the LSD of T_p since the perturbation is of finite rank. Analogous to Johnstone's spiked population model, the LSD G of the sample covariance matrix S_n is still not affected by the spikes. The aim of this work is to identify the effect caused by the spikes (α_k) on a particular subset of sample eigenvalues.

As demonstrated in Baik and Silverstein [5] for Johnston's model, only

a particular subset of the spikes $\{\alpha_k\}$ will generate some sample eigenvalues which will converge to some limiting points outside the support of G. However in the current generalized scheme, because this LSD G can have an arbitrary form, the characterization of these particular spikes need new mathematical tools than those previously introduced in [5]. This paper provide such new tools which are very different from the ones in [5]. In particular, we provide a complete characterization of those particular spikes according to the sign of the derivatives $\{\psi'(\alpha_k)\}$ where ψ is a fundamental function introduced in §3 (though closely related to the Stieltjes transform of G).

Let us mention that after the completion of this paper, we become aware of two recent, unpublished and closely-related works [7] and [8]. These authors consider more general perturbation models including additive and multiplicative ones and there provide important results on point-wisely convergence of extreme eigenvalues [7] as well as on their fluctuations [8]. It is particularly remarked that several asymptotic results on the associated eigenvectors are also established in [7]. However while in the present paper the deformation considered can be viewed as of multiplicative type only, our methods are completely different; moreover the distributions of the matrix entries are more general as they are not required to obey a orthogonal or unitary invariance as in [7] or a log-Sobolev inequality as in [8].

The remaining sections of the paper are organized as following. §2 gives the precise definition of the generalized spiked population model. Next, we use §3 to recall several useful results on the convergence of the ESD from general sample covariance matrices. In §4, we examine the strong point limits of sample eigenvalues associated to spikes. We then introduce a CLT for these sample eigenvalues in §5 using the methodology developed in [4].

2. Generalized spiked population model

In a generalized spiked population model, the population covariance matrix T_p takes the form

$$T_p = \begin{pmatrix} \Sigma & 0 \\ 0 & V_p \end{pmatrix},$$

where Σ and V_p are nonnegative and nonrandom Hermitian matrices of dimension $M \times M$ and $p' \times p'$, respectively, where $p' = p - M$. The submatrix Σ has K eigenvalues $\alpha_1 > \cdots > \alpha_K > 0$ of respective multiplicity (n_k), and V_p has p' eigenvalues $\beta_{n,1} \geq \cdots \geq \beta_{n,p'}$.

Throughout the paper, we assume that the following assumptions hold.

(a) w_{ij}, $i, j = 1, 2, ...$ are i.i.d. complex random variables with $Ew_{11} = 0$, $E|w_{11}|^2 = 1$, and $E|w_{11}|^4 < \infty$.

(b) $n = n(p)$ with $y_n = p/n \to y > 0$ as $n \to \infty$.

(c) The sequence of ESD H_n of (T_p), i.e. generated by the population eigenvalues $\{\alpha_k, \beta_{n,j}\}$, weakly converges to a probability distribution H as $n \to \infty$.

(d) The sequence $(\|T_p\|)$ of spectral norms of (T_p) is bounded.

For any measure μ on $\mathbb{R}$, we denote by Γ_μ the support of μ, a close set.

Definition 2.1. *An eigenvalue α of the matrix Σ is called a* generalized spike eigenvalue *if $\alpha \notin \Gamma_H$.*

To avoid confusion between spikes and non-spike eigenvalues, we further assume that

(e) $\max\limits_{1 \leq j \leq p'} d(\beta_{nj}, \Gamma_H) = \varepsilon_n \to 0,$

where $d(x, A)$ denotes the distance of a point x to a set A. Note that there is a positive constant δ such that $d(\alpha_k, \Gamma_H) > \delta$, for all $k \leq K$.

The above definition for generalized spikes is consistent with Johnstone's original one of (ordinary) spikes, since in that case we have $H_n \equiv H = \delta_{\{1\}}$ and $\alpha \notin \Gamma_H$ simply means $\alpha \neq 1$. Throughout the paper and for any Hermitian matrix A, we order its eigenvalues in a descending order as $\lambda_1^A \geq \lambda_2^A \geq \cdots$.

3. Known results on the spectrum of large sample covariance matrices

3.1. Marčenko-Pastur distributions

In this section y is an arbitrary positive constant and H an arbitrary probability measure on $\mathbb{R}^+$. Define on the set

$$\mathbb{C}^+ := \{ z \in \mathbb{C} \ : \ \Im(z) > 0 \ \} \ ,$$

the map

$$g(s) = g_{y,H}(s) = -\frac{1}{s} + y \int \frac{t}{1+ts} dH(t) \ , \quad s \in \mathbb{C}^+. \tag{3.1}$$

It is well-known ([3, Chap. 5]) that g is a one-to-one map from $\mathbb{C}^+$ onto itself, and the inverse map $m_{y,H} = g_{y,H}^{-1}$ corresponds to the Stieltjes transform of a probability measure $F_{y,H}$ on $[0, \infty)$. Throughout the paper and with a small abuse of language, we refer $F_{y,H}$ as the Marčenko-Pastur (M.P.) distribution with indexes (y, H).

This family of distributions arises naturally as follows. Consider a companion matrix $\underline{S}_n = \frac{1}{n} Z_n^* T_p Z_n$ of the sample covariance matrix S_n. The spectra of S_n and $\underline{S}_n$ are identical except $|n-p|$ zeros. It is then well-known ([12],[3, Chap. 5]) that under Conditions (a)-(d), the ESD of $\underline{S}_n$ converges to the M.P. distribution $F_{y,H}$. The terminology is slightly ambiguous since the classical M.P. distribution refers to the limit of the ESD of S_n when $T_p = I_p$.

Note that we shall always extend a function h defined on $\mathbb{C}^+$ to the real axis $\mathbb{R}$ by taking the limits $\lim_{\varepsilon\to 0_+} h(x+i\varepsilon)$ for real x's whenever these limits exist. For $\alpha \notin \Gamma_H$ and $\alpha \neq 0$ define

$$\psi(\alpha) = \psi_{y,H}(\alpha) := g(-1/\alpha) = \alpha + y\alpha \int \frac{t}{\alpha - t} dH(t) \ . \tag{3.2}$$

Note that this formula could be extended to $\alpha = 0$ when $0 \notin \Gamma_H$. However, there is no much meaning for $\alpha = 0$ since, as we will see below, the values for α are related to the values of type $-1/s(z)$ where s is some Stieltjes transform and $z \in \mathbb{C}^+$. Therefore, the point 0 will always be excluded from the domain of definition of ψ.

Analytical properties of $F_{y,H}$ can be derived from the fundamental equation (3.2). The following lemma, due to Silverstein and Choi [17], characterizes the close relationship between the supports of the generating measure H and the generated M.P. distribution $F_{y,H}$.

Lemma 3.1. *If $\lambda \notin \Gamma_{F_{y,H}}$, then $m_{y,H}(\lambda) \neq 0$ and $\alpha = -1/m_{y,H}(\lambda)$ satisfies*

i. $\alpha \notin \Gamma_H$ and $\alpha \neq 0$ (so that $\psi(\alpha)$ is well-defined);

ii. $\psi'(\alpha) > 0$.

Conversely, if α satisfies (i)-(ii)*, then $\lambda = \psi(\alpha) \notin \Gamma_{F_{y,H}}$.*

It is then possible to determine the support of $F_{y,H}$ by looking at intervals where $\psi' > 0$. As an example, Figure 1 displays the function ψ for the M.P. distribution with indexes $y = 0.3$ and H the uniform distribution on the set $\{1, 4, 10\}$. The function ψ is strictly increasing on the following intervals: $(-\infty,\ 0)$, $(0,\ 0.63)$, $(1.40,\ 2.57)$ and $(13.19,\ \infty)$. According to Lemma 3.1, we get

$$\Gamma^c_{F_{y,H}} \cap \mathbb{R}^* = (0,\ 0.32) \cup (1.37,\ 1.67) \cup (18.00,\ \infty).$$

Hence, taking into account that 0 belongs to the support of $F_{y,H}$, we have

$$\Gamma_{F_{y,H}} = \{0\} \cup [0.32,\ 1.37] \cup [1.67,\ 18.00].$$

We refer to Bai and Silverstein [2] for a complete account of analytical properties of the family of M.P. distributions $\{F_{y,H}\}$ and the maps $\{\psi_{y,H}\}$. In particular, the following conclusions will be useful:

- when restricted to $\Gamma^c_{F_{y,H}}$, $\psi_{y,H}$ has a well-defined inverse function $\psi^{-1}_{y,H}$: $\Gamma^c_{F_{y,H}} \to \Gamma^c_H$ which is strictly increasing on each interval included into $\Gamma^c_{F_{y,H}}$;
- the function $\psi_{y,H}$ tends to the identity function as $y \to 0$.

3.2. Exact separation of sample eigenvalues

We need first quote two results of Bai and Silverstein [1, 2] on exact separation of sample eigenvalues. Recall the ESD's (H_n) of (T_p), $y_n = p/n$, and let $\{F_{y_n,H_n}\}$ be the sequence of associated M.P. distributions. One should not confuse the M.P. distribution $\{F_{y_n,H_n}\}$ with the ESD of $\underline{S}_n$ although both converge to the M.P. distribution $F_{y,H}$ as $n \to \infty$.

Proposition 3.1. *Assume hold Conditions (a)-(d) and the following*

(f) The interval $[a,b]$ with $a > 0$ lies in an open interval (c,d) outside the support of F_{y_n,H_n} for all large n.

Then

$$P(\text{ no eigenvalue of } S_n \text{ appears in } [a,b] \text{ for all large } n\ ) = 1.$$

Roughly speaking, Proposition 3.1 states that a gap in the spectra of the F_{y_n,H_n}'s is also a gap in the spectrum of S_n for large n. Moreover, under Condition (f), we know by Lemma 3.1, that for large n,

$$\psi^{-1}_{y_n,H_n}\{[a,b]\} \subset \psi^{-1}_{y_n,H_n}\{(c,d)\} \subset \Gamma^c_{H_n}\ .$$

By continuity of F_{y_n,H_n} in its indexes, it follows that we have for large n[1]

$$\psi^{-1}\{[a,b]\} = \psi^{-1}_{y,H}\{[a,b]\} \subset \Gamma^c_{H_n}\ .$$

[1]To see this let us choose a', b' such that $c < a' < a < b < b' < d$. We have $\psi_n^{-1}(a') <$

In other words, it holds almost surely for large n that, $\psi^{-1}\{[a,b]\}$ contains no eigenvalue of T_p. Let for these n, the integer $i_n \geq 0$ be such that

$$T_p \text{ has exactly } i_n \text{ eigenvalues larger than } \psi^{-1}(b)\ . \tag{3.3}$$

Proposition 3.2. *Assume Conditions (a)-(d) and (f) hold. If* $y[1-H(0)] \leq 1$*, or* $y[1-H(0)] > 1$ *but* $[a,b]$ *is not contained in* $[0,x_0]$ *where* $x_0 > 0$ *is the smallest value of the support of* $F_{y,H}$*, then with* i_n *defined in (3.3) we have*

$$P(\lambda_{i_n+1}^{S_n} \leq a < b \leq \lambda_{i_n}^{S_n} \quad \textit{for all large } n) = 1.$$

In other words, under these conditions, it happens eventually that the numbers of sample eigenvalues $\{\lambda_i^{S_n}\}$ in both sides of $[a,b]$ match exactly the numbers of populations eigenvalues $\{\alpha_k, \beta_{n,j}\}$ in both sides of the interval $\psi^{-1}\{[a,b]\}$.

4. Almost sure convergence of sample eigenvalues from generalized spikes

From (3.2) we have

$$\psi'(\alpha) = 1 - y\int \frac{t^2}{(\alpha-t)^2} dH(t)\ , \qquad \psi'''(\alpha) = -6y\int \frac{t^2}{(\alpha-t)^4} dH(t)\ .$$

Therefore, ψ' is concave on any interval outside Γ_H. Moreover for a discrete distribution H, $\psi'(\alpha)$ tends to $-\infty$ when α approaches the point masses of H, see also Figure 1.

As we will see, the asymptotic behavior of the sample eigenvalues generated by a generalized spike eigenvalue α depends on the sign of $\psi'(\alpha)$.

$\psi_n^{-1}(a) < \psi_n^{-1}(b) < \psi_n^{-1}(b')$ and then $\psi^{-1}(a') < \psi^{-1}(a) < \psi^{-1}(b) < \psi^{-1}(b')$ in the limits where the strict inequalities follows the fact that ψ is strictly increasing on $[a',b']$. This implies that when n is large, $\psi_n^{-1}(a') < \psi^{-1}(a) < \psi^{-1}(b) < \psi_n^{-1}(b')$ and thus $\psi^{-1}([a,b]) \subset \psi_n^{-1}([a',b']) \subset \Gamma_{H_n}^c$.

Definition 4.1. *We call a generalized spike eigenvalue α, a* distant spike *for the M.P. law $F_{y,H}$ if $\psi'(\alpha) > 0$, and a* close spike *if $\psi'(\alpha) \leq 0$.*

Recall that ψ depend on the parameters (y, H). When H is fixed, and since by (3.2), ψ tends to the identity function as $y \to 0$, a close spike for a given M.P. law $F_{y,H}$ becomes a distant spike for M.P. law $F_{y',H}$ for small enough y'.

As an example, different types of spikes are displayed in Figure 2. The solid curve corresponds to a zoomed view of $\psi_{0.3,H}$ of Figure 1. For $F_{0.3,H}$, the three values α_1, α_2 and α_5 are close spikes; each small enough α (close to zero), or large enough α (not displayed), or a value between u and v (see the figure) is a distant spike. Furthermore, as y decreases from 0.3 to 0.02 (dashed curve), α_1, α_2 and α_5 become all distant spikes.

Throughout this section, for each spike eigenvalue α_k, we denote by $\nu_k + 1, \ldots, \nu_k + n_k$ the descending ranks of α_k among the eigenvalues of T_p (multiplicities of eigenvalues are counted): in other words, there are ν_k eigenvalues of T_p larger than α_k and $p - \nu_k - n_k$ less.

Theorem 4.1. *Assume that the conditions (a)-(e) hold. Let α_k be a generalized spike eigenvalue of multiplicity n_k satisfying $\psi'(\alpha_k) > 0$ (distant spike) with descending ranks $\nu_k + 1, \ldots, \nu_k + n_k$. Then, the n_k consecutive sample eigenvalues $\{\lambda_i^{S_n}\}$, $i = \nu_k + 1, \ldots, \nu_k + n_k$ converge almost surely to $\psi(\alpha_k)$.*

Proof. By definition we have for $\alpha \notin \{\alpha_k, k = 1, \ldots, K;\ \beta_{n,j}, j = 1, \ldots, p'\}$,

$$\psi_n(\alpha) := \psi_{y_n,H_n}(\alpha) = \alpha + y_n\alpha\left[\frac{p'}{p}\int \frac{t}{\alpha - t} dH_n^v(t) \ + \frac{1}{p}\sum_{j=1}^{K} \frac{n_j\alpha_j}{\alpha - \alpha_j}\right], \quad (4.1)$$

where $H_n^v = \frac{1}{p'}\sum_j \delta_{\beta_{n,j}}$ is the ESD of V_p. Its derivative equals

$$\psi_n'(\alpha) = \psi_{y_n,H_n}'(\alpha) = 1 - y_n\left[\frac{p'}{p}\int \frac{t^2}{(\alpha-t)^2} dH_n^v(t) \ + \frac{1}{p}\sum_{j=1}^{K} \frac{n_j\alpha_j^2}{(\alpha-\alpha_j)^2}\right]. \tag{4.2}$$

Since $\psi'(\alpha_k) > 0$ and by continuity, we can always find $d > c > b > a > \alpha_k$ such that $\psi' > 0$ on $[\alpha_k, d]$. Next by condition (e), the eigenvalues $\beta_{n,j}$'s approach the support Γ_H which is at a positive distance from the spike eigenvalues α_ℓ's. It follows that we can choose the above $d > c > b > a$ such that i) $d < \alpha_{k-1}$ (with the convention $\alpha_0 = \infty$); ii) for n large enough, none of the $\beta_{n,j}$'s will appear in the interval $[\alpha_k, d]$.

Next we claim that on $[a, d]$, $(\psi_n)_n$ and $(\psi_n')_n$ converge uniformly to ψ and ψ', respectively. It follows that we have for all n large enough, ψ_n' is positive on $[a, d]$ (with eventually smaller a, b, c, d), and the interval $(\psi(a), \psi(d))$ will be out of the support of F_{y_n,H_n}. Consequently, the interval $[\psi(b), \psi(c)]$ satisfies the conditions of Proposition 3.2 with $i_n = \nu_k$. Therefore, by Proposition 3.2, we have

$$\begin{cases} P(\lambda_{\nu_k+1}^{S_n} \le \psi(b) < \psi(c) \le \lambda_{\nu_k}^{S_n}, \ \text{ for all large } n) = 1 & \text{if } \nu_k > 0; \\ P(\lambda_{\nu_k+1}^{S_n} \le \psi(b), \ \text{ for all large } n) = 1 & \text{otherwise.} \end{cases}$$

Therefore, it holds almost surely

$$\limsup_n \lambda_{\nu_k+1}^{S_n} \le \psi(b),$$

and finally, letting $b \to \alpha_k$,

$$\limsup_n \lambda_{\nu_k+1}^{S_n} \le \psi(\alpha_k). \tag{4.3}$$

Similarly, one can prove that for $e < f < \alpha_k$ sufficiently close to α_k,

$$\begin{cases} P(\lambda_{\nu_k+n_k+1}^{S_n} \le \psi(e) < \psi(f) \le \lambda_{\nu_k+n_k}^{S_n}, \ \text{ for all large } n) = 1 & \text{if } \nu_k + n_k < p, \\ P(\lambda_{\nu_k+n_k}^{S_n} \ge \psi(f), \ \text{ for all large } n) = 1 & \text{otherwise.} \end{cases}$$

Letting $f \to \alpha_k$, we have

$$\liminf_n \lambda_{\nu_k+n_k}^{S_n} \geq \psi(\alpha_k). \tag{4.4}$$

Thus, we proved that almost surely,

$$\lim_n \lambda_{\nu_k+j}^{S_n} = \psi(\alpha_k), \text{ for } j = 1, \cdots, n_k.$$

The proof of Theorem 4.1 will be complete if we prove the above claim for uniform convergence of $(\psi_n)_n$ and $(\psi_n')_n$ on $[a, d]$. For $(\psi_n)_n$ we have

$$\begin{aligned}\psi_n(\alpha) - \psi(\alpha) &= y\alpha \int \frac{t}{\alpha - t} dH_n^v(t) - y\alpha \int \frac{t}{\alpha - t} dH(t) \\ &\quad + \left(y_n \frac{p'}{p} - y \right) \alpha \int \frac{t}{\alpha - t} dH_n^v(t) \\ &\quad + y_n \alpha \frac{1}{p} \sum_{j=1}^{K} \frac{n_j \alpha_j}{\alpha - \alpha_j} \,. \end{aligned} \tag{4.5}$$

First observe that on $[a, d]$

$$\inf_{1 \leq j \leq K, \alpha \in [a,d]} |\alpha - \alpha_j| > 0,$$

so that it is readily seen that the second and the third term in the r.h.s of (4.5) above converge uniformly to 0.

For the first term, let split the measure H_n^v into two parts $H_{n,1}^v$ and $H_{n,2}^v$ according to whether the $\beta_{n,j}$'s are on the left side or the right side of the interval $[a, d]$. For each of these sub-measures, by similar arguments as above, the integrals

$$\alpha \int \frac{t}{\alpha - t} dH_{n,j}^v(t), \quad j = 1, 2$$

converge pointwisely to

$$\alpha \int \frac{t}{\alpha - t} \mathbb{1}_{\{t<a\}} dH(t) \quad \text{and} \quad \alpha \int \frac{t}{\alpha - t} \mathbb{1}_{\{t>d\}} dH(t),$$

respectively. Note that $\mathbb{1}_{\{t<a\}} dH(t) + \mathbb{1}_{\{t>d\}} dH(t) = dH(t)$. Moreover, the functions

$$\alpha \mapsto \alpha \int \frac{t}{\alpha - t} dH_{n,j}^v(t), \quad j = 1, 2$$

are monotonic and continuous. By Dini's theorem, the above pointwise convergence is also uniform on $[a, d]$. This proves the uniform convergence of $(\psi_n)_n$ and the proof for $(\psi_n')_n$ is similar and thus omitted. The proof of Theorem 4.1 is complete. □

Next we consider close spikes.

Theorem 4.2. *Assume that the conditions (a)-(e) hold. Let α_k be a generalized spike eigenvalue of multiplicity n_k satisfying $\psi'(\alpha_k) \leq 0$ (close spike) with descending ranks $\nu_k + 1, \ldots, \nu_k + n_k$. Let I be the maximal interval in Γ_H^c containing α_k.*

i. If I has a sub-interval (u_k, v_k) on which $\psi' > 0$ (then we take this interval to be maximal), then the n_k sample eigenvalues $\{\lambda_j^{S_n}\}$, $j = \nu_k + 1, \ldots, \nu_k + n_k$ converge almost surely to the number $\psi(w)$ where w is one of the endpoints $\{u_k, v_k\}$ nearest to α_k ;

ii. If for all $\alpha \in I$, $\psi'(\alpha) \leq 0$, then the n_k sample eigenvalues $\{\lambda_j^{S_n}\}$, $j = \nu_k + 1, \ldots, \nu_k + n_k$ converge almost surely to the γ-th quantile of G, the LSD of S_n, where $\gamma = H(0, \alpha_k)$.

Proof. The proof refers to the drawing on the bottom of Figure 3.

(i). Suppose α_k is a spike eigenvalue satisfying $\psi'(\alpha_k) \leq 0$ and there is an interval $(u_k, v_k) \subset I$ on which $\psi' > 0$. Without loss of generality, we can assume $\alpha_k \leq u_k$, the argument of the other situation where $\alpha_k > v_k$ being similar. According to Lemma 3.1, $\psi\{(u_k, v_k)\} \subset \Gamma_{F_{y,H}}^c$ and we claim that $\psi(u_k)$ is a boundary point of the support of G (LSD of S_n). To see this, first we observe that u_k is finite and $\psi'(u_k) \leq 0$ (possibly $-\infty$) by continuity and the maximality of the interval (u_k, v_k). Thus $\psi(u_k) \in \Gamma_G$. Moreover, it is necessarily on the boundary of Γ_G, for otherwise we could find an $e > 0$ such that $(\psi(u_k), \psi(u_k + e))$ is in Γ_G and this would imply that $\psi' \leq 0$ on

the interval $(u_k, u_k + e)$ which is clearly impossible.

Choose $u_k < a < b < \tilde{v}$ ($\tilde{v} = \min(v_k, \alpha_{k-1})$ or v_k in accordance with $k > 1$ or not) such that $(a, b) \subset I$, by the argument used in the proof of Theorem 4.1, one can prove that

$$\begin{cases} P(\lambda^{S_n}_{\nu_k+1} \le \psi(a) < \psi(b) \le \lambda^{S_n}_{\nu_k}, \ \text{ for all large } n) = 1 & \text{if } \nu_k > 0; \\ P(\lambda^{S_n}_{\nu_k+1} \le \psi(a), \ \text{ for all large } n) = 1 & \text{otherwise.} \end{cases}$$

This proves that almost surely,

$$\limsup \lambda^{S_n}_{\nu_k+1} \le \psi(u_k) \le \liminf \lambda^{S_n}_{\nu_k}.$$

On the other hand, since $\psi(u_k)$ is a boundary point of the support of G, we know that for any $\varepsilon > 0$, almost surely, the number of $\lambda_i^{S_n}$'s falling into $[\psi(u_k) - \varepsilon, \psi(u_k)]$ tends to infinity **since the LSD has a positive density function on this interval. In particular, almost surely this interval contains $\lambda^{S_n}_{\nu_k+n_k+1}$ for large n.** Therefore,

$$\liminf \lambda^{S_n}_{\nu_k+n_k+1} \ge \psi(u_k) - \varepsilon, \quad \text{a.s..}$$

Since ε is arbitrary, we have finally proved that almost surely,

$$\lim \lambda^{S_n}_{\nu_k+j} = \psi(u_k), \quad j = 1, \cdots, n_k.$$

Thus, the proof of Conclusion (i) of Theorem 4.2 is complete.

Similarly, if the spiked eigenvalue α_k is like α_2, we can show that the n_k corresponding eigenvalues of S_n goes to $\psi(v_k)$.

(ii) If the spiked eigenvalues is like α_5, where the gap of support of LSD disappeared, clearly the corresponding sample eigenvalues $\lambda_{\nu_k+1}, \ldots, \lambda_{\nu_k+n_k}$ tend to the γ-th quantile of the LSD of S_n where

$$\gamma = 1 - \lim \frac{i_n}{\nu_k} = H(0, \alpha_k).$$

□

4.1. Case of Johnstone's spiked population model

In the case of Johnstone's model, H reduces to the Dirac mass δ_1 and the LSD G equals the Marčenko-Pastur law with $\Gamma_G = [a_y, b_y]$. Each $\alpha > 0$, $\alpha \neq 1$ is then a spike eigenvalue. The associated function ψ in (3.2) becomes

$$\psi(\alpha_k) = \alpha_k + \frac{y\alpha_k}{\alpha_k - 1}. \tag{4.6}$$

The function ψ has the following properties, see Figure 4:

- its range equals $(-\infty, a_y] \cup [b_y, \infty)$;
- $\psi(1 - \sqrt{y}) = a_y$, $\psi(1 + \sqrt{y}) = b_y$;
- $\psi'(\alpha) > 0 \Leftrightarrow |\alpha - 1| > \sqrt{y}$.

Therefore, by Theorem 4.1, for any spike eigenvalue satisfying $\alpha_k > 1 + \sqrt{y}$ (large enough) or $\alpha_k < 1 - \sqrt{y}$ (small enough), there is a packet of n_k consecutive eigenvalues $\{\lambda_{n,j}\}$ converging almost surely to $\psi(\alpha_k) \notin [a_y, b_y]$. In other words, assume there are exactly K_1 spikes greater than $1 + \sqrt{y}$ and K_2 spikes smaller than $1 - \sqrt{y}$. By Theorems 4.1 and 4.2 we conclude that

i. the $N_1 := n_1 + \ldots + n_{K_1}$ largest eigenvalues $\{\lambda_j^{S_n}\}$, $j = 1, \ldots, N_1$ tend to their respective limits $\{\psi(\alpha_k)\}$, $k = 1, \ldots, K_1$;

ii. the immediately following largest eigenvalue $\lambda_{N_1+1}^{S_n}$ tends to the right edge b_y;

iii. the $N_2 := n_K + \cdots + n_{K-K_2+1}$ smallest sample eigenvalues $\{\lambda_{n,p-j}^{S_n}\}$, $j = 0, \ldots, N_2 - 1$ tend to their respective limits $\{\psi(\alpha_k)\}$, $k = K, \ldots, K - K_2 + 1$;

iv. the immediately following smallest eigenvalue $\lambda_{p-N_2}^{S_n}$ tends to the left edge a_y.

Hence we have recovered the content of Theorem 1.1 of [5].

4.2. An example of generalized spike eigenvalues

Assume that T_p is diagonal with three base eigenvalues $\{1, 4, 10\}$, nearly $p/3$ times for each of them, and there are four spike eigenvalues $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (15,\ 6,\ 2,\ 0.5)$, with respective multiplicities $(n_k) = (3, 2, 2, 2)$. The limiting population-sample ratio is taken to be $y = 0.3$. The limiting population spectrum H is then the uniform distribution on $\{1, 4, 10\}$. The support of the limiting Marčenko-Pastur distribution $F_{0.3,H}$ contains two intervals [0.32, 1.37] and [1.67, 18], see §3.1. The ψ-function of (3.2) for the current case is displayed in Figure 1. For simulation, we use $p' = 600$ so that T_p has the following 609 eigenvalues:

$$15,\ 15,\ 15,\ \underbrace{10, \ldots, 10}_{200},\ 6,\ 6,\ \underbrace{4, \ldots, 4}_{200},\ 2,\ 2, \underbrace{1, \ldots, 1}_{200},\ 0.5,\ 0.5\ .$$

From the table

spike α_k	15	6	2	0.5
multiplicity n_k	3	2	2	2
$\psi'(\alpha_k)$	$+$	$-$	$+$	$-$
$\psi(\alpha_k)$	18.65	5.82	1.55	0.29
descending ranks	1, 2, 3	204, 205	406, 407	608, 609

we see that 6 is a close spike for H while the three others are distant ones. By Theorems 4.1 and 4.2, we know that

- the 7 sample eigenvalues $\lambda_j^{S_n}$ with $j \in \{1,\ 2,\ 3,\ 406,\ 407,\ 608,\ 609\}$ associated to distant spikes tend to 18.65, 1.55 and 0.29, respectively, which are located outside the support of limiting distribution $F_{0.3,H}$ (or G);
- the two sample eigenvalues $\lambda_j^{S_n}$ with $j = 204, 205$ associated to the close spike 6 tend to a limit located inside the support, the γ-th quantile of the limiting distribution G where $\gamma = H(0, 6) = 2/3$.

These facts are illustrated by a simulation sample displayed in Figure 5.

5. CLT for sample eigenvalues from distant generalized spikes

Following Theorem 4.1, to any distant generalized spike eigenvalue α_k, there is a packet of n_k consecutive sample eigenvalues $\{\lambda_j^{S_n} : j \in J_k\}$ converging to $\psi(\alpha_k) \notin \Gamma_G$ where J_k are the descending ranks of α_k among the eigenvalues of T_p (counting multiplicities). The aim of this section is to introduce a CLT for the n_k-dimensional vector

$$\sqrt{n}\{\lambda_j^{S_n} - \psi(\alpha_k)\}\ , \quad j \in J_k.$$

The method of derivation is exactly the same as in Bai and Yao [4] which considers Johnstone's spiked population model. Therefore, we will give a condensed description of the result and refer to Bai and Yao [4] for technical derivations.

Let us decompose the observation vectors $\mathbf{x}_j = T_p^{1/2}\mathbf{u}_j$, $j = 1, \ldots, n$, where $\mathbf{u}_j = (w_{ij})_{1\le i\le p}$ by blocks,

$$\mathbf{x}_j = \begin{pmatrix} \boldsymbol{\xi}_j \\ \boldsymbol{\eta}_j \end{pmatrix}, \quad \text{with} \quad \boldsymbol{\xi}_j = \Sigma^{1/2}(w_{ij})_{1\le i\le M}, \quad \boldsymbol{\eta}_j = V_p^{1/2}(w_{ij})_{M<i\le p}.$$

Let

$$X_1 = \frac{1}{\sqrt{n}}(\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)_{M\times n} = \frac{1}{\sqrt{n}}\boldsymbol{\xi}_{1:n},\ X_2 = \frac{1}{\sqrt{n}}(\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n)_{p'\times n} = \frac{1}{\sqrt{n}}\boldsymbol{\eta}_{1:n}\,.$$

For $\lambda \notin \Gamma_G$, let us define the following fundamental random matrix

$$R_n = R_n(\lambda) = \frac{1}{\sqrt{n}}\left\{\xi_{1:n}(I + A_n)\xi_{1:n}^* - \Sigma \mathrm{tr}(I + A_n)\right\}, \tag{5.1}$$

with

$$A_n = A_n(\lambda) = X_2^*(\lambda I - X_2X_2^*)^{-1}X_2, \quad \lambda \notin \Gamma_G.$$

For the statement of our result, we first need to find the limit distribution of the sequence $\{R_n(\lambda)\}$. These limit distributions are given in Propositions

3.1 and 3.2 of [4] for the real and complex cases respectively. To ease the reading of the paper, let us give a brief summary. We have for $\lambda \notin \Gamma_G$,

i. if the variables (w_{ij}) are real-valued, the random matrix $R_n(\lambda)$ converges weakly to a symmetric random matrix $R(\lambda) = (R_{ij}(\lambda))$ with zero-mean Gaussian entries having an explicitly known covariance function ;
ii. if the variables (w_{ij}) are complex-valued, the random matrix R_n converges weakly to a zero-mean Hermitian random matrix $R(\lambda) = (R_{ij}(\lambda))$. Moreover, the real and imaginary parts of its upper-triangular bloc $\{R_{ij}(\lambda),\ 1 \le i \le j \le M\}$ form a $2K$-dimensional Gaussian vector with an explicitly known covariance matrix.

Finally, let be the spectral decomposition of Σ,

$$\Sigma = U \begin{pmatrix} \alpha_1 I_{n_1} & \cdots & 0 \\ 0 & \ddots & 0 \\ \cdots & 0 & \alpha_K I_{n_K} \end{pmatrix} U^* \ , \tag{5.2}$$

where U is an unitary matrix. Let $\psi_k = \psi(\alpha_k)$ and $R(\psi_k)$ be the weak Gaussian limit of the sequence of matrices of random forms $[R_n(\psi_k)]_n$ recalled above (in both real and complex variables case). Define

$$\widetilde{R}(\psi_k) = U^* R(\psi_k) U \ ,$$

and

$$m_3(\psi_k) = \int \frac{x}{(\psi_k - x)^2} dG(x).$$

Applying the method introduced in [4], we have the following

Theorem 5.3. *For each distant generalize spike eigenvalue, the n_k-dimensional real vector*

$$\sqrt{n}\{\lambda_j^{S_n} - \psi_k,\ j \in J_k\} \ ,$$

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

$$\frac{1}{1+ym_3(\psi_k)\alpha_k}\widetilde{R}_{kk}(\psi_k).$$

where $\widetilde{R}_{kk}(\psi_k)$ is the k-th diagonal block of $\widetilde{R}(\psi_k)$ corresponding to the indices $\{u, v \in J_k\}$.

It is worth noticing that the limiting distribution of such n_k packed sample extreme eigenvalues are generally *non Gaussian* and asymptotically dependent. Indeed, the limiting distribution of a single sample extreme eigenvalue $\lambda_j^{S_n}$ is Gaussian if and only if the corresponding generalized spike eigenvalue is simple. We refer the reader to [4] for detailed examples illustrating these same facts but for Johnstone's model.

Acknowledgements

We are grateful to Referees for their very careful reading. Their comments have led to significant improvements of the proofs of Theorems 4.1 and 4.2 and a more complete biography on considered subjects.

References

[1] 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.

[2] Z.D. Bai and J.W. Silverstein. Exact separation of eigenvalues of large dimensional sample covariance matrices. *Ann. Probab.*, 27(3):1536–1555, 1999.

[3] Z.D. Bai and J.W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices.* Science Press, Beijing, 2006.

[4] Z.D. Bai and J.F. Yao. Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. H. Poincaré Probab. Statist.*, 44: 447–474, 2008.

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

[6] 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.

[7] F. Benaych-Georges and N. Raj Rao. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. Technical report, `http://arxiv.org/abs/0910.2120`, 2009.

[8] F. Benaych-Georges, A. Guionnet, and M. Maïda. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. Technical report, `http://arxiv.org/abs/1009.0145`, 2010.

[9] Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *Ann. Probab.*, 37 (1):1–47, 2009.

[10] D. Féral and S. Péché. The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.*, 272(1):185–228, 2007.

[11] I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statistics*, 29(2):295–327, 2001.

[12] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, 1:457–483, 1967.

[13] Alexei Onatski. Asymptotics of the principal components estimator of large factor models with weak factors. Technical report, Columbia University, 2005.

[14] Debashis Paul. Asymptotics of sample eigenstruture for a large dimensional spiked covariance mode. *Statistica Sinica*, 17:1617–1642, 2007.

[15] S. Péché. The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields*, 134(1):127–173, 2006.

[16] Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.*, 55(2):331–339, 1995.

[17] Jack W. Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.*, 54(2):295–309, 1995.

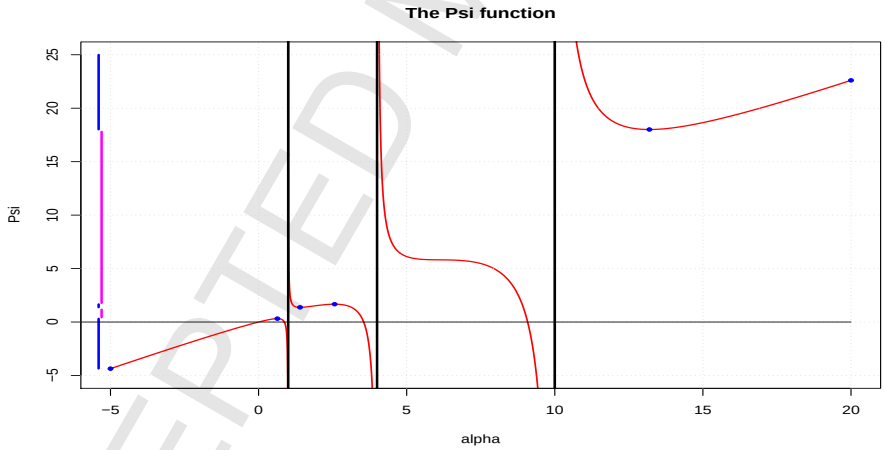

FIGURE 1. *The ψ function for the Marčenko-Pastur distribution $F_{0.3,H}$ with H the uniform distribution on the set $\{1, 4, 10\}$. Blue points indicate intervals where $\psi' > 0$. Singular points of ψ are indicated as vertical lines corresponding to the support of H. On the left, the support set of $F_{0.3,H}$ (except the point 0) and its complementary set are indicated as magenta and blue segments respectively.*

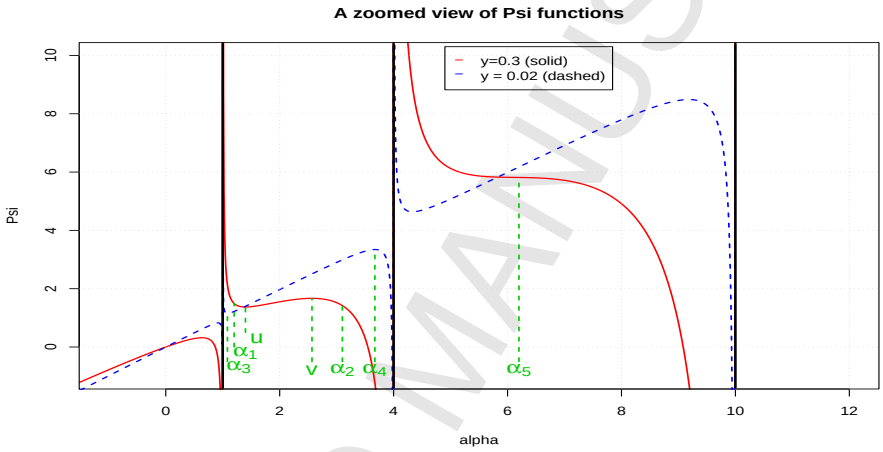

FIGURE 2. *A zoomed view of the ψ functions for the Marčenko-Pastur distribution $F_{0.3,H}$ (solid curve) and $F_{0.02,H}$ (dashed curve) with H the uniform distribution on the set $\{1, 4, 10\}$. The three points α_1, α_2 and α_5 are close spikes for $F_{0.3,H}$ where $\psi'_{0.3,H} \leq 0$. They become all distant spikes for $F_{0.02,H}$ as $\psi'_{0.02,H} > 0$.*

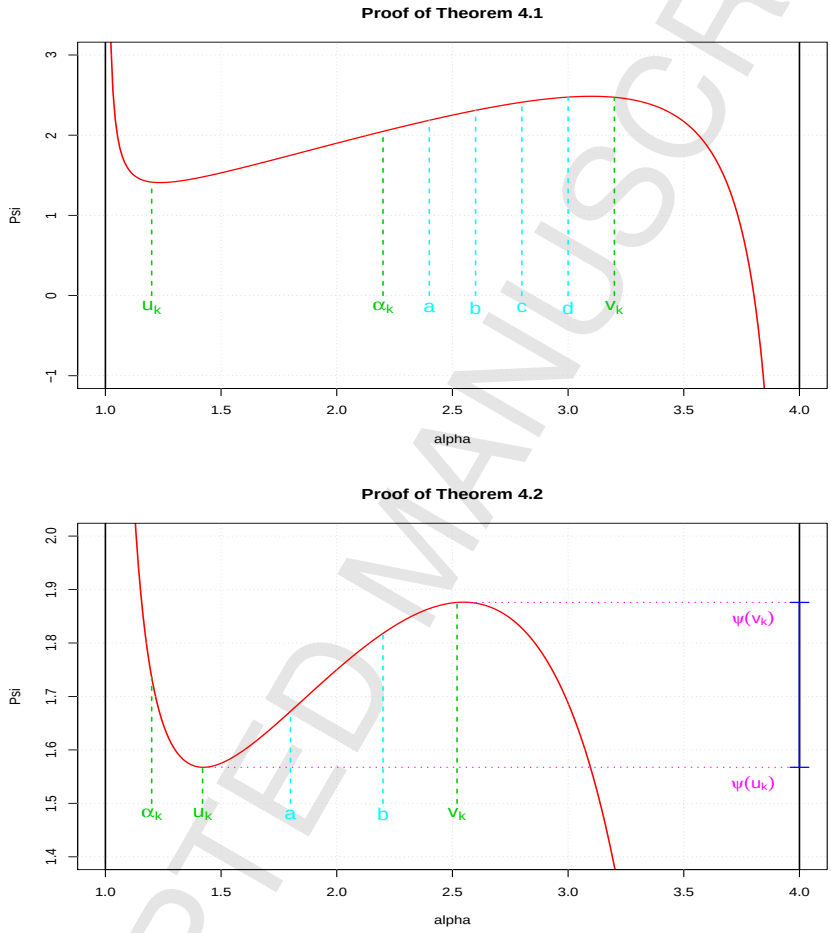

FIGURE 3. *Illustrating (top to bottom) the proofs of Theorems 4.1 and 4.2.*

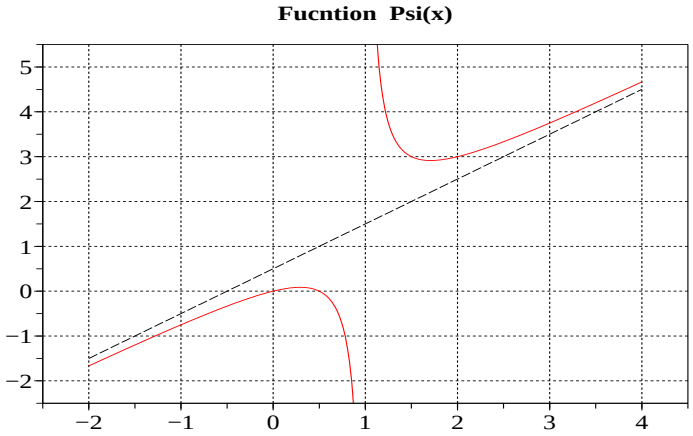

FIGURE 4. *The function* $\alpha \mapsto \psi(\alpha) = \alpha + y\alpha/(\alpha - 1)$ *which maps a spike eigenvalue* α *to the limit of an associated sample eigenvalue in Johnstone's spiked population model. Figure with* $y = \frac{1}{2}$*;* $[1 \mp \sqrt{y}] = [0.293,\ 1.707]$*;* $[(1 \mp \sqrt{y})^2] = [0.086,\ 2.914]$.

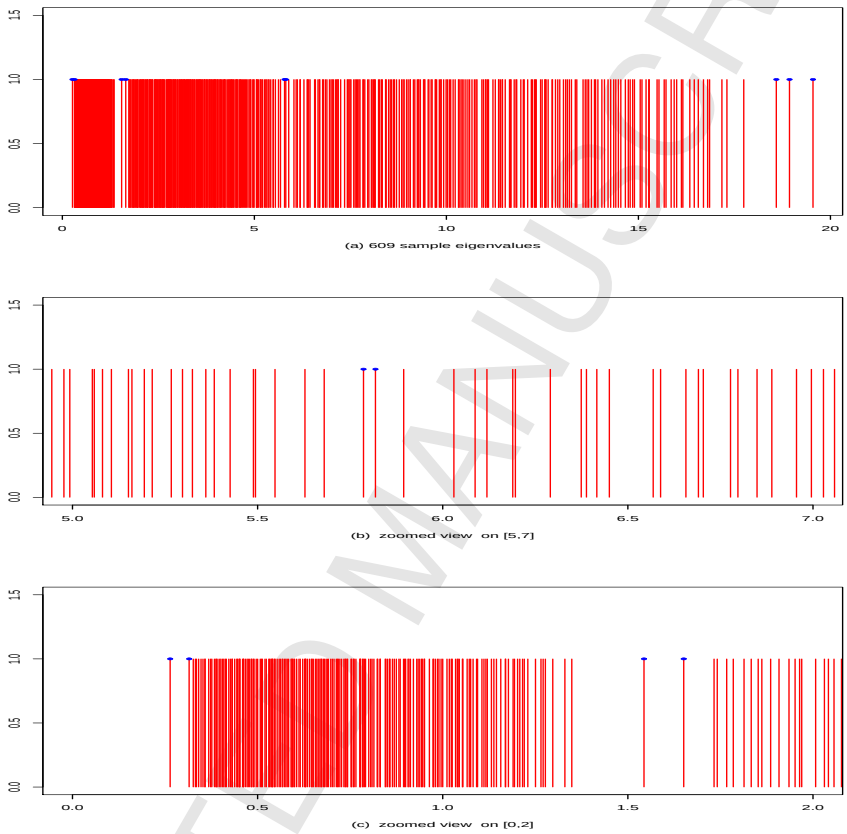

FIGURE 5. *An example of* $p = 609$ *sample eigenvalues (a), and two zoomed views (b) and (c) on [5,7] and [0,2] respectively. The limiting distribution of the ESD has support* $[0.32, 1.37] \cup [1.67, 18.00]$. *The 9 sample eigenvalues* $\{\lambda_j^{S_n},\ j = 1, 2, 3, 204, 205, 406, 407, 608, 609\ \}$ *associated to the spikes are marked with a blue point. Gaussian entries.*