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

Qinwen Wang* **and Jack W. Silverstein†** **and Jian-feng Yao‡**

Qinwen Wang
Department of Mathematics
Zhejiang University
e-mail: wqw8813@gmail.com

Jack W. Silverstein
Department of Mathematics
North Carolina State University
e-mail: jack@ncsu.edu

Jianfeng Yao
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam, Hong Kong
e-mail: jeffyao@hku.hk

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.

AMS 2000 subject classifications: Primary 60F05; secondary 62H15.
Keywords and phrases: Large-dimensional sample covariance matrices, Spiked population model, Central limit theorem, Centering parameter, factor models.

1. Introduction

Let (Σ_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

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

*Research of this author was partly supported by the National Natural Science Foundation of China (Grant No. 11071213), the Natural Science Foundation of Zhejiang Province (No. R6090034), and the Doctoral Program Fund of Ministry of Education (No. J20110031).

†Research of this author was partly supported by the U.S. Army Research Office under Grant W911NF-09-1-0266.

‡Research of this author was partly supported by a HKU start-up grant.

Write $Z_n = (w_{ij})_{1\le i\le p, 1\le j\le n}$, 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 $S_n = \frac{1}{n}\Sigma_p^{1/2} Z_n Z_n^* \Sigma_p^{1/2}$ 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 $\mathbf{u}_j = (w_{ij})_{1\le i\le p}$ denotes the j-th column of Z_n. Note that for any nonnegative definite Hermitian matrix A, $A^{1/2}$ denotes a Hermitian square root and we call the *spectral distribution* (SD) the distribution generated by its eigenvalues.

Assume that the SD H_n of Σ_p converges weakly to a nonrandom probability distribution H on $[0, \infty)$. It is then well-known that the SD F^{S_n} of S_n, generated by its eigenvalues $\lambda_{n,1} \ge \cdots \ge \lambda_{n,p}$, converges to a nonrandom limiting SD G (Marčenko and Pastur, 1967; Silverstein, 1995). The so-called *null case* corresponds to the situation $\Sigma_p \equiv I_p$, so $H_n \equiv \delta_1$ and the limiting SD is the seminal Marčenko-Pastur law G^y with index y and support $[a_y, b_y]$ where $a_y = (1 - \sqrt{y})^2$, $b_y = (1 + \sqrt{y})^2$, and an additional mass at the origin if $y > 1$.

In this paper we consider the *spiked population model* introduced in Johnstone (2001) where the eigenvalues of Σ_p are

$$\underbrace{a_1, \cdots, a_1}_{n_1}, \ldots, \underbrace{a_k, \cdots, a_k}_{n_k}, \underbrace{1, \cdots, 1}_{p-M}. \tag{1.1}$$

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 limiting SD G of S_n is not affected by this perturbation. However, the asymptotic behaviour of the extreme eigenvalues of S_n is significantly different from the null case. The analysis of this new behaviour of extreme eigenvalues has been an active area in the last few years, see e.g. Baik et al. (2005), Baik and Silverstein (2006), Paul (2007), Bai and Yao (2008), Benaych-Georges et al. (2011), Nadakuditi and Silverstein (2010), Benaych-Georges and Nadakuditi (2011) and Bai and Yao (2012). In particular, the base component of the population SD 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.

For statistical applications, besides the principal components analysis which is indeed the origin of spiked models (Johnstone (2001)), 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 the area include, among others, Kritchman and Nadler (2008, 2009), Onatski (2009, 2010, 2012) and Passemier and Yao (2012) 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

$$T_n(f) = \sum_{i=1}^{p} f(\lambda_{n,i}) = F^{S_n}(f)\ , \tag{1.2}$$

where f is a given function. Similarly to the convergence of the SD's, the presence of the spikes does not prevent a central limit theorem for $T_n(f)$; 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. To illustrate the importance of such expansions, we present in Section 3 an application for the determination of the power function for testing the presence of spikes. The Appendix contains some technical derivations.

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

Fluctuations of linear spectral statistics of form (1.2) are indeed covered by a central limit theory initiated in Bai and Silverstein (2004). The theory was later improved by Pan and Zhou (2008) where the restriction $E(|w_{11}|^4) = 3$ matching the real Gaussian case was removed.

Let $f_1, \ldots, f_L$ be L functions analytic on an open domain of the complex plan including the support of the limiting SD. These central limit theorems state that the random vector

$$(X_n(f_1), \cdots, X_n(f_L)) \ ,$$

where

$$X_n(f) = p\left[F^{S_n}(f) - F^{y_n,H_n}(f)\right] = p\int f(x)d(F^{S_n} - F^{y_n,H_n})(x) \ ,$$

converges weakly to a Gaussian vector

$$(X_{f_1}, \cdots, X_{f_L})$$

with known mean function $E[X_f]$ and covariance function $Cov(X_f, X_g)$ that can be calculated from contour integrals involving parameters $\underline{m}(z)$ and H, where $\underline{m}(z)$ is the companion Stieltjes transform corresponding to the limiting SD of $\underline{S}_n = \frac{1}{n}Z_n^*\Sigma_p Z_n$. If the population has a spiked covariance structure, we know that the limit H and $\underline{m}(z)$ remain the same as the non-spiked case, so the limiting parameters $E[X_f]$ and $Cov(X_f, X_g)$ are also unchanged.

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

$$z = -\frac{1}{\underline{m}_n} + y_n\int\frac{t}{1+t\underline{m}_n}dH_n(t) \ . \tag{2.3}$$

This distribution depends on the SD H_n which in turn depends on the spike eigenvalues.

More precisely, the SD H_n of Σ_p is

$$H_n = \frac{p-M}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^{k} n_i \delta_{a_i} \ . \tag{2.4}$$

The term

$$\frac{1}{p}\sum_{i=1}^{k} n_i \delta_{a_i}$$

vanishes when p tends to infinity, so it has no influence when considering limiting spectral distributions. However for the CLT, the term $pF^{y_n,H_n}(f)$ has a p in front, and $\frac{1}{p}\sum_{i=1}^{k} n_i\delta_{a_i}$ times p is of order $O(1)$, thus cannot be neglected.

It is here reminded that, following Baik and Silverstein (2006), for a *distant spike* a_i such that $|a_i - 1| > \sqrt{y}$, the corresponding sample eigenvalue is equal to $\phi(a_i) = a_i + \frac{ya_i}{a_i - 1}$, while for a *close spike* such that $|a_i - 1| \leq \sqrt{y}$, the corresponding sample eigenvalue tends to the edge points a_y and b_y.

Our main result is an asymptotic expansion for this centering parameter.

Theorem 1. *Suppose the population has a spiked population structure as stated in* (1.1) *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 $\phi(a_i)$, $i \leq k_1$. We have:*

$$\begin{aligned} & F^{y_n,H_n}(f) \\ =& -\frac{1}{2\pi i p}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}})(\frac{M}{y_n\underline{m}} - \sum_{i=1}^{k}\frac{n_i a_i^2 \underline{m}}{(1+a_i\underline{m})^2})d\underline{m} && (2.5)\\ &+\frac{1}{2\pi i p}\oint_{\mathcal{C}_1} f'(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}})\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_i\underline{m})(1+\underline{m})}(\frac{1}{\underline{m}} - \frac{y_n\underline{m}}{(1+\underline{m})^2})d\underline{m} && (2.6)\\ &+(1-\frac{M}{p})G^{y_n}(f) + \frac{1}{p}\sum_{i=1}^{k_1} n_i f(\phi(a_i)) + O(\frac{1}{n^2}) \ ; && (2.7) \end{aligned}$$

Here $\underline{m} = \underline{m}_n$ is the companion Stieltjes transform of F^{y_n,H_n} defined in (2.3), *$G^{y_n}(f)$ is the integral of f with respect to the Marčenko-Pastur distribution with index $y_n = p/n$. And*

- *(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 $|a_i - 1| \leq \sqrt{y_n}$, $\mathcal{C}_1$ 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 $y_n \geq 1$, the first k_1 spike eigenvalues $a_i's$ satisfy $a_i - 1 > \sqrt{y_n}$, the remaining $k - k_1$ satisfy $0 < a_i \leq 1 + \sqrt{y_n}$, $\mathcal{C}_1$ 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 (2.7) does not appear.

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 it's 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 $\underline{m}(z)$ (the companion Stieltjes transform of $m(z)$): $\underline{m}(z) = y_n m(z) - \frac{1-y_n}{z}$, we can rewrite

$$\begin{aligned} G^{y_n}(f) &= -\frac{1}{2\pi i}\oint_{\gamma_1} f(z)m(z)dz = -\frac{1}{2\pi i}\oint_{\gamma_1} f(z)\Big(\frac{\underline{m}(z)}{y_n} + \frac{1-y_n}{y_n z}\Big)dz \\ &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\gamma_1} f(z)\underline{m}(z)dz \ . \end{aligned} \tag{2.8}$$

Besides, for $z \notin supp(G^{y_n})$, $\underline{m}(z)$ satisfies the equation:

$$z = -\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}} \ . \tag{2.9}$$

Taking derivatives on both sides with respect to z, we get:

$$dz = (\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2})d\underline{m} \ .$$

Changing the variable from z to $\underline{m}$ in equation (2.8), we get:

$$G^{y_n}(f) = -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}})\underline{m}(z)(\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2})d\underline{m} \ . \tag{2.10}$$

Here, the contour γ_1 of z in equation (2.8) is transformed into a contour of $\underline{m}$ through the mapping (2.9), denoted as $\mathcal{C}_1$.

We present the mapping (2.9) when $0 < y_n < 1$ in Figure 1, restricting z and $\underline{m}$ to the real domain. From Silverstein and Choi (1995), 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 a one-to-one mapping between z and $\underline{m}$ exists. From the figure, we see that when γ_1 is chosen to enclose the support of G^{y_n}: $[a_{y_n}, b_{y_n}]$, the corresponding $\mathcal{C}_1$ will enclose the interval $[\frac{-1}{1-\sqrt{y_n}}, \frac{-1}{1+\sqrt{y_n}}]$, and $\underline{m} = -1$ is the pole contained in this interval. 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.

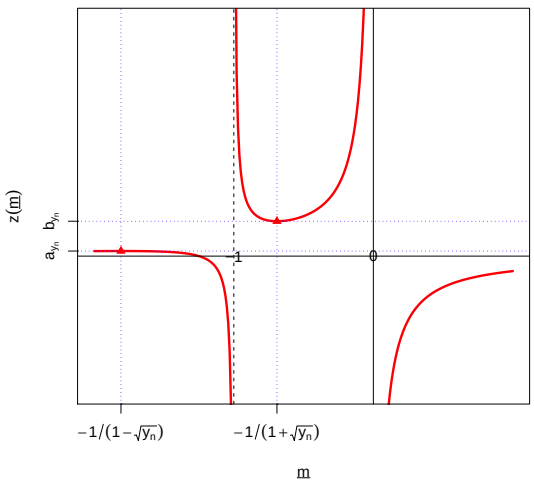

Figure 1: The graph of the transform $z(\underline{m}) = -\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}$ when $0 < y_n < 1$.

When the spiked structure (1.1) exists, by equation (2.3), this time the companion Stieltjes transform $\underline{m} = \underline{m}_n$ of F^{y_n, H_n} satisfies

$$z = -\frac{1}{\underline{m}} + \frac{p-M}{p}\frac{y_n}{1+\underline{m}} + \frac{y_n}{p}\sum_{i=1}^{k}\frac{a_i n_i}{1+a_i\underline{m}}\ , \tag{2.11}$$

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

Repeating the same computation as before, we get:

$$\begin{aligned} F^{y_n,H_n}(f) &= -\frac{n}{p}\frac{1}{2\pi i}\oint_\gamma f(z)\underline{m}(z)dz \\ &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f\left(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+\underline{m})(1+a_i\underline{m})}\right)\underline{m} \\ &\quad\times\left(\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2} + \frac{y_n}{p}\sum_{i=1}^{k} n_i\left[\frac{1}{(1+\underline{m})^2} - \frac{a_i^2}{(1+a_i\underline{m})^2}\right]\right)d\underline{m} \end{aligned} \tag{2.12}$$

where γ is a positively oriented contour of z that encloses the support of F^{S_n} and its limit F^S. From Baik and Silverstein (2006), we know that under the spiked structure (1.1), the support of F^{S_n} consists of the support of M-P distribution: $[a_{y_n}, b_{y_n}]$ plus small intervals near $\phi(a_i) = a_i + \frac{y_n a_i}{a_i - 1}$ $(i = 1, \cdots, k_1)$. 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 (2.11), which can also be divided into $\mathcal{C}_1$ plus $\mathcal{C}_{a_i}$ $(i = 1, \cdots, k_1)$, 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:

$$\begin{aligned}
F^{y_n,H_n}(f) &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2})d\underline{m}\\
&-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})\frac{y_n}{p}\sum_{i=1}^{k}n_i\left[\frac{1}{(1+\underline{m})^2}-\frac{a_i^2}{(1+a_i\underline{m})^2}\right]\underline{m}d\underline{m}\\
&+\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f'(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})\frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+\underline{m})(1+a_i\underline{m})}(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2})d\underline{m}\\
&+O(\frac{1}{n^2})\ . \qquad (2.13)
\end{aligned}$$

Then, we replace $\mathcal{C}$ appearing in equation (2.13) 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 (2.13) separately by calculating the integrals on the contour $\mathcal{C}_1$ and each $\mathcal{C}_{a_i}$ $(i = 1, \cdots, k_1)$. If there are no distant spikes then we will have just $\mathcal{C} = \mathcal{C}_1$.

The first term in equation (2.13) is equal to

$$-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})(\frac{1}{\underline{m}}-\frac{y_n\underline{m}}{(1+\underline{m})^2})d\underline{m} \qquad (2.14)$$

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

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

The second term of equation (2.13) with the contour being $\mathcal{C}_{a_i}$ is equal to

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

and the third term of equation (2.13) with the contour being $\mathcal{C}_{a_i}$ is equal to

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

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

$$\frac{1}{p}\sum_{i=1}^{k_1} n_i f(\phi(a_i))\ . \tag{2.15}$$

So in the remaining part, we only need to consider the integral along the contour $\mathcal{C}_1$. Consider the second term of (2.13) with the contour being $\mathcal{C}_1$:

$$\begin{aligned}
& -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})\frac{y_n}{p}\sum_{i=1}^{k} n_i\Big[\frac{1}{(1+\underline{m})^2}-\frac{a_i^2}{(1+a_i\underline{m})^2}\Big]\underline{m}d\underline{m} \\
= & -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})\Big[\frac{1}{y_n}(\frac{M\underline{m}y_n}{(1+\underline{m})^2}-\frac{M}{\underline{m}})+\frac{1}{y_n}\frac{M}{\underline{m}}-\sum_{i=1}^{k}\frac{n_i a_i^2\underline{m}}{(1+a_i\underline{m})^2}\Big]d\underline{m} \\
= & -\frac{M}{p}\frac{n}{2\pi ip}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})(\frac{\underline{m}y_n}{(1+\underline{m})^2}-\frac{1}{\underline{m}})d\underline{m} \\
& -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})\Big(\frac{M}{\underline{m}y_n}-\sum_{i=1}^{k}\frac{n_i a_i^2\underline{m}}{(1+a_i\underline{m})^2}\Big)d\underline{m}\ .
\end{aligned}\tag{2.16}$$

Combining Equations (2.10), (2.14), (2.15) and (2.16), we get:

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

Case of $y_n > 1$:

We also present the mapping (2.9) when $y_n > 1$ in Figure 2 below.

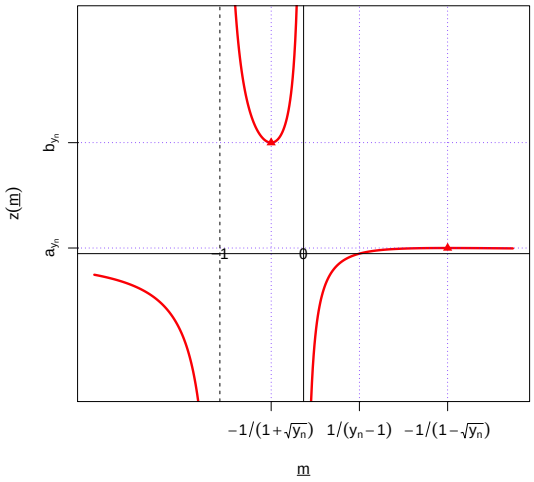

Figure 2: The graph of the transform $z(\underline{m}) = -\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}}$ when $y_n > 1$.

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 (2.8),

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

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 Figure 3.

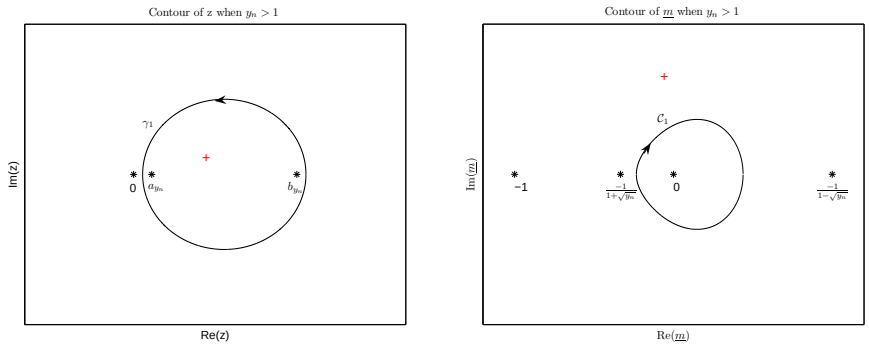

Figure 3: Contours of z and $\underline{m}$ when $y_n > 1$.

We have in place of (2.10),

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

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 Figure 4. On

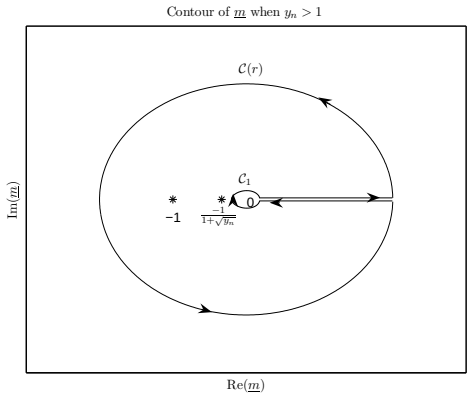

Figure 4: The new contour of $\underline{m}$ when $y_n > 1$

$\mathcal{C}(r)$ we have using the dominated convergence theorem

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

Therefore

$$G_{y_n}(f)=-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}}+\frac{y_n}{1+\underline{m}})\underline{m}(\frac{1}{\underline{m}^2}-\frac{y_n}{(1+\underline{m})^2})d\underline{m}. \tag{2.17}$$

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 $a_i>1+\sqrt{y_n}$. 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 as when there are no spikes. Therefore we get exactly (2.12) 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 $a_i>1+\sqrt{y_n}$: $-\frac{1}{a_i}$ $(i=1,\cdots,k_1)$, 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. □

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

In Bai et al. (2009), a corrected likelihood ratio statistic $\tilde{L}^*$ is proposed to test the hypothesis

$$H_0:\ \Sigma=I_p \quad vs. \quad H_1:\ \Sigma\neq I_p\ .$$

They prove that under H_0,

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

where

$$\begin{aligned}
&\tilde{L}^* = trS_n - \log|S_n| - p \ , \\
&G^{y_n,H_n}(g) = 1 - \frac{y_n - 1}{y_n}\log(1-y_n) \ , \\
&m(g) = -\frac{\log(1-y)}{2} \ , \\
&v(g) = -2\log(1-y) - 2y.
\end{aligned}$$

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

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

$$H_1^* : \ \Sigma_p \text{ has the spiked structure (1.1).}$$

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 1 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 $\tilde{L}^*$, the centering term $F^{y_n,H_n}(f)$ can be found to be

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

thanks to the following formulas

$$F^{y_n,H_n}(x) = 1 + \frac{1}{p}\sum_{i=1}^{k} n_i a_i - \frac{M}{p} + O(\frac{1}{n^2}) \tag{3.18}$$

and

$$F^{y_n,H_n}(\log x) = \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - 1 + (1 - \frac{1}{y_n})\log(1-y_n) + O(\frac{1}{n^2}) \ . \tag{3.19}$$

The details of derivation of these formulas are given in the Appendix A.

Therefore we have obtained that under H_1^*,

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

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

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

which is exactly the formula (5.6) found in Onatski et al. (2011). Note that these authors have found this formula using a sophisticated tools of asymptotic contiguity and Le Cam's first and third lemmas, our derivation is in a sense much more direct.

Appendix A: Additional proofs of (3.18) and (3.19)

The likelihood ratio test works only when $0 < y_n < 1$, and when $k_1 + 1 \leq i \leq k$, the corresponding a_i satisfy $|a_i - 1| \leq \sqrt{y_n}$, which is equivalent to $-\frac{1}{a_i} \in [\frac{-1}{1-\sqrt{y_n}}, \frac{-1}{1+\sqrt{y_n}}]$, so poles of $\{\underline{m} = -1\}$, $\{\underline{m} = -\frac{1}{a_i}, i = (k_1+1, \cdots, k)\}$ and $\{\underline{m} = \frac{1}{y_n - 1}\}$ (pole of the function $logz$) should be included in $\mathcal{C}_1$. In all the following, we write m to stand for $\underline{m}$ for convenience.

A.1. Proof of (3.18)

We have

$$(2.5) = -\frac{1}{2\pi i p}\oint_{\mathcal{C}_1}(-\frac{1}{m} + \frac{y_n}{1+m})(\frac{M}{y_n m} - \sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2})dm , \qquad (A.20)$$

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

$$\frac{M}{p} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{n_i a_i^2}{(1-a_i)^2} . \qquad (A.21)$$

$$(2.6) = \frac{1}{2\pi i p}\oint_{\mathcal{C}_1}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_i m)(1+m)}\left(\frac{1}{m} - \frac{y_n m}{(1+m)^2}\right)dm , \qquad (A.22)$$

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

$$\begin{aligned} & \frac{1}{p}\sum_{i=1}^{k}\left[- n_i - \frac{1}{2}(1-a_i)n_i y_n \frac{\partial}{\partial m^2}\Big(\frac{m}{1+a_i m}\Big)^2\Big|_{m=-1}\right] \\ = \quad & \frac{1}{p}\sum_{i=1}^{k}\left[- n_i + \frac{a_i n_i y_n}{(1-a_i)^2}\right] . \end{aligned} \qquad (A.23)$$

Besides, the residual of (A.20)+(A.22) at $m=-\frac{1}{a_i}$, $i=(k_1+1,\cdots,k)$ can be calculated as

$$\frac{1}{p}n_i\left(a_i+\frac{y_n a_i}{a_i-1}\right). \tag{A.24}$$

$$(2.7) \quad = \quad 1-\frac{M}{p}+\frac{1}{p}\sum_{i=1}^{k_1}n_i\left(a_i+\frac{y_n a_i}{a_i-1}\right)+O(\frac{1}{n^2}). \tag{A.25}$$

Combine (A.21), (A.23), (A.24) and (A.25), we get:

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

A.2. Proof of (3.19)

We first calculate (2.5) and (2.6) by considering their residuals at $m=-1$.

$$\begin{aligned}(2.5) \quad &= \quad \frac{-1}{2\pi i p y_n}\oint_{\mathcal{C}_1}\frac{\log(\frac{y_n-1}{m})+\log(\frac{m-\frac{1}{y_n-1}}{m+1})}{m}(M-\sum_{i=1}^{k}\frac{n_i a_i^2 y_n m^2}{(1+a_i m)^2})dm\\ &= \quad \frac{-M}{2\pi i p y_n}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\frac{1}{m}dm\\ &\quad +\frac{1}{2\pi i p y_n}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\sum_{i=1}^{k}\frac{n_i a_i^2 y_n m}{(1+a_i m)^2}dm\\ &\triangleq \quad A+B\,. \end{aligned} \tag{A.26}$$

$$\begin{aligned}A \quad &= \quad \frac{-M}{2\pi i p y_n}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\cdot d\log m\\ &= \quad \frac{M}{2\pi i p y_n}\oint_{\mathcal{C}_1}\log m\cdot d\log(\frac{m-\frac{1}{y_n-1}}{m+1})\\ &= \quad \frac{M}{2\pi i p y_n}\cdot\frac{y_n}{y_n-1}\oint_{\mathcal{C}_1}\frac{\log m}{(m+1)(m-\frac{1}{y_n-1})}dm\\ &= \quad -\frac{M}{p y_n}\log(1-y_n)\,, \end{aligned} \tag{A.27}$$

$$\begin{aligned}B \quad &= \quad \frac{1}{2\pi i p}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2}dm\\ &= \quad \frac{1}{2\pi i p}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})n_i a_i(\frac{1}{1+a_i m}-\frac{1}{(1+a_i m)^2})dm\\ &\triangleq \quad C-D\,, \end{aligned} \tag{A.28}$$

where

$$\begin{aligned} C &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\frac{n_i a_i}{1+a_i m}dm \\ &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1} n_i\log(\frac{m-\frac{1}{y_n-1}}{m+1})\cdot d\log(1+a_i m) \\ &= \frac{-1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1} n_i\log(1+a_i m)\cdot d\log(\frac{m-\frac{1}{y_n-1}}{m+1}) \\ &= \frac{-1}{2\pi ip}\cdot\frac{y_n}{y_n-1}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\frac{n_i\log(1+a_i m)}{(m+1)(m-\frac{1}{y_n-1})}dm \\ &= \frac{1}{p}\sum_{i=1}^{k} n_i\log(1-a_i)-\frac{1}{p}\sum_{i=1}^{k} n_i\log(1+\frac{a_i}{y_n-1})\ , \end{aligned} \tag{A.29}$$

and

$$\begin{aligned} D &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\frac{n_i a_i}{(1+a_i m)^2}dm \\ &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\frac{n_i}{1+a_i m}\cdot d\log(\frac{m-\frac{1}{y_n-1}}{m+1}) \\ &= \frac{y_n}{2\pi ip(y_n-1)}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\frac{n_i}{(1+a_i m)(m-\frac{1}{y_n-1})(m+1)}dm \\ &= \frac{1}{p}\sum_{i=1}^{k}(\frac{n_i}{1+\frac{a_i}{y_n-1}}-\frac{n_i}{1-a_i})\ . \end{aligned} \tag{A.30}$$

Combine (A.26), (A.27), (A.28), (A.29) and (A.30), we get the residual of (2.5) at $m=-1$:

$$\begin{aligned} &-\frac{M}{py_n}\log(1-y_n)+\frac{1}{p}\sum_{i=1}^{k} n_i\log(1-a_i)-\frac{1}{p}\sum_{i=1}^{k} n_i\log(1+\frac{a_i}{y_n-1}) \\ &-\frac{1}{p}\sum_{i=1}^{k}\frac{n_i}{1+\frac{a_i}{y_n-1}}+\frac{1}{p}\sum_{i=1}^{k}\frac{n_i}{1-a_i}\ . \end{aligned} \tag{A.31}$$

Then, we consider the part (2.6) in the general formula influenced by the pole $m=-1$:

$$\begin{aligned}(2.6) &= -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f^{'}(-\frac{1}{m}+\frac{y_n}{1+m})\sum_{i=1}^{k}(\frac{n_i a_i}{1+a_i m}-\frac{n_i}{1+m})(\frac{1}{m}-\frac{y_n m}{(1+m)^2})dm \\ &= -\frac{1}{2\pi ip}\sum_{i=1}^{k} n_i \oint_{\mathcal{C}_1}\frac{m(m+1)}{y_n m-m-1}(\frac{a_i}{1+a_i m}-\frac{1}{1+m})(\frac{1}{m}-\frac{y_n m}{(1+m)^2})dm \\ &\triangleq \frac{-1}{2\pi ip(y_n-1)}\sum_{i=1}^{k} n_i(E-F-G+H)\ ,\end{aligned}$$

where

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

Collecting these four terms, we have the residual of (2.6) at $m=-1$:

$$\frac{1}{p}\sum_{i=1}^{k}(\frac{1}{a_i-1}-\frac{a_i}{y_n+a_i-1})n_i\ . \tag{A.32}$$

Then we consider the influence of (2.5)+(2.6) caused by the pole $m=-\frac{1}{a_i}$, $i=k_1+1,\cdots,k$, which can be calculated similarly as

$$\frac{n_i}{p}\log(a_i+\frac{y_n a_i}{a_i-1})\ . \tag{A.33}$$

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 Bai and Silverstein (2004), and combine (A.31), (A.32), (A.33) and (2.7), we get

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

References

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

Bai, Z.D. and Yao, J.F. (2008). CLT for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.*, **44**(3), 447–474.

Bai, Z.D., Jiang, D.D., Yao, J.F. and Zheng, S.R. (2009). Corrections to LRT on large dimensional covariance matrix by RMT. *Ann. Statist.* **37**, 3822-3840.

Bai, Z.D. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices* (2nd edition). Springer, 20.

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

Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**(5), 1643–1697.

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

Benaych-Georges, F., Guionnet, A. and Maida, M. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.*, **16**, 1621–1662.

Benaych-Georges, F. and Nadakuditi, R.R. (2011). The eigenvalues and eigenvectors of finite low rank perturbations of large random matrices. *Adv. Math.*, **227**(2), 494–521.

Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**(2), 295–327.

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

Kritchman, S. and Nadler, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process.* **57**(10), 3930–3941.

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

Nadakuditi, R.R. and Silverstein, J.W. (2010). Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *IEEE J. Sel. Topics Signal Processing.* **4**(3), 468-480.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, **77**(5), 1447-1479.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, **92**(4), 1004-1016.

Onatski, A., Moreira, M.J. and Hallin, M. (2011). Asymptotic power of sphericity tests for high-dimensional data. *Preprint*, available at `arXiv:1210.5663v1`.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics*, **168**, 244-258.

Pan, G.M. and Zhou, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.*, **18**, 1232-1270.

Passemier, D. and Yao, J.F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrix: Theory and Applciations*, **1**, 1150002

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

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

Silverstein, J.W. and Choi, S.I. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.*, **54**(2), 295–309.