Journal of Multivariate Analysis 130 (2014) 194–207

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A note on the CLT of the LSS for sample covariance matrix from a spiked population model

Qinwen Wang [a,*], Jack W. Silverstein [b], Jian-feng Yao [c]

[a] Department of Mathematics, Zhejiang University, China

[b] Department of Mathematics, North Carolina State University, United States

[c] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong

ARTICLE INFO

Article history:
Received 16 August 2013
Available online 14 May 2014

AMS 2000 subject classifications:
primary 62H15
secondary 60F05

Keywords:
Large-dimensional sample covariance matrices
Spiked population model
Central limit theorem
Centering parameter
Factor models

ABSTRACT

In this note, we establish an asymptotic expansion for the centering parameter appearing in the central limit theorems for linear spectral statistic of large-dimensional sample covariance matrices when the population has a spiked covariance structure. As an application, we provide an asymptotic power function for the corrected likelihood ratio statistic for testing the presence of spike eigenvalues in the population covariance matrix. This result generalizes an existing formula from the literature where only one simple spike exists.

1. Introduction

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

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

Write $Z_n = (z_{ij})_{1\leq i\leq p, 1\leq j\leq n}$, where $p = p(n)$ is related to n such that when $n \to \infty$, $p/n \to y \in (0, +\infty)$. Then the matrix $S_n = \frac{1}{n}\Sigma_p^{1/2} Z_n Z_n^* \Sigma_p^{1/2}$ 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 $\mathbf{x}_j = \Sigma_p^{1/2}\mathbf{u}_j$ where $\mathbf{u}_j = (z_{ij})_{1\leq i\leq p}$ denotes the jth column of Z_n. Note that for any nonnegative definite $p \times p$ Hermitian matrix A, $A^{1/2}$ denotes a Hermitian square root and we call the *empirical spectral distribution* (ESD) F^A the distribution generated by its eigenvalues, that is, $F^A = \frac{1}{p}\sum_{j=1}^{p} \delta_{\alpha_j}$, where $\{\alpha_j\}$ are the eigenvalues of A.

We first recall some useful results from random matrix theory. For a thorough presentation of these results, the reader is referred to [5]. Assume that the ESD H_n of Σ_p (we denote $H_n = F^{\Sigma_p}$ for short) converges weakly to a nonrandom probability distribution H on $[0, \infty)$. It is then well-known that the ESD F^{S_n} of S_n, generated by its eigenvalues $\lambda_{n,1} \geq \cdots \geq \lambda_{n,p}$,

* Corresponding author.
E-mail addresses: wqw8813@gmail.com (Q. Wang), jack@ncsu.edu (J.W. Silverstein), jeffyao@hku.hk (J.-f. Yao).

http://dx.doi.org/10.1016/j.jmva.2014.04.021
0047-259X/

converges to a nonrandom *limiting spectral distribution* (LSD) $F^{y,H}$ [15,24], whose Stieltjes transform $m(z) \equiv m_{F^{y,H}(z)}$ is the unique solution to

$$m = \int \frac{1}{\lambda(1-y-yzm)-z} dH(\lambda) \tag{1.1}$$

in the set $\{m \in \mathbb{C} : -\frac{1-y}{z} + ym \in \mathbb{C}^+\}$ (the Stieltjes transform of a distribution G is defined as: $m_G(z) \equiv \int \frac{1}{\lambda - z} dG(\lambda)$, $\Im z \neq 0$). Eq. (1.1) takes a simpler form when $F^{y,H}$ is replaced by $\underline{F}^{y,H} \equiv (1-y)\delta_0 + yF^{y,H}$, which is the LSD of $\underline{S}_n \equiv \frac{1}{n} Z_n^* \Sigma_p Z_n$. Its Stieltjes transform

$$\underline{m}(z) \equiv m_{\underline{F}^{y,H}}(z) = -\frac{1-y}{z} + ym(z)$$

has the inverse

$$z = z(\underline{m}) = -\frac{1}{\underline{m}} + y \int \frac{t}{1+t\underline{m}} dH(t). \tag{1.2}$$

We call this $\underline{m}(z)$ the companion Stieltjes transform of $m(z)$. The so-called *null case* corresponds to the situation $\Sigma_p \equiv I_p$, so $H_n \equiv H \equiv \delta_1$ and the LSD G^y of S_n ($G^y \equiv F^{y,\delta_1}$) has an explicit density function:

$$dG^y = \frac{1}{2\pi xy} \sqrt{(b_y - x)(x - a_y)}, \quad a_y \leq x \leq b_y, \tag{1.3}$$

which is the seminal Marčenko–Pastur law with index y and support $[a_y, b_y]$, where $a_y = (1-\sqrt{y})^2$, $b_y = (1+\sqrt{y})^2$. In case of $y > 1$, the distribution has an additional mass of weight $1 - 1/y$ at the origin.

In this paper we consider the *spiked population model* introduced in [12] where the eigenvalues of Σ_p are:

$$\underbrace{a_1, \ldots, a_1}_{n_1}, \ldots, \underbrace{a_k, \ldots, a_k}_{n_k}, \underbrace{1, \ldots, 1}_{p-M}. \tag{1.4}$$

Here M and the multiplicity numbers (n_k) are fixed and satisfy $n_1 + \cdots + n_k = M$. In other words, all the population eigenvalues are unit except some fixed number of them (the spikes). The model can be viewed as a finite-rank perturbation of the null case. Obviously, the LSD of S_n is not affected by this perturbation. However, the asymptotic behavior of the extreme eigenvalues of S_n is significantly different from the null case. The analysis of this new behavior of extreme eigenvalues has been an active area in the last few years, see e.g. [3,4,23,7,9,17,10] and [8]. In particular, the base component of the population ESD H_n in the last three references has been extended to a form more general than the simple Dirac mass δ_1 of the null case. Beyond the sample covariance matrix, there are also in the literature several closely related works on the behavior of the extreme eigenvalues of a Wigner matrix or general Hermitian matrix perturbated, in multiplicative or additive form, by a low rank matrix, see [11].

For statistical applications, besides the principal components analysis which is indeed the origin of spiked models [12], large-dimensional strict factor models are equivalent to a spiked population model and can be analyzed using the above-mentioned results. Related recent contributions in this area include, among others, Mestre [16] Kritchman and Nadler [13,14] Onatski [18–20] and Passemier and Yao [22] and they all concern the problem of estimation and testing the number of factors (or spikes).

In this note, we analyze the effects caused by the spike eigenvalues on the fluctuations of linear spectral statistics of the form

$$\frac{1}{p} \sum_{i=1}^{p} f(\lambda_{n,i}) = \int f(x) dF^{S_n}(x) \equiv F^{S_n}(f), \tag{1.5}$$

where f is a given function. Similarly to the convergence of the ESD's, the presence of the spikes does not prevent a central limit theorem for $F^{S_n}(f)$; however as we will see, the centering term in the CLT will be modified according to the values of the spikes. As this term has no explicit form, our main result is an asymptotic expansion presented in Section 2. Section 3 shows how to deal with these integrals appearing in the main theorem by detailing the computation of three frequently used functions f. To illustrate the importance of such expansions, we present in Section 4 an application for the determination of the power function for testing the presence of spikes. Section 5 contains the proof of the main result.

2. Centering parameter in the CLT of the LSS from a spiked population model

Fluctuations of linear spectral statistics of form (1.5) are indeed covered by a central limit theory initiated in [6], which says that the rate $\int f(x) dF^{S_n}(x) - \int f(x) dF^{y_n, H_n}(x)$ (F^{y_n,H_n} is the finite counterpart of $F^{y,H}$, where $y_n \equiv p/n$, and $H_n \equiv F^{\Sigma_p}$) approaches zero is essentially $1/p$. Define $X_n(f) = p(F^{S_n}(f) - F^{y_n,H_n}(f))$, the main result stated in that paper is

the following:

Theorem 1 ([6]). *Assume that the Z-variables satisfy the condition*

$$\frac{1}{np}\sum_{ij}\mathbb{E}|z_{ij}^4|I(|z_{ij}|\geq\sqrt{n}\eta)\longrightarrow 0$$

for any fixed $\eta>0$ and that the following additional conditions hold:

(a) *For each n, $z_{ij}=z_{ij}^{(n)}$, $i\leq p, j\leq n$ are independent. $\mathbb{E}z_{ij}=0$, $\mathbb{E}|z_{ij}|^2=1$, $\max_{i,j,n}\mathbb{E}|z_{ij}|^4<\infty$, $p/n\to y$.*

(b) *Σ_p is $p\times p$ nonrandom Hermitian nonnegative definite with spectral norm bounded in p, with $F^{\Sigma_p}\xrightarrow{\mathcal{D}}H$ a proper c.d.f. Let $f_1,\ldots,f_k$ be functions analytic on an open region containing the interval*

$$\left[\liminf_n\lambda_{\min}^{\Sigma_p}I_{(0,1)}(y)(1-\sqrt{y})^2,\limsup_n\lambda_{\max}^{\Sigma_p}(1+\sqrt{y})^2\right].$$

Then

(i) *the random vector*

$$(X_n(f_1),\ldots,X_n(f_k)) \tag{2.6}$$

forms a tight sequence in n.

(ii) *If z_{ij} and Σ_p are real and $\mathbb{E}(z_{ij}^4)=3$, then (2.6) converges weakly to a Gaussian vector $(X_{f_1},\ldots,X_{f_k})$ with means*

$$\mathbb{E}X_f=-\frac{1}{2\pi i}\int_{\mathcal{C}}f(z)\frac{y\int\frac{\underline{m}(z)^3t^2dH(t)}{(1+t\underline{m}(z))^3}}{\left(1-y\int\frac{\underline{m}(z)^2t^2dH(t)}{(1+t\underline{m}(z))^2}\right)^2}dz \tag{2.7}$$

and covariance function

$$Cov(X_f,X_g)=-\frac{1}{2\pi^2}\int_{\mathcal{C}_1}\int_{\mathcal{C}_2}\frac{f(z_1)g(z_2)}{(\underline{m}(z_1)-\underline{m}(z_2))^2}\underline{m}'(z_1)\underline{m}'(z_2)dz_1dz_2 \tag{2.8}$$

($f,g\in\{f_1,\ldots,f_k\}$). The contours in (2.7) and (2.8) (two in (2.8), which may be assumed to be non-overlapping) are closed and are taken in the positive direction in the complex plane, each enclosing the support of $F^{y,H}$.

(iii) *If z_{ij} is complex with $\mathbb{E}(z_{ij}^2)=0$ and $\mathbb{E}(|z_{ij}|^4)=2$, then* (ii) *also holds, except the means are zero and the covariance function is $1/2$ times the function given in (2.8).*

In particular, the limiting mean function $\mathbb{E}X_f$ and covariance function $Cov(X_f,X_g)$ could be calculated from contour integrals involving parameters $\underline{m}(z)$ and H, which are both related to the LSD $F^{y,H}$ and the pre-given function f.

For the centering parameter $pF^{y_n,H_n}(f)$, it depends on a particular distribution F^{y_n,H_n} which is a finite-horizon proxy for the LSD of S_n. The difficulty is that F^{y_n,H_n} has no explicit form; it is indeed *implicitly* defined through $\underline{m}_n(z)$, which solves the equation:

$$z=-\frac{1}{\underline{m}_n}+y_n\int\frac{t}{1+t\underline{m}_n}dH_n(t) \tag{2.9}$$

(substitute $y_n=p/n$ for y and H_n for H in (1.2)). This distribution depends on the ESD H_n with spike eigenvalues.

More precisely, the ESD H_n of Σ_p under (1.4) is

$$H_n=\frac{p-M}{p}\delta_1+\frac{1}{p}\sum_{i=1}^k n_i\delta_{a_i}, \tag{2.10}$$

which converges to the Dirac mass δ_1 (corresponding to the null case $\Sigma_p=I_p$). So anything that is related to the LSD remains the same, such as the limiting parameters $\mathbb{E}X_f$ and $Cov(X_f,X_g)$. However for the centering term $pF^{y_n,H_n}(f)$, it is still not enough if we expand $F^{y_n,H_n}(f)$ only to the order $O(1/p)$, which will lead to a bias of order $O(1)$. As an example, let us consider the simplest case that $f(x)=x$, it is known that $F^{y_n,\delta_1}(f)=1$ (see [6]). Our result shows that $F^{y_n,H_n}(f)=1+\frac{1}{p}\sum_{i=1}^k n_ia_i-\frac{M}{p}+O(1/n^2)$ (see (3.20)). The difference between these two terms (multiply by p) is

$$p\left(F^{y_n,H_n}(x)-F^{y_n,\delta_1}(x)\right)=p\left(\frac{1}{p}\sum_{i=1}^k n_ia_i-\frac{M}{p}\right)=\sum_{i=1}^k n_i(a_i-1),$$

which is actually a constant that cannot be neglected.

The following main result gives an asymptotic expansion of this centering term. It is here reminded that, following Baik and Silverstein [4], for a *distant spike* a_i such that $|a_i-1|>\sqrt{y}$, the corresponding sample eigenvalue is fluctuating around the value of $\phi(a_i)=a_i+\frac{ya_i}{a_i-1}$; while for a *close spike* such that $|a_i-1|\leq\sqrt{y}$, the corresponding sample eigenvalue tends to the edge points a_y if $1-\sqrt{y}\leq a_i<1$ and b_y if $1<a_i\leq 1+\sqrt{y}$.

Theorem 2. *Suppose the population has a spiked structure as stated in (1.4) with k_1 distant spikes and $k - k_1$ close spikes (arranged in decreasing order). Let f be any analytic function on an open domain including the support of M-P distribution G^y and all the $\phi(a_i)$, $i \le k_1$. We have:*

$$F^{y_n,H_n}(f)$$

$$= -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}\right)\left(\frac{M}{y_n\underline{m}} - \sum_{i=1}^{k}\frac{n_i a_i^2 \underline{m}}{(1+a_i\underline{m})^2}\right)d\underline{m} \tag{2.11}$$

$$+ \frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f'\left(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}\right)\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_i\underline{m})(1+\underline{m})}\left(\frac{1}{\underline{m}} - \frac{y_n\underline{m}}{(1+\underline{m})^2}\right)d\underline{m} \tag{2.12}$$

$$+ \left(1-\frac{M}{p}\right)G^{y_n}(f) + \frac{1}{p}\sum_{i=1}^{k_1} n_i f(\phi(a_i)) + O\left(\frac{1}{n^2}\right). \tag{2.13}$$

Here $\underline{m}$ is short for $\underline{m}_n$, which is defined in (2.9), $G^{y_n}(f)$ is the integral of f with respect to the Marčenko–Pastur distribution in (1.3), with index y replaced by $y_n = p/n$.

(i) *When $0 < y_n < 1$, the first k_1 spike eigenvalues a_i's satisfy $|a_i - 1| > \sqrt{y_n}$, the remaining $k - k_1$ satisfy $|a_i - 1| \le \sqrt{y_n}$, $\mathcal{C}_1$ is a contour counterclockwise, when restricted to the real axes, encloses the interval $[\frac{-1}{1-\sqrt{y_n}}, \frac{-1}{1+\sqrt{y_n}}]$.*

(ii) *When $y_n \ge 1$, the first k_1 spike eigenvalues a_i's satisfy $a_i - 1 > \sqrt{y_n}$, the remaining $k - k_1$ satisfy $0 < a_i \le 1 + \sqrt{y_n}$, $\mathcal{C}_1$ is a contour clockwise, when restricted to the real axes, encloses the interval $[-1, \frac{-1}{1+\sqrt{y_n}}]$.*

If there are no distant spikes then the second term in (2.13) does not appear.

The proof is given in Section 5.

3. Detailed examples of expansion for some popular functions

In this section, we derive in detail the computation of asymptotic expansions of the centering terms for three popular functions: $f(x) = x$, $f(x) = x^2$ and $f(x) = \log(x)$ when $0 < y_n < 1$, which frequently appear as part of some well known statistics like LRT, empirical moments, etc. Such statistics can be found in Section 4. For general function $f(x) = x^l$, the computation can be done similarly with the help of a symbolic computation software like Mathematica and we provide the formula with $l = 4$ in the last subsection for reference. When calculating residues, we should find the poles inside the integral region, and it should be noticed that when the index $i \in [k_1 + 1, k]$, the corresponding a_i's satisfy $|a_i - 1| \le \sqrt{y_n}$, which is equivalent to $-\frac{1}{a_i} \in [\frac{-1}{1-\sqrt{y_n}}, \frac{-1}{1+\sqrt{y_n}}]$, so poles of $\{\underline{m} = -1\}$, $\{\underline{m} = -\frac{1}{a_i}, i = (k_1+1, \ldots, k)\}$ and $\{\underline{m} = \frac{1}{y_n - 1}\}$ (pole of the function $log(z)$) should be included in $\mathcal{C}_1$. Besides, from (i) in Theorem 2, $\mathcal{C}_1$ is counterclockwise.

Notice that in all the sections, $m = \underline{m} = \underline{m}_n$ denotes the Stieltjes transform defined in (2.9).

3.1. Example of $F^{y_n,H_n}(x)$

We first calculate (2.11) and (2.12) by considering their residues at $m = -1$.

$$(2.11) = -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1}\left(-\frac{1}{m} + \frac{y_n}{1+m}\right)\left(\frac{M}{y_n m} - \sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2}\right)dm, \tag{3.14}$$

and its residue at $m = -1$ equals to

$$\frac{M}{p} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{n_i a_i^2}{(1-a_i)^2}. \tag{3.15}$$

$$(2.12) = \frac{1}{2\pi ip}\oint_{\mathcal{C}_1}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_i m)(1+m)}\left(\frac{1}{m} - \frac{y_n m}{(1+m)^2}\right)dm, \tag{3.16}$$

and its residue at $m = -1$ equals to

$$\frac{1}{p}\sum_{i=1}^{k}\left[-n_i - \frac{1}{2}(1-a_i)n_i y_n \frac{\partial}{\partial m^2}\left(\frac{m}{1+a_i m}\right)^2\Big|_{m=-1}\right] = \frac{1}{p}\sum_{i=1}^{k}\left[-n_i + \frac{a_i n_i y_n}{(1-a_i)^2}\right]. \tag{3.17}$$

Besides, the residue of (2.11) + (2.12) at $m = -\frac{1}{a_i}, i = (k_1+1, \ldots, k)$ can be calculated as

$$\frac{1}{p} n_i \left(a_i + \frac{y_n a_i}{a_i - 1} \right). \tag{3.18}$$

$$(2.13) = 1 - \frac{M}{p} + \frac{1}{p} \sum_{i=1}^{k_1} n_i \left(a_i + \frac{y_n a_i}{a_i - 1} \right) + O\left(\frac{1}{n^2} \right). \tag{3.19}$$

Combine (3.15) and (3.17)–(3.19), we get:

$$F^{y_n, H_n}(x) = 1 + \frac{1}{p} \sum_{i=1}^{k} n_i a_i - \frac{M}{p} + O\left(\frac{1}{n^2} \right). \tag{3.20}$$

3.2. Example of $F^{y_n, H_n}(x^2)$

We first calculate (2.11) and (2.12) by considering their residues at $m = -1$.

$$(2.11) = -\frac{1}{2\pi i p} \oint_{\mathcal{C}_1} \left(-\frac{1}{m} + \frac{y_n}{1+m} \right)^2 \left(\frac{M}{y_n m} - \sum_{i=1}^{k} \frac{n_i a_i^2 m}{(1 + a_i m)^2} \right) dm, \tag{3.21}$$

and its residue at $m = -1$ equals to

$$\frac{2M}{p} + \frac{M}{n} + \frac{y_n^2}{p} \sum_{i=1}^{k} \frac{(1+a_i) n_i a_i^2}{(1-a_i)^3} - \frac{2y_n}{p} \sum_{i=1}^{k} \frac{n_i a_i^2}{(1-a_i)^2}. \tag{3.22}$$

$$(2.12) = \frac{2}{2\pi i p} \oint_{\mathcal{C}_1} \left(-\frac{1}{m} + \frac{y_n}{1+m} \right) \sum_{i=1}^{k} \frac{(1-a_i) n_i}{(1+a_i m)(1+m)} \left(\frac{1}{m} - \frac{y_n m}{(1+m)^2} \right) dm, \tag{3.23}$$

and its residue at $m = -1$ equals to

$$-\frac{2M}{p} - \frac{2y_n M}{p} + \frac{2y_n}{p} \sum_{i=1}^{k} \frac{n_i a_i^2}{(a_i - 1)^2} + \frac{2y_n}{p} \sum_{i=1}^{k} \frac{a_i n_i}{1 - a_i} - \frac{2}{p} \sum_{i=1}^{k} \frac{y_n^2 a_i^2 n_i}{(1-a_i)^3}. \tag{3.24}$$

Besides, the residue of (2.11) + (2.12) at $m = -\frac{1}{a_i}$ $(i > k_1)$ equals to

$$\frac{1}{p} n_i \left(a_i + \frac{y_n a_i}{a_i - 1} \right)^2. \tag{3.25}$$

And

$$(2.13) = \left(1 - \frac{M}{p} \right) (1 + y_n) + \frac{1}{p} \sum_{i=1}^{k_1} n_i \left(a_i + \frac{y_n a_i}{a_i - 1} \right)^2 + O\left(\frac{1}{n^2} \right). \tag{3.26}$$

Combine (3.22) and (3.24)–(3.26), we get:

$$F^{y_n, H_n}(x^2) = \frac{2}{n} \sum_{i=1}^{k} a_i n_i - \frac{2}{n} M + 1 + y_n - \frac{M}{p} + \frac{1}{p} \sum_{i=1}^{k} n_i a_i^2 + O\left(\frac{1}{n^2} \right).$$

3.3. Example of $F^{y_n, H_n}(\log(x))$

We first calculate (2.11) and (2.12) by considering their residues at $m = -1$.

$$\begin{aligned} (2.11) &= \frac{-1}{2\pi i p y_n} \oint_{\mathcal{C}_1} \frac{\log(\frac{y_n - 1}{m}) + \log(\frac{m - \frac{1}{y_n - 1}}{m+1})}{m} \left(M - \sum_{i=1}^{k} \frac{n_i a_i^2 y_n m^2}{(1 + a_i m)^2} \right) dm \\ &= \frac{-M}{2\pi i p y_n} \oint_{\mathcal{C}_1} \log\left(\frac{m - \frac{1}{y_n - 1}}{m+1} \right) \frac{1}{m} dm + \frac{1}{2\pi i p y_n} \oint_{\mathcal{C}_1} \log\left(\frac{m - \frac{1}{y_n - 1}}{m+1} \right) \sum_{i=1}^{k} \frac{n_i a_i^2 y_n m}{(1 + a_i m)^2} dm \\ &\triangleq A + B. \end{aligned} \tag{3.27}$$

$$\begin{aligned} A &= \frac{-M}{2\pi i p y_n} \oint_{\mathcal{C}_1} \log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \cdot d\log(m) \\ &= \frac{M}{2\pi i p y_n} \oint_{\mathcal{C}_1} \log(m) \cdot d\log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \\ &= \frac{M}{2\pi i p y_n} \cdot \frac{y_n}{y_n-1} \oint_{\mathcal{C}_1} \frac{\log(m)}{(m+1)(m-\frac{1}{y_n-1})} dm \\ &= -\frac{M}{p y_n} \log(1-y_n), \end{aligned} \tag{3.28}$$

$$\begin{aligned} B &= \frac{1}{2\pi i p} \oint_{\mathcal{C}_1} \log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \sum_{i=1}^{k} \frac{n_i a_i^2 m}{(1+a_i m)^2} dm \\ &= \frac{1}{2\pi i p} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} \log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) n_i a_i \left(\frac{1}{1+a_i m} - \frac{1}{(1+a_i m)^2}\right) dm \\ &\triangleq C - D, \end{aligned} \tag{3.29}$$

where

$$\begin{aligned} C &= \frac{1}{2\pi i p} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} \log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \frac{n_i a_i}{1+a_i m} dm \\ &= \frac{1}{2\pi i p} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} n_i \log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \cdot d\log(1+a_i m) \\ &= \frac{-1}{2\pi i p} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} n_i \log(1+a_i m) \cdot d\log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \\ &= \frac{-1}{2\pi i p} \cdot \frac{y_n}{y_n-1} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} \frac{n_i \log(1+a_i m)}{(m+1)(m-\frac{1}{y_n-1})} dm \\ &= \frac{1}{p} \sum_{i=1}^{k} n_i \log(1-a_i) - \frac{1}{p} \sum_{i=1}^{k} n_i \log\left(1+\frac{a_i}{y_n-1}\right), \end{aligned} \tag{3.30}$$

and

$$\begin{aligned} D &= \frac{1}{2\pi i p} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} \log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \frac{n_i a_i}{(1+a_i m)^2} dm \\ &= \frac{1}{2\pi i p} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} \frac{n_i}{1+a_i m} \cdot d\log\left(\frac{m-\frac{1}{y_n-1}}{m+1}\right) \\ &= \frac{y_n}{2\pi i p (y_n-1)} \sum_{i=1}^{k} \oint_{\mathcal{C}_1} \frac{n_i}{(1+a_i m)\left(m-\frac{1}{y_n-1}\right)(m+1)} dm \\ &= \frac{1}{p} \sum_{i=1}^{k} \left(\frac{n_i}{1+\frac{a_i}{y_n-1}} - \frac{n_i}{1-a_i}\right). \end{aligned} \tag{3.31}$$

Combine (3.27)–(3.31), we get the residue of (2.11) at $m=-1$:

$$\begin{aligned} &-\frac{M}{p y_n} \log(1-y_n) + \frac{1}{p} \sum_{i=1}^{k} n_i \log(1-a_i) - \frac{1}{p} \sum_{i=1}^{k} n_i \log\left(1+\frac{a_i}{y_n-1}\right) \\ &-\frac{1}{p} \sum_{i=1}^{k} \frac{n_i}{1+\frac{a_i}{y_n-1}} + \frac{1}{p} \sum_{i=1}^{k} \frac{n_i}{1-a_i}. \end{aligned} \tag{3.32}$$

Then, we consider the part (2.12) in the general formula influenced by the pole $m=-1$:

$$\begin{aligned}(2.12) &= -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f'\left(-\frac{1}{m}+\frac{y_n}{1+m}\right)\sum_{i=1}^{k}\left(\frac{n_i a_i}{1+a_i m}-\frac{n_i}{1+m}\right)\left(\frac{1}{m}-\frac{y_n m}{(1+m)^2}\right)dm\\ &= -\frac{1}{2\pi ip}\sum_{i=1}^{k} n_i\oint_{\mathcal{C}_1}\frac{m(m+1)}{y_n m-m-1}\left(\frac{a_i}{1+a_i m}-\frac{1}{1+m}\right)\left(\frac{1}{m}-\frac{y_n m}{(1+m)^2}\right)dm\\ &\triangleq \frac{-1}{2\pi ip(y_n-1)}\sum_{i=1}^{k} n_i(E-F-G+H),\end{aligned}$$

where

$$\begin{aligned}E &= \oint_{\mathcal{C}_1}\frac{a_i(m+1)}{(1+a_i m)(m-\frac{1}{y_n-1})} = 2\pi i\frac{y_n a_i}{y_n+a_i-1},\\ F &= \oint_{\mathcal{C}_1}\frac{a_i y_n m^2}{(m+1)(1+a_i m)(m-\frac{1}{y_n-1})} = 2\pi i\left(\frac{a_i(y_n-1)}{a_i-1}+\frac{a_i}{y_n+a_i-1}\right),\\ G &= \oint_{\mathcal{C}_1}\frac{1}{m-\frac{1}{y_n-1}} = 2\pi i,\\ H &= \oint_{\mathcal{C}_1}\frac{y_n m^2}{(m+1)^2(m-\frac{1}{y_n-1})}dm = 2\pi i y_n.\end{aligned}$$

Collecting these four terms, we have the residue of (2.12) at $m=-1$:

$$\frac{1}{p}\sum_{i=1}^{k}\left(\frac{1}{a_i-1}-\frac{a_i}{y_n+a_i-1}\right)n_i. \tag{3.33}$$

Then we consider the influence of (2.11) + (2.12) caused by the pole $m=-\frac{1}{a_i}$, $i=k_1+1,\ldots,k$, which can be calculated similarly as

$$\frac{n_i}{p}\log\left(a_i+\frac{y_n a_i}{a_i-1}\right). \tag{3.34}$$

Finally, using the known result that $G^{y_n}(\log(x))=(1-\frac{1}{y_n})\log(1-y_n)-1$, which has been calculated in [6], and combine (3.32)–(3.34) and (2.13), we get

$$F^{y_n,H_n}(\log(x)) = \frac{1}{p}\sum_{i=1}^{k} n_i\log(a_i)-1+\left(1-\frac{1}{y_n}\right)\log(1-y_n)+O\left(\frac{1}{n^2}\right).$$

3.4. Example of $F^{y_n,H_n}(x^l)$

Consider the general case of $f(x)=x^l$, and combination of (2.11) and (2.12) in Theorem 2 leads to:

$$F^{y_n,H_n}(x^l) = \frac{1}{2\pi ip}\oint_{\mathcal{C}_1} g(m)dm+\left(1-\frac{M}{p}\right)G^{y_n}(f)+\frac{1}{p}\sum_{i=1}^{k_1} n_i f(\phi(a_i))+O\left(\frac{1}{n^2}\right),$$

where

$$\begin{aligned}g(m) :=\ & l\left(-\frac{1}{m}+\frac{y_n}{1+m}\right)^{l-1}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_i m)(1+m)}\left(\frac{1}{m}-\frac{y_n m}{(1+m)^2}\right)\\ &-\left(-\frac{1}{m}+\frac{y_n}{1+m}\right)^{l}\left(\frac{M}{y_n m}-\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2}\right).\end{aligned}$$

Then the main task is to calculate the residue of $g(m)$ at $m=-1$ and $m=-1/a_i$, $i=(k_1+1,\ldots,k)$, which can be done with the help of a symbolic computation software like Mathematica.

We present the result of $l = 4$ in the following for reference while skipping calculation details:

$$F^{y_n,H_n}(x^4) = \frac{1}{p}\sum_{i=1}^{k} n_i a_i^4 + \frac{4}{n}\sum_{i=1}^{k} n_i a_i^3 + \left(\frac{4}{n} + \frac{6y}{n}\right)\sum_{i=1}^{k} n_i a_i^2 + \left(\frac{4}{n} + \frac{12y}{n} + \frac{4y^2}{n}\right)\sum_{i=1}^{k} n_i a_i - \frac{3M}{n}(2 + 4y + y^2) + \left(1 - \frac{M}{p}\right)(1 + y)(1 + 5y + y^2).$$

4. An application to the test of presence of spike eigenvalues

Suppose that $\{\mathbf{x}\}$ follows a p-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma_p)$ and we want to test

$$H_0 : \Sigma_p = I_p \quad \text{vs.} \quad H_1 : \Sigma_p \neq I_p,$$

where I_p denotes the p-dimensional identity matrix. This test has been detailed in textbook like [1, Chapter 10]. Given a sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ from $\mathbf{x}$, the likelihood ratio criterion is

$$\lambda = \frac{\max L(I_p)}{\max_{\Sigma_p} L(\Sigma_p)}, \tag{4.35}$$

where the likelihood function is

$$L(\Sigma_p) = (2\pi)^{-\frac{1}{2}pn} |\Sigma_p|^{-\frac{1}{2}n} exp\left[-\frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i^* \Sigma_p^{-1} \mathbf{x}_i\right].$$

More explicitly,

$$\lambda = \left(\frac{e}{n}\right)^{\frac{1}{2}pn} |nS_n|^{\frac{1}{2}n} e^{-\frac{n}{2}\mathrm{tr}S_n}, \tag{4.36}$$

where S_n is the sample covariance matrix defined as

$$S_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^*.$$

Further, taking the log function on (4.36) and multiplying by $-2/n$ leads to the statistic: $L^* = \mathrm{tr}S_n - \log|S_n| - p$. Denote $T_n = n \cdot L^*$, classical theory states that T_n converges to the $\chi^2_{1/2p(p+1)}$ distribution under H_0 when the value of p is fixed while letting $n \to \infty$. However, this classical approximation was shown incorrect when dealing with large-dimensional data, say, in the scheme of "$p \to \infty, n \to \infty, p/n \to y \in (0, 1)$" by Bai et al. [2]. In this scheme, the limiting dimension-to-sample size ratio y should be kept smaller than 1 to avoid null eigenvalues appearing in the term $\log|S_n|$.

The main reason that classical asymptotic theory fails is that for large p, T_n approaches infinity. Therefore, Bai et al. [2] modify the limit distribution under H_0 to cope with large-dimensional data. Since $L^* = \mathrm{tr}S_n - \log|S_n| - p = \sum_{i=1}^{p}\left(\lambda_{n,i} - \log\lambda_{n,i} - 1\right)$, using the CLT for LSS derived in Bai and Silverstein [6] with the function $f(x) = x - \log x - 1$, Bai et al. [2] prove that under H_0,

$$L^* - pG^{y_n,H_n}(f) \Rightarrow N(m(f), v(f)),$$

where

$$G^{y_n,H_n}(f) = 1 - \frac{y_n - 1}{y_n}\log(1 - y_n), \tag{4.37}$$

$$m(f) = -\frac{\log(1 - y)}{2}, \tag{4.38}$$

$$v(f) = -2\log(1 - y) - 2y. \tag{4.39}$$

At a significance level α (usually 0.05), the test will reject H_0 when $L^* - pG^{y_n,H_n}(f) > m(f) + \Phi^{-1}(1 - \alpha)\sqrt{v(f)}$ where Φ is the standard normal cumulative distribution function.

However, the power function of this test remains unknown because the distribution of L^* under the general alternative hypothesis H_1 is ill-defined. Let us consider this general test as a way to test the null hypothesis H_0 above against an alternative hypothesis of the form:

H_1^* : Σ_p has the spiked structure (1.4),

which corresponds to the existence of a low-dimensional structure or signal in the data. In other words, we want to test the absence against the presence of possible spike eigenvalues in the population covariance matrix. The general asymptotic expansion in Theorem 2 helps to find the power function of the test.

More precisely, under the alternative H_1^* and for $f(x) = x - \log x - 1$ used in the statistic L^*, the centering term $F^{y_n,H_n}(f)$ can be found to be

$$1 + \frac{1}{p}\sum_{i=1}^{k} n_i a_i - \frac{M}{p} - \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - \left(1 - \frac{1}{y_n}\right)\log(1 - y_n) + O\left(\frac{1}{n^2}\right), \tag{4.40}$$

thanks to the expansion found in Section 3:

$$F^{y_n,H_n}(x) = 1 + \frac{1}{p}\sum_{i=1}^{k} n_i a_i - \frac{M}{p} + O\left(\frac{1}{n^2}\right)$$

and

$$F^{y_n,H_n}(\log x) = \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - 1 + \left(1 - \frac{1}{y_n}\right)\log(1 - y_n) + O\left(\frac{1}{n^2}\right).$$

Therefore we have obtained that under H_1^*,

$$L^* - pF^{y_n,H_n}(f) \Rightarrow N(m(f), v(f)),$$

where this time, the value of the centering term $F^{y_n,H_n}(f)$ is given in (4.40) while the values of $m(f)$ and $v(f)$ remain the same as in (4.38) and (4.39). It follows that the asymptotic power function of the test is

$$\beta(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^{k} n_i(a_i - 1 - \log a_i)}{\sqrt{-2\log(1-y) - 2y}}\right).$$

In the particular case where the spiked model has only one simple close spike, i.e. $k = 1, k_1 = 0, n_1 = 1$, the above power function becomes

$$\beta(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{a_1 - 1 - \log a_1}{\sqrt{-2\log(1-y) - 2y}}\right), \tag{4.41}$$

which is exactly the formula (5.6) found in [21]. Note that these authors have found this formula using sophisticated tools of asymptotic contiguity and Le Cam's first and third lemmas, our derivation is in a sense much more direct.

5. Proof of Theorem 2

Proof. We divide the proof into three parts according to whether $0 < y_n < 1, y_n > 1$ or $y_n = 1$.

Case of $0 < y_n < 1$:

Recall that $G^{y_n}(f) = \int f(x)dG^{y_n}(x)$ when no spike exists, where G^{y_n} is the M-P distribution with index y_n. And by the Cauchy integral formula, it can be expressed as $-\frac{1}{2\pi i}\oint_{\gamma_1} f(z)m(z)dz$, where the integral contour γ_1 is chosen to be positively oriented, enclosing the support of G^{y_n} and its limit G^y. Due to the restriction that $0 < y_n < 1$, we choose γ_1 such that the origin $\{z = 0\}$ is not enclosed inside.

Using the relationship between $m(z)$ and $\underline{m}(z)$: $\underline{m}(z) = y_n m(z) - \frac{1-y_n}{z}$, we can rewrite

$$\begin{aligned} G^{y_n}(f) &= -\frac{1}{2\pi i}\oint_{\gamma_1} f(z)m(z)dz = -\frac{1}{2\pi i}\oint_{\gamma_1} f(z)\Big(\frac{\underline{m}(z)}{y_n} + \frac{1-y_n}{y_n z}\Big)dz \\ &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\gamma_1} f(z)\underline{m}(z)dz. \end{aligned} \tag{5.42}$$

Besides, for $z \notin supp(G^{y_n})$, $\underline{m}(z)$ satisfies the equation (with $dH_n(t)$ replaced by δ_1 in (2.9)):

$$z = -\frac{1}{\underline{m}} + \frac{y_n}{1 + \underline{m}}. \tag{5.43}$$

If we solve this Eq. (5.43), then the solution $\underline{m}(z)$ will involve the square root of some function of z. So, if we are trying to deal with the integral with respect to z in Eq. (5.42), it will be more intricate. For this reason, we choose to change the variable from z to $\underline{m}$ in Eq. (5.42), making it much easier to compute.

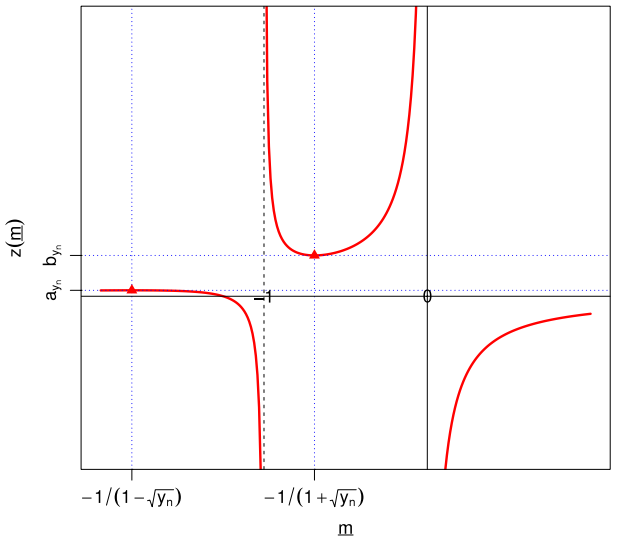

Fig. 1. The graph of the transform $z(\underline{m}) = -\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}$ when $0 < y_n < 1$.

We present the mapping (5.43) when $0 < y_n < 1$ in Fig. 1, restricting z and $\underline{m}$ to the real domain. From [25], we know that the z's such that $z'(m) > 0$ are not in the support of G^{y_n}. Therefore, we shall focus on the increasing intervals, where $z \to \underline{m}(z)$ is a one-to-one analytic map on $\mathbb{C}^+$, which admits a unique extension to $\mathbb{C} \setminus supp(G^{y_n})$. Moreover, when z runs along the contour γ_1 clockwise, which encloses $[a_{y_n}, b_{y_n}]$, $\underline{m}(z)$ will run along the contour, which we denote as $\mathcal{C}_1$, enclosing the interval $[\frac{-1}{1-\sqrt{y_n}}, \frac{-1}{1+\sqrt{y_n}}]$ (according to Fig. 1, it could be understood in this way: if we draw two horizontal lines, one is above $z(\underline{m}) = b_{y_n}$ and the other is below $z(\underline{m}) = a_{y_n}$, then these two horizontal lines will intersect with the red curve (the increasing parts). And we project the two points of intersection to the $\underline{m}$-axes, which will enclose the interval $[\frac{-1}{1-\sqrt{y_n}}, \frac{-1}{1+\sqrt{y_n}}]$). On these contours, we have by taking derivatives on both sides of (5.43) with respect to z, combining with Eq. (5.42):

$$G^{y_n}(f) = -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}\right)\underline{m}(z)\left(\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2}\right)d\underline{m}. \tag{5.44}$$

The point on γ_1 intersecting the real line to the left of a_{y_n} (right of b_{y_n}) maps to a point to the left of $\frac{-1}{1-\sqrt{y_n}}$ (right of $\frac{-1}{1+\sqrt{y_n}}$). Since the imaginary part of $\underline{m}(z)$ is the same sign as the imaginary part of z, we see that $\mathcal{C}_1$ is also oriented counterclockwise. Besides, $\underline{m} = -1$ is the pole contained in $\mathcal{C}_1$.

When the spiked structure (1.4) exists, by Eq. (2.9), this time the companion Stieltjes transform $\underline{m} = \underline{m}_n$ of F^{y_n, H_n} satisfies

$$\begin{aligned} z &= -\frac{1}{\underline{m}} + \frac{p-M}{p}\frac{y_n}{1+\underline{m}} + \frac{y_n}{p}\sum_{i=1}^{k}\frac{a_i n_i}{1+a_i\underline{m}}, \\ dz &= \left(\frac{1}{\underline{m}^2} - \frac{p-M}{p}\frac{y_n}{(1+\underline{m})^2} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{a_i^2 n_i}{(1+a_i\underline{m})^2}\right)d\underline{m}. \end{aligned} \tag{5.45}$$

Repeating the same computation as before, we get:

$$\begin{aligned} F^{y_n,H_n}(f) &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\gamma} f(z)\underline{m}(z)dz \\ &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f\left(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+\underline{m})(1+a_i\underline{m})}\right)\underline{m} \\ &\quad \times\left(\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2} + \frac{y_n}{p}\sum_{i=1}^{k} n_i\left[\frac{1}{(1+\underline{m})^2} - \frac{a_i^2}{(1+a_i\underline{m})^2}\right]\right)d\underline{m}, \end{aligned} \tag{5.46}$$

where γ is a positive oriented contour of z that encloses the support of F^{S_n} and its limit F^S. From [4], we know that under the spiked structure (1.4), the support of F^{S_n} consists of the support of M-P distribution: $[a_{y_n}, b_{y_n}]$ plus small intervals enclosing the points $\phi(a_i) = a_i + \frac{y_n a_i}{a_i - 1}$ $(i = 1, \ldots, k_1)$. Therefore, the contour γ can be expressed as $\gamma_1 \bigoplus(\bigoplus_{i=1}^{k_1}\gamma_{a_i})$ (γ_{a_i} is denoted

as the contour that encloses the point of $\phi(a_i)$). Moreover, $\mathcal{C}$ is the image of γ under the mapping (5.45), which can also be divided into $\mathcal{C}_1$ plus $\mathcal{C}_{a_i}$ $(i = 1, \ldots, k_1)$, with $\mathcal{C}_{a_i}$ enclosing $-\frac{1}{a_i}$ and all the contours are non-overlapping and positively oriented.

The term

$$\frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+\underline{m})(1+a_i\underline{m})}$$

is of order $O(\frac{1}{n})$, so we can take the Taylor expansion of f around the value of $-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}$, and the term

$$\frac{y_n}{p}\sum_{i=1}^{k}n_i\left[\frac{1}{(1+\underline{m})^2}-\frac{a_i^2}{(1+a_i\underline{m})^2}\right]$$

is also of order $O(\frac{1}{n})$. This gives rise to:

$$\begin{aligned}F^{y_n,H_n}(f) = &-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\left(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2}\right)d\underline{m}\\ &-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\frac{y_n}{p}\sum_{i=1}^{k}n_i\left[\frac{1}{(1+\underline{m})^2}-\frac{a_i^2}{(1+a_i\underline{m})^2}\right]\underline{m}d\underline{m}\\ &+\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f'\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+\underline{m})(1+a_i\underline{m})}\left(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2}\right)d\underline{m}+O\left(\frac{1}{n^2}\right).\end{aligned}\tag{5.47}$$

Then, we replace $\mathcal{C}$ appearing in Eq. (5.47) by $\mathcal{C}_1\bigoplus(\bigoplus_{i=1}^{k_1}\mathcal{C}_{a_i})$ as mentioned above, and thus we can calculate the value of (5.47) separately by calculating the integrals on the contour $\mathcal{C}_1$ and each $\mathcal{C}_{a_i}$ $(i = 1, \ldots, k_1)$. If there are no distant spikes then we will have just $\mathcal{C} = \mathcal{C}_1$.

The first term in Eq. (5.47) is equal to

$$-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\left(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2}\right)d\underline{m}\tag{5.48}$$

for the reason that the only poles: $\underline{m} = 0$ and $\underline{m} = -1$ are not enclosed in the contours C_{a_i} $(i = 1, \ldots, k_1)$.

Next, we consider these integrals on $\mathcal{C}_{a_i}(i = 1, \ldots, k_1)$.

The second term of Eq. (5.47) with the contour being $\mathcal{C}_{a_i}$ is equal to

$$\begin{aligned}&-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_{a_i}} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\frac{y_n}{p}\sum_{i=1}^{k}n_i\left[\frac{1}{(1+\underline{m})^2}-\frac{a_i^2}{(1+a_i\underline{m})^2}\right]\underline{m}d\underline{m}\\ &=\frac{n}{p}\frac{1}{2\pi in}\oint_{\mathcal{C}_{a_i}} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\sum_{i=1}^{k}\frac{a_i^2n_i\underline{m}}{(1+a_i\underline{m})^2}d\underline{m}\\ &=\frac{1}{2\pi ip}\oint_{\mathcal{C}_{a_i}}\frac{f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})\underline{m}n_i}{(\underline{m}+\frac{1}{a_i})^2}d\underline{m}\\ &=\frac{n_i}{p}\left[f(\phi(a_i))-f'(\phi(a_i))\left(a_i-\frac{y_na_i}{(a_i-1)^2}\right)\right],\end{aligned}$$

and the third term of Eq. (5.47) with the contour being $\mathcal{C}_{a_i}$ is equal to

$$\begin{aligned}&\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_{a_i}} f'\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+\underline{m})(1+a_i\underline{m})}\left(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2}\right)d\underline{m}\\ &=\frac{-1}{2\pi ip}\oint_{\mathcal{C}_{a_i}} f'\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\frac{n_i(1-a_i)}{(\underline{m}+\frac{1}{a_i})a_i(\underline{m}+1)}\left(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2}\right)d\underline{m}\\ &=\frac{1}{p}n_if'(\phi(a_i))\left(a_i-\frac{y_na_i}{(a_i-1)^2}\right).\end{aligned}$$

Combining these two terms, we get the influence of the distant spikes, that is, the integral on the contours $\bigcup_{i=1,\ldots,k_1} \mathcal{C}_{a_i}$, which equals to:

$$\frac{1}{p}\sum_{i=1}^{k_1} n_i f(\phi(a_i)). \tag{5.49}$$

So in the remaining part, we only need to consider the integral along the contour $\mathcal{C}_1$. Consider the second term of (5.47) with the contour being $\mathcal{C}_1$:

$$\begin{aligned}
&-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\frac{y_n}{p}\sum_{i=1}^{k} n_i\left[\frac{1}{(1+\underline{m})^2}-\frac{a_i^2}{(1+a_i\underline{m})^2}\right]\underline{m}d\underline{m}\\
&=-\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\left[\frac{1}{y_n}\left(\frac{M\underline{m}y_n}{(1+\underline{m})^2}-\frac{M}{\underline{m}}\right)+\frac{1}{y_n}\frac{M}{\underline{m}}-\sum_{i=1}^{k}\frac{n_i a_i^2\underline{m}}{(1+a_i\underline{m})^2}\right]d\underline{m}\\
&=-\frac{M}{p}\frac{n}{2\pi ip}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\left(\frac{\underline{m}y_n}{(1+\underline{m})^2}-\frac{1}{\underline{m}}\right)d\underline{m}\\
&\quad-\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\left(\frac{M}{\underline{m}y_n}-\sum_{i=1}^{k}\frac{n_i a_i^2\underline{m}}{(1+a_i\underline{m})^2}\right)d\underline{m}.
\end{aligned}\tag{5.50}$$

Combining Eqs. (5.44) and (5.48)–(5.50), we get:

$$\begin{aligned}
F^{y_n,H_n}(f) =& -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\left(\frac{M}{\underline{m}y_n}-\sum_{i=1}^{k}\frac{n_i a_i^2\underline{m}}{(1+a_i\underline{m})^2}\right)d\underline{m}\\
&+\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f'\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+\underline{m})(1+a_i\underline{m})}\left(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2}\right)d\underline{m}\\
&+\left(1-\frac{M}{p}\right)G^{y_n}(f)+\sum_{i=1}^{k_1}\frac{n_i}{p}f(\phi(a_i))+O\left(\frac{1}{n^2}\right).
\end{aligned}$$

Case of $y_n > 1$:

We also present the mapping (5.43) when $y_n > 1$ in Fig. 2.

When $y_n > 1$ there will be mass $1 - 1/y_n$ at zero. Assume first that f is analytic on an open interval containing 0 and b_{y_n} and let γ_1 be a contour covering $[a_{y_n}, b_{y_n}]$. Then we have in place of (5.42),

$$\begin{aligned}
G^{y_n}(f) &= \left(1-\frac{1}{y_n}\right)f(0)-\frac{1}{2\pi i}\oint_{\gamma_1} f(z)m(z)dz\\
&= \left(1-\frac{1}{y_n}\right)f(0)-\frac{1}{2\pi iy_n}\oint_{\gamma_1} f(z)\underline{m}(z)dz.
\end{aligned}$$

This time the $\underline{m}$ value corresponding to a_{y_n}, namely $\frac{-1}{1-\sqrt{y_n}}$, is positive, and so when changing variables the new contour $\mathcal{C}$ covers $[c_n, d_n]$ where $c_n < 0$ is slightly to the right of $\frac{-1}{1+\sqrt{y_n}}$, and $d_n > 0$ is slightly to the left of $\frac{-1}{1-\sqrt{y_n}}$, This interval includes the origin and not -1, and is oriented in a clockwise direction. We present these two contours γ_1 and $\mathcal{C}_1$ in Fig. 3.

We have in place of (5.44),

$$G_{y_n}(f) = \left(1-\frac{1}{y_n}\right)f(0)-\frac{1}{2\pi iy_n}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\underline{m}\left(\frac{1}{\underline{m}^2}-\frac{y_n}{(1+\underline{m})^2}\right)d\underline{m}.$$

Extend $\mathcal{C}_1$ to the following contour. On the right side on the real line continue $\mathcal{C}_1$ to a number large number r, then go on a circle $\mathcal{C}(r)$ with radius r in a counterclockwise direction until it returns to the point $r - i0$, then go left till it hits $\mathcal{C}_1$. This new contour covers pole -1 and not the origin, see Fig. 4. On $\mathcal{C}(r)$ we have using the dominated convergence theorem

$$\begin{aligned}
&\frac{1}{2\pi iy_n}\oint_{\mathcal{C}(r)} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\underline{m}\left(\frac{1}{\underline{m}^2}-\frac{y_n}{(1+\underline{m})^2}\right)d\underline{m}\quad(\text{with }\underline{m}=re^{i\theta})\\
&=\frac{1}{2\pi y_n}\int_0^{2\pi} f\left(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}}\right)\left(1-\frac{y_n\underline{m}^2}{(1+\underline{m})^2}\right)d\theta\to\frac{1-y_n}{y_n}f(0)\quad(\text{as } r\to\infty).
\end{aligned}$$

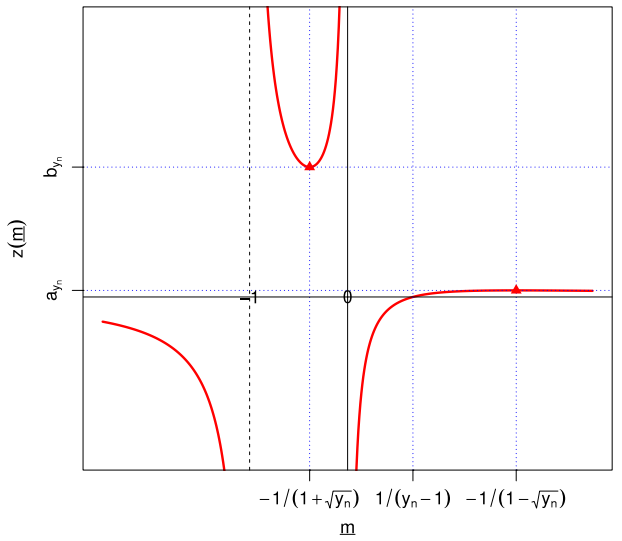

Fig. 2. The graph of the transform $z(\underline{m}) = -\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}$ when $y_n > 1$.

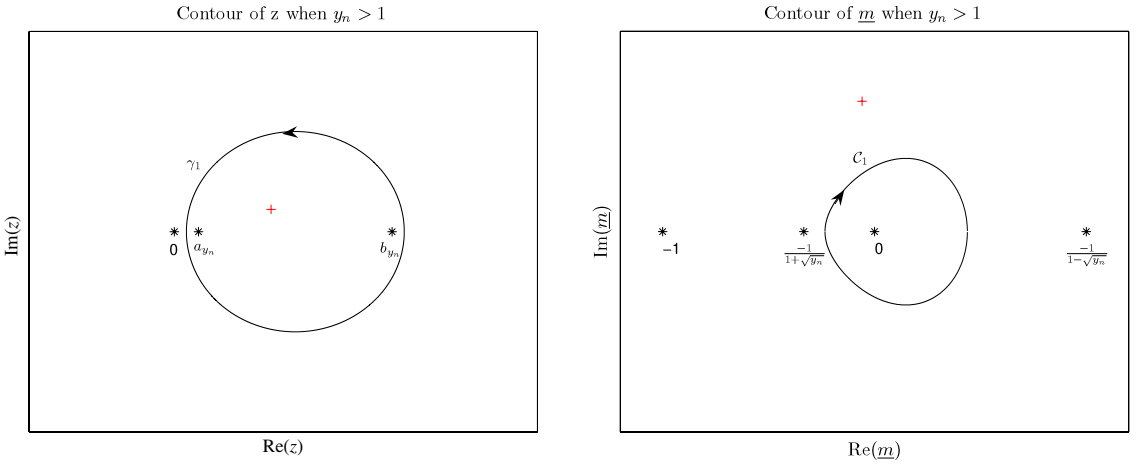

Fig. 3. Contours of z and $\underline{m}$ when $y_n > 1$.

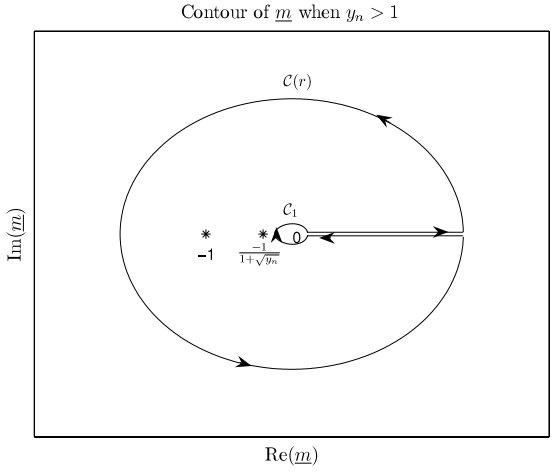

Fig. 4. The new contour of $\underline{m}$ when $y_n > 1$.

Therefore

$$G_{y_n}(f) = -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f\left(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}\right)\underline{m}\left(\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2}\right)d\underline{m} \tag{5.51}$$

where $\mathcal{C}_1$ just covers $[-1, \frac{-1}{1+\sqrt{y_n}}]$.

When there are spikes the only distant ones are those for which $a_i > \sqrt{y_n}+1$. We will get after the change of variable to $\underline{m}$ a contour which covers now $[c_n', d_n']$ where $c_n' < 0$ is to the right of the largest of $-\frac{1}{a_i}$ among the distant spikes (to the right of $\frac{-1}{1+\sqrt{y_n}}$ if there are no distant spikes), and $d_n' > 0$ is to the left of $\frac{-1}{1-\sqrt{y_n}}$, and oriented clockwise. We can extend the contour as we did before and get the same limit on the circle when there are no spikes. Therefore we get exactly (5.46) where now the contour $\mathcal{C}$ contains -1 and the largest of $-\frac{1}{a_i}$ among the distant spikes (contain $\frac{-1}{1+\sqrt{y_n}}$ if there are no distant spikes). Next, we can follow the same proof as for the case $0 < y_n < 1$, by slitting the contour $\mathcal{C}$ into $\mathcal{C} = \mathcal{C}_1 \bigoplus(\bigoplus_{i=1}^{k_1} \mathcal{C}_{a_i})$, where now $\mathcal{C}_1$ just contains the interval $[-1, \frac{-1}{1+\sqrt{y_n}}]$ and the contours $\mathcal{C}_{a_i}$ contain the influence of k_1 distant spikes $a_i > 1+\sqrt{y_n}$: $-\frac{1}{a_i}$ $(i = 1, \ldots, k_1)$, respectively. We thus obtain the same formula as in the case $0 < y_n < 1$. Therefore Theorem 1 follows where $\mathcal{C}_1$ contains just $[-1, \frac{-1}{1+\sqrt{y_n}}]$, and none of the $-\frac{1}{a_i}$ among the distant spikes ($-\frac{1}{a_i}$ are enclosed in the contour $\mathcal{C}_{a_i}$ as the case of $0 < y_n < 1$).

Case of $y_n = 1$:

For $y_n = 1$ we have $m(z) = \underline{m}(z)$, and the contour defining $G_1(f)$ must contain the interval $[0, 4]$. The contour in $\underline{m}$ contains $[c_n, d_n]$ where $-\frac{1}{2} < c_n < 0$, $d_n > 0$ and again is oriented in the clockwise direction. Extending again this contour we find the limit of the integral on the circle is zero for both $G_1(f)$ and $F^{1,H_n}(f)$, and we get again Theorem 1 where $\mathcal{C}_1$ is a contour containing $[-1, -\frac{1}{2}]$, and not the origin.

The proof of the theorem is complete. □

Acknowledgments

The first author's research was partially supported by the National Natural Science Foundation of China (Grant No. 11371317), the Natural Science Foundation of Zhejiang Province (No. R6090034), and the Doctoral Program Fund of Ministry of Education (No. J20110031). The second author's research was partially supported by the U.S. Army Research Office under Grant W911NF-09-1-0266. The third author's research was partially supported by RGC grant HKU 705413p.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, second ed., Wiley, New York, 1984.
[2] Z.D. Bai, D.D. Jiang, J.F. Yao, S.R. Zheng, Corrections to LRT on large dimensional covariance matrix by RMT, Ann. Statist. 37 (2009) 3822–3840.
[3] J. Baik, G. Ben Arous, S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, Ann. Probab. 33 (5) (2005) 1643–1697.
[4] J. Baik, J.W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, J. Multivariate Anal. 97 (2006) 1382–1408.
[5] Z.D. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, second ed., Springer, 2010, p. 20.
[6] Z.D. Bai, J.W. Silverstein, CLT for linear spectral statistics of large-dimensional sample covariance matrices, Ann. Probab. 32 (2004) 553–605.
[7] Z.D. Bai, J.F. Yao, CLT for eigenvalues in a spiked population model, Ann. Inst. Henri Poincaré Probab. Stat. 44 (3) (2008) 447–474.
[8] Z.D. Bai, J.F. Yao, On sample eigenvalues in a generalized spiked population model, J. Multivariate Anal. 106 (2012) 167–177.
[9] F. Benaych-Georges, A. Guionnet, M. Maida, Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices, Electron. J. Probab. 16 (2011) 1621–1662.
[10] F. Benaych-Georges, R.R. Nadakuditi, The eigenvalues and eigenvectors of finite low rank perturbations of large random matrices, Adv. Math. 227 (2) (2011) 494–521.
[11] Donati-Martin Capitaine, Février Féral, et al., Free convolution with a semi-circular distribution and eigenvalues of spiked deformations of Wigner matrices, Electron. J. Probab. 16 (2011) 1750–1792.
[12] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist. 29 (2) (2001) 295–327.
[13] S. Kritchman, B. Nadler, Determining the number of components in a factor model from limited noisy data, Chem. Int. Lab. Syst. 94 (2008) 19–32.
[14] S. Kritchman, B. Nadler, Non-parametric detection of the number of signals: hypothesis testing and random matrix theory, IEEE Trans. Signal Process. 57 (10) (2009) 3930–3941.
[15] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR-Sb 1 (1967) 457–483.
[16] X. Mestre, Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates, IEEE Trans. Infor. Theory 54 (11) (2008) 5113–5129.
[17] R.R. Nadakuditi, J.W. Silverstein, Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples, IEEE J. Sel. Top. Signal Process. 4 (3) (2010) 468–480.
[18] A. Onatski, Testing hypotheses about the number of factors in large factor models, Econometrica 77 (5) (2009) 1447–1479.
[19] A. Onatski, Determining the number of factors from empirical distribution of eigenvalues, Rev. Econ. Stat. 92 (4) (2010) 1004–1016.
[20] A. Onatski, Asymptotics of the principal components estimator of large factor models with weakly influential factors, J. Econometrics 168 (2012) 244–258.
[21] A. Onatski, M.J. Moreira, M. Hallin, Asymptotic power of sphericity tests for high-dimensional data, Ann. Statist. 41 (3) (2013) 1204–1231.
[22] D. Passemier, J.F. Yao, On determining the number of spikes in a high-dimensional spiked population model, Random Matrix: Theory Appl. 1 (2012) 1150002.
[23] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statist. Sinica 17 (2007) 1617–1642.
[24] J.W. Silverstein, Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices, J. Multivariate Anal. 55 (2) (1995) 331–339.
[25] J.W. Silverstein, S.I. Choi, Analysis of the limiting spectral distribution of large dimensional random matrices, J. Multivariate Anal. 54 (2) (1995) 295–309.