Journal of Statistical Planning and Inference 142 (2012) 2241–2256

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Likelihood ratio tests for covariance matrices of high-dimensional normal distributions

Dandan Jiang [a,1], Tiefeng Jiang [b,2,*], Fan Yang [b,c]

[a] School of Mathematics, Jilin University, Changchun 130012, China
[b] School of Statistics, University of Minnesota, 224 Church Street, Minneapolis, MN 55455, United States
[c] Boston Scientific, 1 Scimed Place, Maple Grove, MN 55311, United States

ARTICLE INFO

Article history:
Received 27 August 2011
Received in revised form
27 February 2012
Accepted 28 February 2012
Available online 7 March 2012

Keywords:
High-dimensional data
Testing on covariance matrices
Selberg integral
Gamma function

ABSTRACT

For a random sample of size n obtained from a p-variate normal population, the likelihood ratio test (LRT) for the covariance matrix equal to a given matrix is considered. By using the Selberg integral, we prove that the LRT statistic converges to a normal distribution under the assumption $p/n \to y \in (0,1]$. The result for $y=1$ is much different from the case for $y \in (0,1)$. Another test is studied: given two sets of random observations of sample size n_1 and n_2 from two p-variate normal distributions, we study the LRT for testing the two normal distributions having equal covariance matrices. It is shown through a corollary of the Selberg integral that the LRT statistic has an asymptotic normal distribution under the assumption $p/n_1 \to y_1 \in (0,1]$ and $p/n_2 \to y_2 \in (0,1]$. The case for $\max\{y_1,y_2\}=1$ is much different from the case $\max\{y_1,y_2\}<1$.

1. Introduction

In their pioneer work, Bai et al. (2009) studied two Likelihood Ratio Tests (LRTs) by using Random Matrix Theory. The limiting distributions of the LRT test statistics are derived. There are two purposes in this paper. We first use the Selberg integral, a different method, to revisit the two problems. We then prove two theorems which cover the critical cases that are not studied in Bai et al. (2009). Now we review the two tests and present our results.

Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be i.i.d. $\mathbb{R}^p$-valued random variables with normal distribution $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. Consider the test

$$H_0 : \boldsymbol{\Sigma} = \mathbf{I}_p \quad \text{vs} \quad H_a : \boldsymbol{\Sigma} \neq \mathbf{I}_p, \tag{1.1}$$

with $\boldsymbol{\mu}$ unspecified. Any test $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ with known non-singular $\boldsymbol{\Sigma}_0$ and unspecified $\boldsymbol{\mu}$ can be reduced to (1.1) by transforming data $\mathbf{y}_i = \boldsymbol{\Sigma}_0^{-1/2}\mathbf{x}_i$ for $i=1,2,\ldots,n$ (then $\mathbf{y}_1,\ldots,\mathbf{y}_n$ are i.i.d. with distribution $N_p(\tilde{\boldsymbol{\mu}},\mathbf{I}_p)$, where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\mu}$). Recall

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \text{and} \quad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^*. \tag{1.2}$$

* Corresponding author.
E-mail addresses: jiangdandan@jlu.edu.cn (D. Jiang), jiang040@umn.edu (T. Jiang), yang0712@umn.edu (F. Yang).
[1] Supported in part by NSFC 11101181 and RFDP 20110061120005.
[2] Supported in part by NSF #DMS-0449365.

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2012.02.057

Of course $\boldsymbol{S}$ is a $p\times p$ matrix. After scaling and taking logarithm, a LRT statistic for (1.1) is chosen to be in the following form:

$$L_n^* = \mathrm{tr}(\mathbf{S}) - \log|\mathbf{S}| - p = \frac{1}{n}\sum_{i=1}^{p}(\lambda_i - n\log\lambda_i) + p\log n - p, \tag{1.3}$$

where $\lambda_1,\ldots,\lambda_p$ are the eigenvalues of $n\mathbf{S}$. See, for example, p. 355 from Muirhead (1982) for this. The notation log above stands for the natural logarithm $\log_e$ throughout the paper.

For fixed p, it is known from the classical multivariate analysis theory that a (constant) linear transform of nL_n^* converges to $\chi^2_{p(p+1)/2}$ as $n\to\infty$. See, e.g., p. 359 from Muirhead (1982). When p is large, particularly as $n\to\infty$ and $p/n\to y\in(0,1)$, there are some results on the improvement of the convergence, see, e.g., Bai and Saranadasa (1996). The fact that dimension p is large and is proportional to the sample size n is a common practice in modern data. A failure for a similar LRT test in the high dimensional case (p is large) is observed by Dempster (1958) in as early as 1958. It is due to this reason that Bai et al. (2009) study the statistic L_n^* in (1.3) when both n and p are large and are proportional to each other.

Now, we state our results in this paper next.

Theorem 1. *Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be i.i.d. random vectors with normal distribution $\mathrm{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Let L_n^* be as in (1.3). Assume H_0 in (1.1) holds. If $n>p=p_n$ and $\lim_{n\to\infty}p/n=y\in(0,1]$, then $(L_n^*-\mu_n)/\sigma_n$ converges in distribution to $\mathrm{N}(0,1)$ as $n\to\infty$, where*

$$\mu_n = \left(n-p-\frac{3}{2}\right)\log\left(1-\frac{p}{n}\right)+p-y \quad \text{and} \quad \sigma_n^2 = -2\left[\frac{p}{n}+\log\left(1-\frac{p}{n}\right)\right].$$

A simulation study was made for the quantity $(L_n^*-\mu_n)/\sigma_n$ as in Theorem 1. We chose $p/n=0.9$ in Fig. 1 with different values of n. The figure shows that the convergence becomes more accurate as n increases. To see the convergence rate for the case $y=1$, we chose an extreme scenario with $p=n-4$ in Fig. 2. As n increases, the convergence rate seems quite decent too.

Now, note that $\sigma_n^2\to -2y-2\log(1-y)$ if $p/n\to y\in(0,1)$. We obviously have the following corollary.

Corollary 1.1. *Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be i.i.d. random vectors with normal distribution $\mathrm{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Let L_n^* be as in (1.3). Assume H_0 in (1.1) holds. If $n>p=p_n$ and $\lim_{n\to\infty}p/n=y\in(0,1)$, then $L_n^*-\mu_n$ converges in distribution to $\mathrm{N}(0,\sigma^2)$ as $n\to\infty$, where $\sigma^2=-2y-2\log(1-y)$ and*

$$\mu_n = (n-p)\log\left(1-\frac{p}{n}\right)+p-y-\frac{3}{2}\log(1-y).$$

Looking at Theorem 1, it is obvious that $\sigma_n^2\sim -2\log(1-(p/n))$ as $p/n\to 1$. We then get the following.

Corollary 1.2. *Assume all the conditions in Theorem 1 hold with $y=1$. Let $r_n=(-\log(1-(p/n)))^{1/2}$. Then*

$$\frac{L_n^*-p-(p-n+1.5)r_n^2}{\sqrt{2}r_n} \text{ converges in distribution to } \mathrm{N}(0,1) \text{ as } n\to\infty.$$

The above result studies the critical case for $y=1$, which is not covered in Bai et al. (2009). In fact, the random matrix tool by Bai and Silverstein (2004) is used to derive the results in Bai et al. (2009). Their tool fails when $y=1$.

For a practical testing procedure, we would use Theorem 1 directly instead of using Corollaries 1.1 and 1.2, which deal with the cases $y\in(0,1)$ and $y=1$ separately. This is because, for a real set of data, sometimes it is hard to judge when p/n goes to 1 or when it goes to a number less than 1.

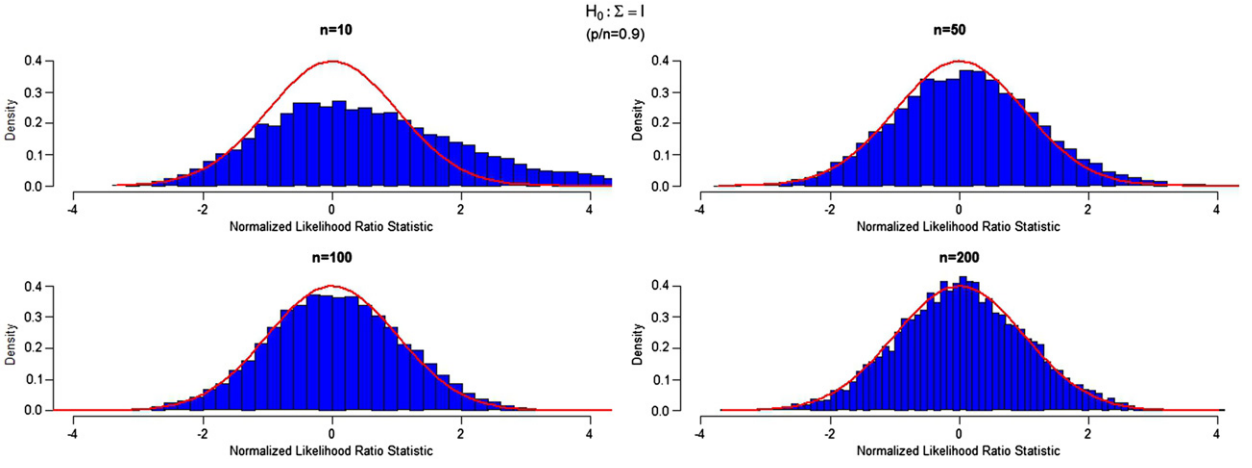

Fig. 1. Histograms were constructed based on 10,000 simulations of the normalized likelihood ratio statistic $(L_n^*-\mu_n)/\sigma_n$ according to Theorem 1 under the null hypothesis $\Sigma=\mathbf{I}_p$ with $p/n=0.9$. The curves on the top of the histograms are the standard normal curve.

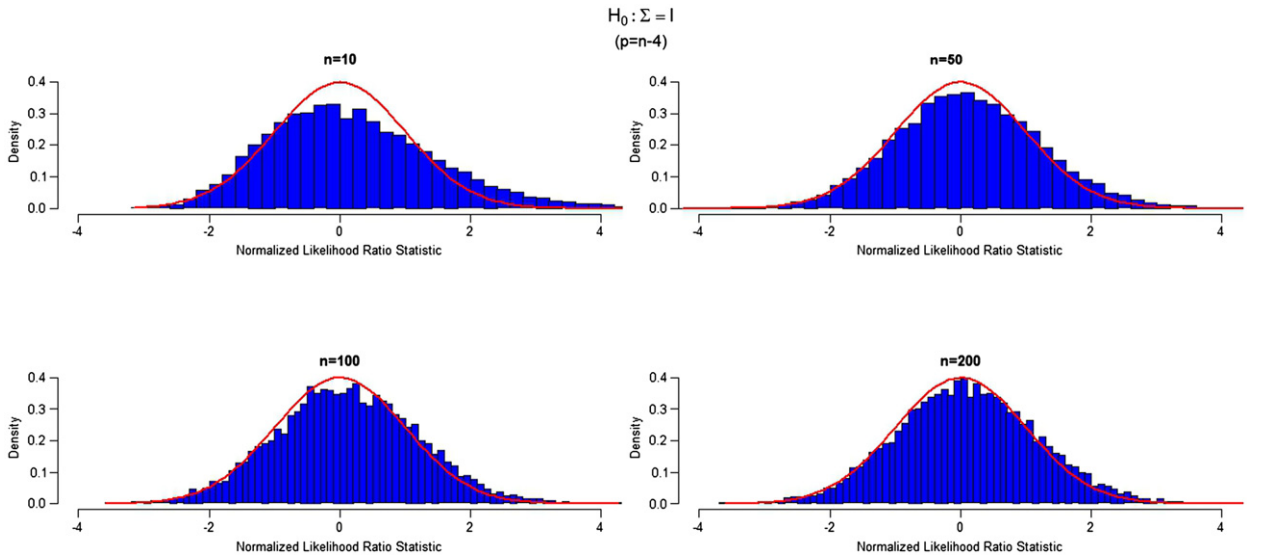

Fig. 2. Histograms were constructed based on 10,000 simulations of the normalized likelihood ratio statistic $(L_n^*-\mu_n)/\sigma_n$ according to Theorem 1 under the null hypothesis $\Sigma=\mathbf{I}_p$ with $p=n-4$. The curves on the top of the histograms are the standard normal curve.

Now we study another likelihood test. For two p-dimensional normal distributions $N(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, $k=1,2$, where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are non-singular and unknown, we wish to test

$$H_0:\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2 \quad \text{vs} \quad H_a:\boldsymbol{\Sigma}_1\neq\boldsymbol{\Sigma}_2, \tag{1.4}$$

with unspecified $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. The data are given as follows: $\mathbf{x}_1,\ldots,\mathbf{x}_{n_1}$ is a random sample from $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$; $\mathbf{y}_1,\ldots,\mathbf{y}_{n_2}$ is a random sample from $N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, and two sets of random vectors are independent. The two relevant covariance matrices are

$$\mathbf{A}=\frac{1}{n_1}\sum_{i=1}^{n_1}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^* \quad \text{and} \quad \mathbf{B}=\frac{1}{n_2}\sum_{i=1}^{n_2}(\mathbf{y}_i-\overline{\mathbf{y}})(\mathbf{y}_i-\overline{\mathbf{y}})^*, \tag{1.5}$$

where

$$\overline{\mathbf{x}}=\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_i \quad \text{and} \quad \overline{\mathbf{y}}=\frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{y}_i. \tag{1.6}$$

Let $N=n_1+n_2$ and $c_k=n_k/N$ for $k=1,2$. The likelihood ratio test statistic is

$$T_N=-2\log L_1 \quad \text{where } L_1=\frac{|\mathbf{A}|^{n_1/2}\cdot|\mathbf{B}|^{n_2/2}}{|c_1\mathbf{A}+c_2\mathbf{B}|^{N/2}}. \tag{1.7}$$

See, e.g., Section 8.2 from Muirhead (1982) for this. The second main result in this paper is as follows.

Theorem 2. *Let $n_i>p$ for $i=1,2$ and T_N be as in (1.7). Assume H_0 in (1.4) holds. If $n_1\to\infty, n_2\to\infty$ and $p\to\infty$ with $p/n_i\to y_i\in(0,1]$ for $i=1,2$, then*

$$\frac{1}{\sigma_n}\left(\frac{T_N}{N}-\mu_n\right) \textit{ converges in distribution to } N(0,1),$$

where

$$\mu_n=(p-N+2.5)\log\left(1-\frac{p}{N}\right)-\sum_{i=1}^{2}\frac{(p-n_i+1.5)n_i}{N}\log\left(1-\frac{p}{n_i}\right);$$

$$\sigma_n^2=2\log\left(1-\frac{p}{N}\right)-2\sum_{i=1}^{2}\frac{n_i^2}{N^2}\log\left(1-\frac{p}{n_i}\right). \tag{1.8}$$

We did some simulations for the statistic $(T_N/N-\mu_n)/\sigma_n$ as in Theorem 2. In Fig. 3, we chose $p/n_1=p/n_2=0.9$, the picture shows that the convergence rate is quite robust with the value of n_1,n_2 and p increases even though the ratio 0.9 is close to 1. To see the convergence rate for the case that $\max\{y_1,y_2\}=1$, we chose an extreme situation with $p=n_1-4=n_2-4$ in Fig. 4. The convergence rate looks well too although it is not as fast as the case $p/n_1=p/n_2=0.9$ presents.

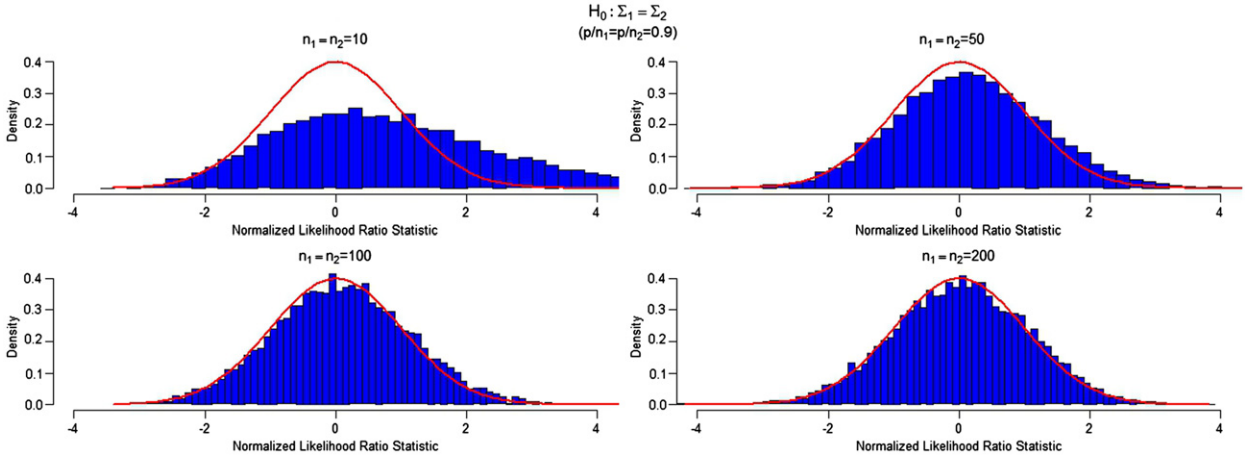

Fig. 3. Histograms were constructed based on 10,000 simulations of the normalized likelihood ratio statistic $(T_N/N-\mu_n)/\sigma_n$ according to Theorem 2 under the null hypothesis $\Sigma_1=\Sigma_2$ with $p/n_1=p/n_2=0.9$. The curves on the top of the histograms are the standard normal curve.

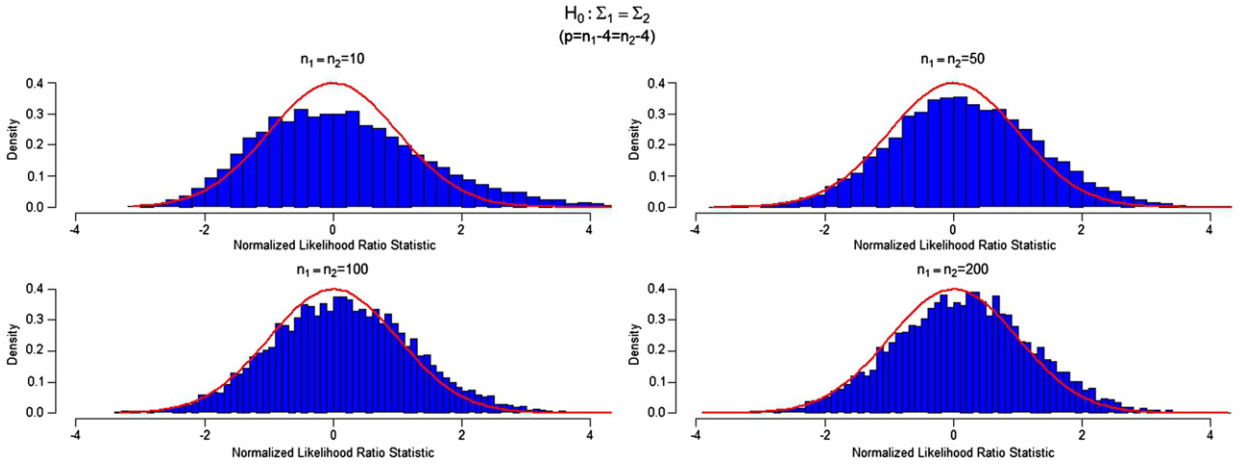

Fig. 4. Histograms were constructed based on 10,000 simulations of the normalized likelihood ratio statistic $(T_N/N-\mu_n)/\sigma_n$ according to Theorem 2 under the null hypothesis $\Sigma_1=\Sigma_2$ with $p=n_1-4=n_2-4$. The curves on the top of the histograms are the standard normal curve.

According to the notation in Theorem 2, we know that $p/N=((n_1/p)+(n_2/p))^{-1}\to y_1y_2/(y_1+y_2)$ and $n_i/N=n_i/p\cdot((n_1/p)+(n_2/p))^{-1}\to y_i^{-1}/(y_1^{-1}+y_2^{-1})$ for $i=1, 2$. We easily get the following corollary.

Corollary 1.3. *Let $n_i>p$ for $i=1,2$ and T_N be as in (1.7). Assume H_0 in (1.4) holds. If $n_1\to\infty, n_2\to\infty$ and $p\to\infty$ with $p/n_i\to y_i\in(0,1)$ for $i=1,2$, then*

$$\frac{T_N}{N}-v_n \text{ converges in distribution to } \mathrm{N}(\mu,\sigma^2),$$

where

$$\mu=\tfrac{1}{2}[5\log(1-y)-3\gamma_1\log(1-y_1)-3\gamma_2\log(1-y_2)];$$

$$\sigma^2=2[\log(1-y)-\gamma_1^2\log(1-y_1)-\gamma_2^2\log(1-y_2)];$$

$$v_n=(p-N)\log\left(1-\frac{p}{N}\right)-\frac{(p-n_1)n_1}{N}\log\left(1-\frac{p}{n_1}\right)-\frac{(p-n_2)n_2}{N}\log\left(1-\frac{p}{n_2}\right), \tag{1.9}$$

with $\gamma_1=y_2(y_1+y_2)^{-1}$, $\gamma_2=y_1(y_1+y_2)^{-1}$ and $y=y_1y_2(y_1+y_2)^{-1}$.

Our method of proving the above results is much different from Bai et al. (2009). The random matrix theories, developed by Bai and Silverstein (2004) for the Wishart matrices and Zheng (2008) for the *F*-matrices, are used in Bai et al. (2009). The tools are universal in the sense that no normality assumption is needed. However, the requirements that $y<1$

as in Corollary 1.1 and $\max\{y_1,y_2\}<1$ as in Corollary 1.3 are crucial. Technically, the study for critical cases that $y=1$ and that $\max\{y_1,y_2\}=1$ are more challenging.

Under the normality assumption, without relying on the random matrix theories similar to Bai and Silverstein (2004) and Zheng (2008), we are able to use analysis tools. In fact, the Selberg integral is used in the proof of both theorems. Through the Selberg integral, some close forms of the moment generating functions of the two likelihood ratio test statistics are obtained. We then study the moment generating functions to derive the central limit theorems for the two likelihood ratio test statistics. In particular, our results study the cases that $y\le 1$ and that $\max\{y_1,y_2\}\le 1$. As shown in Corollary 1.2, the result for $y=1$ and the result for $y\in(0,1)$ are much different. The same applies for the second test.

We develop a tool on the product of a series of Gamma functions (Proposition 2.1). It is powerful in analyzing the moment generating functions of the two log-likelihood ratio statistics studied in this paper.

The organization of the rest of the paper is as follows. In Section 2, we derive a tool to study the product of a series of the Gamma functions. The proofs of the main theorems stated above are given in Section 3.

2. Auxiliary results

Proposition 2.1. *Let $n>p=p_n$ and $r_n=(-\log(1-(p/n)))^{1/2}$. Assume that $p/n\to y\in(0,1]$ and $t=t_n=\mathrm{O}(1/r_n)$ as $n\to\infty$. Then, as $n\to\infty$,*

$$\log\prod_{i=n-p}^{n-1}\frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}=pt(1+\log 2)-pt\log n+r_n^2(t^2+(p-n+1.5)t)+\mathrm{o}(1).$$

The proposition is proved through the following three lemmas.

Lemma 2.1. *Let $b:=b(x)$ be a real-valued and bounded function defined on $(0,\infty)$. Then*

$$\log\frac{\Gamma(x+b)}{\Gamma(x)}=b\log x+\frac{b^2-b}{2x}+\mathrm{O}\left(\frac{1}{x^2}\right),$$

as $x\to+\infty$, where $\Gamma(x)$ is the gamma function.

Proof. Recall the Stirling formula (see, e.g., p. 368 from Gamelin, 2001 or (37) on p. 204 from Ahlfors, 1979):

$$\log\Gamma(z)=z\log z-z-\frac{1}{2}\log z+\log\sqrt{2\pi}+\frac{1}{12z}+\mathrm{O}\left(\frac{1}{x^3}\right),$$

as $x=\mathrm{Re}(z)\to+\infty$. It follows that

$$\log\frac{\Gamma(x+b)}{\Gamma(x)}=(x+b)\log(x+b)-x\log x-b-\frac{1}{2}(\log(x+b)-\log x)+\frac{1}{12}\left(\frac{1}{x+b}-\frac{1}{x}\right)+\mathrm{O}\left(\frac{1}{x^3}\right),\tag{2.1}$$

as $x\to+\infty$. First, use the fact that $\log(1+t)\sim t-(t^2/2)+\mathrm{O}(t^3)$ as $t\to 0$ to get

$$\begin{aligned}(x+b)\log(x+b)-x\log x&=(x+b)\left(\log x+\log\left(1+\frac{b}{x}\right)\right)-x\log x\\&=(x+b)\left(\log x+\frac{b}{x}-\frac{b^2}{2x^2}+\mathrm{O}(x^{-3})\right)-x\log x\\&=b\log x+b+\frac{b^2}{2x}+\mathrm{O}\left(\frac{1}{x^2}\right),\end{aligned}$$

as $x\to+\infty$. Evidently,

$$\log(x+b)-\log x=\log\left(1+\frac{b}{x}\right)=\frac{b}{x}+\mathrm{O}\left(\frac{1}{x^2}\right)\quad\text{and}\quad\frac{1}{x+b}-\frac{1}{x}=\mathrm{O}\left(\frac{1}{x^2}\right),$$

as $x\to+\infty$. Plugging these two assertions into (2.1), we have

$$\log\frac{\Gamma(x+b)}{\Gamma(x)}=b\log x+\frac{b^2-b}{2x}+\mathrm{O}\left(\frac{1}{x^2}\right),$$

as $x\to+\infty$. □

Lemma 2.2. *Let $n>p=p_n$. Assume that $\lim_{n\to\infty}p/n=y\in(0,1)$ and $\{t_n;n\geq 1\}$ is bounded. Then, as $n\to\infty$,*

$$\log\prod_{i=n-p}^{n-1}\frac{\Gamma\left(\frac{i}{2}-t_n\right)}{\Gamma\left(\frac{i}{2}\right)}=pt_n(1+\log 2)-t_n n\log n+t_n(n-p)\log(n-p)-\left(t_n^2+\frac{3t_n}{2}\right)\log(1-y)+o(1). \tag{2.2}$$

Proof. Since $p/n\to y\in(0,1)$, then $n-p\to+\infty$ as $n\to\infty$. By Lemma 2.1, there exists integer $C_1\geq 2$ such that

$$\log\frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}=-t\log\frac{i}{2}+\frac{t^2+t}{i}+\varphi(i)\quad\text{and}\quad|\varphi(i)|\leq\frac{C_1}{i^2},$$

for all $i\geq n-p$ as n is sufficiently large, where here and later in this proof we write t for t_n for short notation. Notice $-t\log(i/2)=t\log 2-t\log i$. Then,

$$\begin{aligned}\sum_{i=n-p}^{n-1}\log\frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}&=pt\log 2-t\sum_{i=n-p}^{n-1}\log i+(t^2+t)\sum_{i=n-p}^{n-1}\frac{1}{i}+\sum_{i=n-p}^{n-1}\varphi(i)\\&=pt\log 2+(t^2+t)\sum_{i=n-p}^{n-1}\frac{1}{i}-t\log\frac{n!}{(n-p)!}+t\log\frac{n}{(n-p)}+O\left(\frac{1}{n}\right)\\&=pt\log 2+(t^2+t)\sum_{i=n-p}^{n-1}\frac{1}{i}-t\log(1-y)-t\log\frac{n!}{(n-p)!}+o(1),\end{aligned} \tag{2.3}$$

since $\sum_{i=n-p}^{n-1}\varphi(i)=O(1/n)$ and $\log(n/(n-p))\to-\log(1-y)$ as $n\to\infty$. First,

$$\sum_{i=n-p}^{n-1}\frac{1}{i}\leq\sum_{i=n-p}^{n-1}\int_{i-1}^{i}\frac{1}{x}\,dx=\int_{n-p-1}^{n-1}\frac{1}{x}\,dx.$$

By working on the lower bound similarly, we have

$$\log\frac{n}{n-p}=\int_{n-p}^{n}\frac{1}{x}\,dx\leq\sum_{i=n-p}^{n-1}\frac{1}{i}\leq\int_{n-p-1}^{n-1}\frac{1}{x}\,dx=\log\frac{n-1}{n-p-1}.$$

This implies, by assumption $p/n\to y$, that

$$\sum_{i=n-p}^{n-1}\frac{1}{i}\to-\log(1-y), \tag{2.4}$$

as $n\to\infty$. Second, by the Stirling formula (see, e.g., p. 210 from Freitag and Busam, 2005), there are some $\theta_n,\theta_n'\in(0,1)$,

$$\begin{aligned}\log\frac{n!}{(n-p)!}&=\log\frac{\sqrt{2\pi n}n^n e^{-n+(\theta_n/12n)}}{\sqrt{2\pi(n-p)}(n-p)^{n-p}e^{-n+p+(\theta_n'/12(n-p))}}=n\log n-(n-p)\log(n-p)-p+\frac{1}{2}\log\frac{n}{n-p}+o(1)\\&=n\log n-(n-p)\log(n-p)-p-\frac{1}{2}\log(1-y)+o(1),\end{aligned}$$

as $n\to\infty$. Join this with (2.3) and (2.4), we arrive at

$$\begin{aligned}\log\prod_{i=n-p}^{n-1}\frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}&=pt\log 2-(t^2+t)\log(1-y)-t\log(1-y)-tn\log n+t(n-p)\log(n-p)+tp+\frac{t}{2}\log(1-y)+o(1)\\&=pt(1+\log 2)-\left(t^2+\frac{3t}{2}\right)\log(1-y)-tn\log n+t(n-p)\log(n-p)+o(1),\end{aligned}$$

as $n\to\infty$. The proof is then completed. □

Lemma 2.3. *Let $n>p=p_n$ and $r_n=(-\log(1-(p/n)))^{1/2}$. Assume that $\lim_{n\to\infty}p/n=1$ and $t=t_n=O(1/r_n)$ as $n\to\infty$. Then, as $n\to\infty$,*

$$\log\prod_{i=n-p}^{n-1}\frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}=pt(1+\log 2)-pt\log n+r_n^2(t^2+(p-n+1.5)t)+o(1).$$

Proof. Obviously, $\lim_{n\to\infty} r_n = +\infty$. Hence, $\{t_n; n \geq 2\}$ is bounded. By Lemma 2.1, there exist integers $C_1 \geq 2$ and $C_2 \geq 2$ such that

$$\log \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)} = -t \log \frac{i}{2} + \frac{t^2+t}{i} + \varphi(i) \quad \text{and} \quad |\varphi(i)| \leq \frac{C_1}{i^2}, \tag{2.5}$$

for all $i \geq C_2$.

We will use (2.5) to estimate $\prod_{i=n-p}^{n-1} \Gamma(i/2-t)/\Gamma(i/2)$. However, when $n-p$ is small, say, 2 or 3 (which is possible since $p/n \to 1$), the identity (2.5) cannot be directly applied to estimate each term in the product of $\prod_{i=n-p}^{n-1} \Gamma(i/2-t)/\Gamma(i/2)$. We next use a truncation to solve the problem thanks to the fact that $\Gamma(i/2-t)/\Gamma(i/2) \to 1$ as $n \to \infty$ for fixed i.

Fix $M \geq C_2$. Write

$$a_i = \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)} \quad \text{for } i \geq 1 \quad \text{and} \quad \gamma_n = \begin{cases} 1 & \text{if } n-p \geq M; \\ \prod_{i=n-p}^{M-1} a_i & \text{if } n-p < M. \end{cases}$$

Then,

$$\prod_{i=n-p}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)} = \gamma_n \cdot \prod_{i=(n-p)\vee M}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}. \tag{2.6}$$

Easily,

$$\left(\min_{1\leq i\leq M}(1\wedge a_i)\right)^M \leq \gamma_n \leq \left(\max_{1\leq i\leq M}(1\vee a_i)\right)^M,$$

for all $n \geq 1$. Note that, for each $i \geq 1$, $a_i \to 1$ as $n \to \infty$ since $\lim_{n\to\infty} t_n = 0$. Thus, since M is fixed, the two bounds above go to 1 as $n \to \infty$. Consequently, $\lim_{n\to\infty} \gamma_n = 1$. This and (2.6) say that

$$\prod_{i=n-p}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)} \sim \prod_{i=(n-p)\vee M}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}, \tag{2.7}$$

as $n \to \infty$. By (2.5), as n is sufficiently large, we know

$$\log \prod_{i=(n-p)\vee M}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)} = \sum_{i=(n-p)\vee M}^{n-1} \left(-t \log \frac{i}{2} + \frac{t^2+t}{i} + \varphi(i)\right),$$

with $|\varphi(i)| \leq C_1 i^{-2}$ for $i \geq C_2$. Write $-t \log(i/2) = -t \log i + t \log 2$. It follows that

$$\log \prod_{i=(n-p)\vee M}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)} = (n-(n-p)\vee M)t \log 2 - t \sum_{i=(n-p)\vee M}^{n-1} \log i + (t^2+t) \sum_{i=(n-p)\vee M}^{n-1} \frac{1}{i} + \sum_{i=(n-p)\vee M}^{n-1} \varphi(i) := A_n - B_n + C_n + D_n \tag{2.8}$$

as n is sufficiently large. Now we analyze the four terms above.

By distinguishing the cases $n-p > M$ and $n-p \leq M$, we get

$$|A_n - pt \log 2| \leq (t \log 2) \cdot |n-p-M| \cdot I(n-p \leq M) \leq (M \log 2)t. \tag{2.9}$$

Now we estimate B_n. By the same argument as in (2.9), we get

$$\left| \sum_{i=(n-p)\vee M}^{n-1} h(i) - \sum_{i=(n-p)}^{n-1} h(i) \right| \leq \sum_{i=1}^{M} |h(i)| \tag{2.10}$$

for $h(x) = \log x$ or $h(x) = 1/x$ on $x \in (0,\infty)$. By the Stirling formula (see, e.g., Freitag and Busam, 2005, p. 210), $n! = \sqrt{2\pi n} n^n e^{-n+\theta_n/12n}$ with $\theta_n \in (0,1)$ for all $n \geq 1$. It follows that for some $\theta_n, \theta_n' \in (0,1)$,

$$\sum_{i=n-p}^{n-1} \log i = \log \frac{n!}{(n-p)!} + \log \frac{n-p}{n} = \log \frac{\sqrt{2\pi n} n^n e^{-n+(\theta_n/12n)}}{\sqrt{2\pi(n-p)}(n-p)^{n-p} e^{-n+p+(\theta_n'/12(n-p))}}$$

$$+\log\frac{n-p}{n}=n\log n-(n-p)\log(n-p)-p+\frac{1}{2}\log\frac{n-p}{n}+R_n,$$

with $|R_n|\le 1$ as n is sufficiently large. Recall $B_n=t\sum_{i=(n-p)\vee M}^{n-1}\log i$. We know from (2.10) that

$$\left|B_n-\left(tn\log n-t(n-p)\log(n-p)-tp+\frac{t}{2}\log\frac{n-p}{n}\right)\right|\le Ct, \tag{2.11}$$

where C here and later stands for a constant and can be different from line to line.

Now we estimate C_n. Recall the identity $s_n:=\sum_{i=1}^{n}(1/i)=\log n+c_n$ for all $n\ge 1$ and $\lim_{n\to\infty}c_n=c$, where $c\sim 0.577$ is the Euler constant. Thus, $|(s_n-s_{n-p})-\log(n/(n-p))|\le c_n+c_{n-p}$. Moreover,

$$\sum_{i=n-p+1}^{n}\frac{1}{i}=s_n-s_{n-p}\quad\text{and}\quad\left|\sum_{i=n-p}^{n-1}\frac{1}{i}-\sum_{i=n-p+1}^{n}\frac{1}{i}\right|\le 1.$$

Therefore,

$$\left|\sum_{i=n-p}^{n-1}\frac{1}{i}-\log\frac{n}{n-p}\right|\le C.$$

Consequently, since $C_n=(t^2+t)\sum_{i=(n-p)\vee M}^{n-1}(1/i)$, we know from (2.10) that

$$\left|C_n-(t^2+t)\log\frac{n}{n-p}\right|\le(t^2+t)C. \tag{2.12}$$

Finally, it is easy to see from the second fact in (2.5) that

$$|D_n|\le C_1\sum_{i=M}^{\infty}\frac{1}{i^2}, \tag{2.13}$$

for all $n\ge 2$. Now, reviewing that $t=t_n\to 0$ as $n\to\infty$, we have from (2.7)–(2.9), (2.11) and (2.12) that, for fixed integer $M>0$,

$$\begin{aligned}A_n-B_n+C_n+D_n&=pt\log 2-\left(tn\log n-t(n-p)\log(n-p)-tp+\frac{t}{2}\log\frac{n-p}{n}\right)+(t^2+t)\log\frac{n}{n-p}+D_n+o(1)\\&=\underbrace{pt(1+\log 2)+\left(t^2+\frac{3t}{2}-nt\right)\log n-\left(t^2+\frac{3t}{2}-(n-p)t\right)\log(n-p)}_{E_n}+D_n+o(1),\end{aligned}$$

as $n\to\infty$. Write $\log(n-p)=\log n-r_n^2$. Then

$$E_n=pt(1+\log 2)-pt\log n+r_n^2\left(t^2+\frac{3t}{2}-(n-p)t\right).$$

From (2.13) we have that

$$\limsup_{n\to\infty}|(A_n-B_n+C_n+D_n)-E_n|\le C_1\sum_{i=M}^{\infty}\frac{1}{i^2},$$

for any $M\ge C_2$. Recalling (2.7) and (2.8), letting $M\to\infty$, we eventually obtain the desired conclusion. □

Proof of Proposition 2.1. The conclusion corresponding to the case $y=1$ follows from Lemma 2.3. If $y\in(0,1)$, then $\lim_{n\to\infty}r_n=(-\log(1-y))^{1/2}$, and hence $\{t_n:n\ge 1\}$ is bounded. It follows that

$$pt(1+\log 2)-pt\log n+r_n^2(t^2+(p-n+1.5)t)=pt(1+\log 2)-pt\log n-t(p-n)\log\left(1-\frac{p}{n}\right)-\left(t^2+\frac{3t}{2}\right)\log\left(1-\frac{p}{n}\right).$$

The last term above is identical to $(t^2+(3t/2))\log(1-y)+o(1)$ since $p/n\to y$ as $n\to\infty$. Moreover,

$$-pt\log n-t(p-n)\log\left(1-\frac{p}{n}\right)=-pt\log n+t(n-p)(\log(n-p)-\log n)=-nt\log n+t(n-p)\log(n-p).$$

The above three assertions conclude

$$pt(1+\log 2)-pt\log n+r_n^2(t^2+(p-n+1.5)t)=pt(1+\log 2)-nt\log n+(n-p)t\log(n-p)-\left(t^2+\frac{3t}{2}\right)\log(1-y)+o(1),$$

as $n\to\infty$. This is exactly the right hand side of (2.2). □

3. Proof of main results

We first prove Theorem 1. To do that, we need to make a preparation. Assume that $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are $\mathbb{R}^p$-valued random variables. Recall

$$\mathbf{S}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^* \quad \text{where } \overline{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i. \tag{3.1}$$

The following is from Theorem 3.1.2 and Corollary 3.2.19 in Muirhead (1982).

Lemma 3.1. *Assume $n>p$. Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be i.i.d. $\mathbb{R}^p$-valued random variables with distribution $\mathrm{N}_p(\boldsymbol{\mu},\mathbf{I}_p)$. Then $n\mathbf{S}$ and $\mathbf{Z}^*\mathbf{Z}$ have the same distribution, where $\mathbf{Z}:=(z_{ij})_{(n-1)\times p}$ and z_{ij}'s are i.i.d. with distribution $\mathrm{N}(0,1)$. Further, $\lambda_1,\ldots,\lambda_p$ have joint density function*

$$f(\lambda_1,\ldots,\lambda_p)=\mathit{Const}\cdot\prod_{1\le i<j\le p}|\lambda_i-\lambda_j|\cdot\prod_{i=1}^{p}\lambda_i^{(n-p-2)/2}\cdot e^{-(1/2)\sum_{i=1}^{p}\lambda_i},$$

for all $\lambda_1>0,\lambda_2>0,\ldots,\lambda_p>0$.

Recall the β-Laguerre ensemble as follows:

$$f_{\beta,a}(\lambda_1,\ldots,\lambda_p)=c_L^{\beta,a}\cdot\prod_{1\le i<j\le p}|\lambda_i-\lambda_j|^{\beta}\cdot\prod_{i=1}^{p}\lambda_i^{a-q}\cdot e^{-(1/2)\sum_{i=1}^{p}\lambda_i}, \tag{3.2}$$

for all $\lambda_1>0,\lambda_2>0,\ldots,\lambda_p>0$, where

$$c_L^{\beta,a}=2^{-pa}\prod_{j=1}^{p}\frac{\Gamma\left(1+\frac{\beta}{2}\right)}{\Gamma\left(1+\frac{\beta}{2}j\right)\Gamma\left(a-\frac{\beta}{2}(p-j)\right)}, \tag{3.3}$$

$\beta>0$, $p\ge 2$, $a>(\beta/2)(p-1)$ and $q=1+(\beta/2)(p-1)$. See, e.g., Dumitriu and Edelman (2002) and Jiang for further details. It is known that $f_{\beta,a}(\lambda_1,\ldots,\lambda_p)$ is a probability density function, i.e., $\int\cdots\int_{[0,\infty)^p}f_{\beta,a}(\lambda_1,\ldots,\lambda_p)\,d\lambda_1\cdots d\lambda_p=1$. See (17.6.5) from Mehta (2004) (which is essentially a corollary of the Selberg integral in (3.23) below). Evidently, the density function in Lemma 3.1 corresponds to the β-Laguerre ensemble in (3.2) with

$$\beta=1,\quad a=\tfrac{1}{2}(n-1)\quad\text{and}\quad q=1+\tfrac{1}{2}(p-1). \tag{3.4}$$

Lemma 3.2. *Let $n>p$ and L_n^* be as in* (1.3). *Assume $\lambda_1,\ldots,\lambda_p$ have density function $f_{\beta,a}(\lambda_1,\ldots,\lambda_p)$ as in* (3.2) *with $a=(\beta/2)(n-1)$ and $q=1+(\beta/2)(p-1)$. Then*

$$Ee^{tL_n^*}=e^{(\log n-1)pt}\cdot\left(1-\frac{2t}{n}\right)^{p(t-(\beta/2)(n-1))}\cdot 2^{-pt}\cdot\prod_{j=0}^{p-1}\frac{\Gamma\left(a-t-\frac{\beta}{2}j\right)}{\Gamma\left(a-\frac{\beta}{2}j\right)},$$

for any $t\in(-(1/2)\beta,(1/2)(\beta\wedge n))$.

Proof. Recall

$$L_n^*=\frac{1}{n}\sum_{j=1}^{p}(\lambda_j-n\log\lambda_j)+p\log n-p.$$

We then have

$$\begin{aligned}Ee^{tL_n^*}&=e^{(\log n-1)pt}\int_{[0,\infty)^p}e^{(t/n)\sum_{j=1}^{p}\lambda_j}\cdot\prod_{j=1}^{p}\lambda_j^{-t}\cdot f_{\beta,a}(\lambda_1,\ldots,\lambda_p)\,d\lambda_1\cdots d\lambda_p\\&=e^{(\log n-1)pt}\cdot c_L^{\beta,a}\int_{[0,\infty)^p}e^{-((1/2)-(t/n))\sum_{j=1}^{p}\lambda_j}\cdot\prod_{j=1}^{p}\lambda_j^{(a-t)-q}\cdot\prod_{1\le k<l\le p}|\lambda_k-\lambda_l|^{\beta}\,d\lambda_1\cdots d\lambda_p.\end{aligned} \tag{3.5}$$

For $t\in(-\frac{1}{2}\beta,\frac{1}{2}(\beta\wedge n))$, we know $(1/2)-(t/n)>0$. Make transforms $\mu_j=(1-(2t/n))\lambda_j$ for $1\le j\le p$. It follows that the above is identical to

$$e^{(\log n-1)pt}\cdot c_L^{\beta,a}\cdot\left(1-\frac{2t}{n}\right)^{-p(a-t-q)-(\beta/2)p(p-1)-p}\cdot\int_{[0,\infty)^p}e^{-(1/2)\sum_{j=1}^{p}\mu_j}\cdot\prod_{j=1}^{p}\mu_j^{(a-t)-q}\cdot\prod_{1\le k<l\le p}|\mu_k-\mu_l|^{\beta}\,d\mu_1\cdots d\mu_p. \tag{3.6}$$

Since $t \in (-\frac{1}{2}\beta, \frac{1}{2}(\beta \wedge n))$ and $n-p \geq 1$, we know

$$t < \frac{\beta}{2} \leq \frac{\beta}{2}(n-p) = \frac{\beta}{2}(n-1) - \frac{\beta}{2}(p-1) = a - \frac{\beta}{2}(p-1).$$

That is, $a-t > (\beta/2)(p-1)$. Therefore the integral in (3.6) is equal to $1/c_L^{\beta,a-t}$ by (3.2) and (3.3). It then from (3.5) and (3.6) that

$$\begin{aligned} Ee^{tL_n^*} &= e^{(\log n-1)pt} \cdot \left(1-\frac{2t}{n}\right)^{-p(a-t-q)-(\beta/2)p(p-1)-p} \cdot \frac{c_L^{\beta,a}}{c_L^{\beta,a-t}} \\ &= e^{(\log n-1)pt} \cdot \left(1-\frac{2t}{n}\right)^{-p(a-t-q)-(\beta/2)p(p-1)-p} \cdot 2^{-pt} \cdot \prod_{j=1}^{p} \frac{\Gamma\left(a-t-\frac{\beta}{2}(p-j)\right)}{\Gamma\left(a-\frac{\beta}{2}(p-j)\right)}. \end{aligned}$$

Now, use $a = (\beta/2)(n-1)$ and $q = 1+(\beta/2)(p-1)$ to obtain that

$$Ee^{tL_n^*} = e^{(\log n-1)pt} \cdot \left(1-\frac{2t}{n}\right)^{p(t-(\beta/2)(n-1))} \cdot 2^{-pt} \cdot \prod_{j=0}^{p-1} \frac{\Gamma\left(a-t-\frac{\beta}{2}j\right)}{\Gamma\left(a-\frac{\beta}{2}j\right)}.$$

The proof is completed. □

Let $\{Z, Z_n;\ n \geq 1\}$ be a sequence of random variables. It is known that

$$Z_n \text{ converges to } Z \text{ in distribution if } \lim_{n\to\infty} Ee^{tZ_n} = Ee^{tZ} < \infty, \tag{3.7}$$

for all $t \in (-t_0, t_0)$, where $t_0 > 0$ is a constant. See, e.g., p. 408 from Billingsley (1986).

Proof of Theorem 1. First, since $\log(1-x) < -x$ for all $x < 1$, we know $\sigma_n^2 > 0$ for all $n > p \geq 1$. Now, by assumption, it is easy to see

$$\lim_{n\to\infty} \sigma_n^2 = \begin{cases} -2[y+\log(1-y)] & \text{if } y \in (0,1); \\ +\infty & \text{if } y = 1. \end{cases} \tag{3.8}$$

Trivially, the limit is always positive. Consequently,

$$\delta_0 := \inf\{\sigma_n; n > p \geq 1\} > 0.$$

To finish the proof, by (3.7) it is enough to show that

$$E \exp\left\{\frac{L_n^* - \mu_n}{\sigma_n} s\right\} \to e^{s^2/2} = Ee^{sN(0,1)}, \tag{3.9}$$

as $n \to \infty$ for all s such that $|s| < \delta_0/2$.

Fix s such that $|s| < \delta_0/2$. Set $t = t_n = s/\sigma_n$. Then $|t_n| < 1/2$ for all $n > p \geq 1$. In Lemma 3.2, take $\beta = 1$ and $a = (n-1)/2$, by (3.4),

$$Ee^{tL_n^*} = e^{(\log n-1)pt} \cdot \left(1-\frac{2t}{n}\right)^{pt-(np/2)+(p/2)} \cdot 2^{-pt} \cdot \prod_{j=0}^{p-1} \frac{\Gamma\left(\frac{n-j-1}{2}-t\right)}{\Gamma\left(\frac{n-j-1}{2}\right)}.$$

Letting $i = n-j-1$, we get

$$Ee^{tL_n^*} = 2^{-pt} \cdot e^{(\log n-1)pt} \cdot \left(1-\frac{2t}{n}\right)^{pt-(np/2)+(p/2)} \cdot \prod_{i=n-p}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}, \tag{3.10}$$

for $n > p$. Then

$$\log Ee^{tL_n^*} = pt(\log n - 1 - \log 2) + p\left(t+\frac{1-n}{2}\right)\log\left(1-\frac{2t}{n}\right) + \log \prod_{i=n-p}^{n-1} \frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}.$$

Now, use identity $\log(1-x) = -x-(x^2/2)+O(x^3)$ as $x \to 0$ to have

$$p\left(t+\frac{1-n}{2}\right)\log\left(1-\frac{2t}{n}\right) = p\left(t+\frac{1-n}{2}\right)\left(-\frac{2t}{n}-\frac{2t^2}{n^2}+O\left(\frac{1}{n^3}\right)\right)$$

$$=-\frac{2pt}{n}\left(t+\frac{1-n}{2}\right)\left(1+\frac{t}{n}\right)+o(1)=-\frac{2pt}{n}\left(\frac{1}{2}t+\frac{1-n}{2}+O\left(\frac{1}{n}\right)\right)+o(1)=-\frac{p}{n}t^2+pt-yt+o(1),$$

as $n\to\infty$. Recall $r_n=(-\log(1-(p/n)))^{1/2}$. We know $t=t_n=s/\sigma_n=O(1/r_n)$ as $n\to\infty$. By Proposition 2.1,

$$\log\prod_{i=n-p}^{n-1}\frac{\Gamma\left(\frac{i}{2}-t\right)}{\Gamma\left(\frac{i}{2}\right)}=pt(1+\log 2)-pt\log n+r_n^2(t^2+(p-n+1.5)t)+o(1),$$

as $n\to\infty$. Join all the assertions from (3.10) to the above to obtain that

$$\begin{aligned}\log Ee^{tL_n^*}&=pt(\log n-1-\log 2)-\frac{p}{n}t^2+pt-yt+pt(1+\log 2)-pt\log n+r_n^2(t^2+(p-n+1.5)t)+o(1)\\&=\left(-\frac{p}{n}+r_n^2\right)t^2+[p+r_n^2(p-n+1.5)-y]t+o(1),\end{aligned}\tag{3.11}$$

as $n\to\infty$. Noticing

$$p+r_n^2(p-n+1.5)-y=\left(n-p-\frac{3}{2}\right)\log\left(1-\frac{p}{n}\right)+p-y=\mu_n,$$

and from the definition of σ_n and notation $t=s/\sigma_n$, we know $(-(p/n)+r_n^2)t^2=s^2/2$. Hence, it follows from (3.11) that

$$\log E\exp\left\{\frac{L_n^*-\mu_n}{\sigma_n}s\right\}=\log Ee^{tL_n^*}-\mu_n t\to\frac{s^2}{2},$$

as $n\to\infty$. This implies (3.9). The proof is completed. □

Now we start to prove Theorem 2. The following lemma says that the distribution of L_1 in (1.7) does not depend on the mean vectors or covariance matrices of the population distributions where random samples $\mathbf{x}_i$'s and $\mathbf{y}_j$'s come from.

Lemma 3.3. *Let L_1 be defined as in (1.7) with $n_1>p$ and $n_2>p$. Then, under H_0 in (1.4), L_1 and*

$$\tilde{L}_1:=\frac{(n_1+n_2)^{(n_1+n_2)p/2}}{n_1^{n_1p/2}n_2^{n_2p/2}}|\mathbf{C}|^{n_1/2}\cdot|\mathbf{I}-\mathbf{C}|^{n_2/2}\tag{3.12}$$

have the same distribution, where

$$\mathbf{C}=(\mathbf{U}^*\mathbf{U}+\mathbf{V}^*\mathbf{V})^{-1/2}(\mathbf{U}^*\mathbf{U})(\mathbf{U}^*\mathbf{U}+\mathbf{V}^*\mathbf{V})^{-1/2},\tag{3.13}$$

with $\mathbf{U}=(u_{ij})_{(n_1-1)\times p}$ and $\mathbf{V}=(v_{ij})_{(n_2-1)\times p}$, and $\{u_{ij},v_{kl}\}$ are i.i.d. random variables with distribution $N(0,1)$.

Proof. Recall that $\mathbf{x}_1,\ldots,\mathbf{x}_{n_1}$ is a random sample from population $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, and $\mathbf{y}_1,\ldots,\mathbf{y}_{n_2}$ is a random sample from population $N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, and the two sets of random variables are independent. Under H_0 in (1.4), $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$ is non-singular. Set

$$\tilde{\mathbf{x}}_i=\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i\quad\text{and}\quad\tilde{\mathbf{y}}_j=\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_j,$$

for $1\le i\le n_1$ and $1\le j\le n_2$. Then $\{\tilde{\mathbf{x}}_i;\ 1\le i\le n_1\}$ are i.i.d. with distribution $N_p(\tilde{\boldsymbol{\mu}}_1,\mathbf{I}_p)$ where $\tilde{\boldsymbol{\mu}}_1=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_1$; $\{\tilde{\mathbf{y}}_j;\ 1\le j\le n_2\}$ are i.i.d. with distribution $N_p(\tilde{\boldsymbol{\mu}}_2,\mathbf{I}_p)$ where $\tilde{\boldsymbol{\mu}}_2=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_2$. Further, $\{\tilde{\mathbf{x}}_i;\ 1\le i\le n_1\}$ and $\{\tilde{\mathbf{y}}_j;\ 1\le j\le n_2\}$ are obviously independent. Similar to (1.5) and (1.6), define

$$\tilde{\mathbf{A}}=\frac{1}{n_1}\sum_{i=1}^{n_1}(\tilde{\mathbf{x}}_i-\bar{\tilde{\mathbf{x}}})(\tilde{\mathbf{x}}_i-\bar{\tilde{\mathbf{x}}})^*\quad\text{and}\quad\tilde{\mathbf{B}}=\frac{1}{n_2}\sum_{i=1}^{n_2}(\tilde{\mathbf{y}}_i-\bar{\tilde{\mathbf{y}}})(\tilde{\mathbf{y}}_i-\bar{\tilde{\mathbf{y}}})^*,\tag{3.14}$$

where

$$\bar{\tilde{\mathbf{x}}}=\frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{\mathbf{x}}_i\quad\text{and}\quad\bar{\tilde{\mathbf{y}}}=\frac{1}{n_2}\sum_{i=1}^{n_2}\tilde{\mathbf{y}}_i.\tag{3.15}$$

It is easy to check that

$$\mathbf{A}=\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{A}}\boldsymbol{\Sigma}^{1/2}\quad\text{and}\quad\mathbf{B}=\boldsymbol{\Sigma}^{1/2}\tilde{\mathbf{B}}\boldsymbol{\Sigma}^{1/2}.\tag{3.16}$$

By Lemma 3.1,

$$n_1\tilde{\mathbf{A}}\overset{d}{=}\mathbf{U}^*\mathbf{U}\quad\text{and}\quad n_2\tilde{\mathbf{B}}\overset{d}{=}\mathbf{V}^*\mathbf{V},\tag{3.17}$$

where $\mathbf{U}=(u_{ij})_{(n_1-1)\times p}$ and $\mathbf{V}=(v_{ij})_{(n_2-1)\times p}$, and $\{u_{ij},v_{kl};\ i,j,k,l\ge 1\}$ are i.i.d. random variables with distribution $N(0,1)$. Review (1.7),

$$L_1=\frac{|\mathbf{A}|^{n_1/2}\cdot|\mathbf{B}|^{n_2/2}}{|c_1\mathbf{A}+c_2\mathbf{B}|^{N/2}}=\frac{N^{Np/2}}{n_1^{n_1p/2}n_2^{n_2p/2}}\cdot\frac{|n_1\mathbf{A}|^{n_1/2}\cdot|n_2\mathbf{B}|^{n_2/2}}{|n_1\mathbf{A}+n_2\mathbf{B}|^{N/2}}$$

$$= \frac{N^{Np/2}}{n_1^{n_1p/2}n_2^{n_2p/2}} \cdot \frac{|n_1\tilde{\mathbf{A}}|^{n_1/2} \cdot |n_2\tilde{\mathbf{B}}|^{n_2/2}}{|n_1\tilde{\mathbf{A}}+n_2\tilde{\mathbf{B}}|^{N/2}}, \tag{3.18}$$

since $|n_1\mathbf{A}| = |n_1\tilde{\mathbf{A}}| \cdot |\mathbf{\Sigma}|$ and $|n_2\mathbf{B}| = |n_2\tilde{\mathbf{B}}| \cdot |\mathbf{\Sigma}|$ and

$$|n_1\mathbf{A}+n_2\mathbf{B}| = |\mathbf{\Sigma}^{1/2}(n_1\tilde{\mathbf{A}}+n_2\tilde{\mathbf{B}})\mathbf{\Sigma}^{1/2}| = |n_1\tilde{\mathbf{A}}+n_2\tilde{\mathbf{B}}| \cdot |\mathbf{\Sigma}|,$$

by (3.16), and hence the term $|\mathbf{\Sigma}|^{(n_1+n_2)/2}$ in the numerator canceled $|\mathbf{\Sigma}|^{N/2}$ in the denominator. Define $\tilde{\mathbf{C}} = (n_1\tilde{\mathbf{A}}+n_2\tilde{\mathbf{B}})^{-1/2}(n_1\tilde{\mathbf{A}})(n_1\tilde{\mathbf{A}}+n_2\tilde{\mathbf{B}})^{-1/2}$. We see from the independence between $n_1\tilde{\mathbf{A}}$ and $n_2\tilde{\mathbf{B}}$ and the independence between $\mathbf{U}^*\mathbf{U}$ and $\mathbf{V}^*\mathbf{V}$ that

$$\tilde{\mathbf{C}} \overset{d}{=} \mathbf{C}, \tag{3.19}$$

where $\mathbf{C}$ is as in (3.13). It is obvious that

$$|\tilde{\mathbf{C}}| = |n_1\tilde{\mathbf{A}}| \cdot |n_1\tilde{\mathbf{A}}+n_2\tilde{\mathbf{B}}|^{-1} \quad \text{and} \quad |\mathbf{I}-\tilde{\mathbf{C}}| = |n_2\tilde{\mathbf{B}}| \cdot |n_1\tilde{\mathbf{A}}+n_2\tilde{\mathbf{B}}|^{-1}.$$

Hence we have from (3.18) that

$$L_1 = \frac{N^{Np/2}}{n_1^{n_1p/2}n_2^{n_2p/2}} \cdot |\tilde{\mathbf{C}}|^{n_1/2} \cdot |\mathbf{I}-\tilde{\mathbf{C}}|^{n_2/2}. \tag{3.20}$$

Finally, we get the desired conclusion from (3.19) and (3.20). □

Let $\lambda_1,\ldots,\lambda_p$ be the eigenvalues of the β-Jacobi ensemble or the β-MANOVA matrix, that is, they have the joint probability density function

$$f(\lambda_1,\ldots,\lambda_p) = c_J^{\beta,a_1,a_2} \prod_{1 \le i < j \le p} |\lambda_i-\lambda_j|^{\beta} \cdot \prod_{i=1}^{p} \lambda_i^{a_1-q}(1-\lambda_i)^{a_2-q}, \tag{3.21}$$

for $0 \le \lambda_1,\ldots,\lambda_p \le 1$, where $a_1,a_2 > (\beta/2)(p-1)$ are parameters, $q = 1+(\beta/2)(p-1)$, and

$$c_J^{\beta,a_1,a_2} = \prod_{j=1}^{p} \frac{\Gamma\left(1+\frac{\beta}{2}\right)\Gamma\left(a_1+a_2-\frac{\beta}{2}(p-j)\right)}{\Gamma\left(1+\frac{\beta}{2}j\right)\Gamma\left(a_1-\frac{\beta}{2}(p-j)\right)\Gamma\left(a_2-\frac{\beta}{2}(p-j)\right)}, \tag{3.22}$$

with $a_1 = (\beta/2)(n_1-1)$ and $a_2 = (\beta/2)(n_2-1)$. The fact that $f(\lambda_1,\ldots,\lambda_p)$ is a probability density function follows from the Selberg integral (see, e.g., Forrester and Warnaar, 2008; Mehta, 2004):

$$\int_{[0,1]^p} \prod_{1 \le i < j \le p} |\lambda_i-\lambda_j|^{\beta} \cdot \prod_{i=1}^{p} \lambda_i^{a_1-q}(1-\lambda_i)^{a_2-q}\, d\lambda_1 \cdots \lambda_p = \frac{1}{c_J^{\beta,a_1,a_2}}. \tag{3.23}$$

It is known that the eigenvalues of $\mathbf{C}$ defined in (3.13) has density function $f(\lambda_1,\ldots,\lambda_p)$ in (3.21)

$$\text{with } \beta = 1, \quad a_1 = \tfrac{1}{2}(n_1-1), \quad a_2 = \tfrac{1}{2}(n_2-1) \quad \text{and} \quad q = 1+\tfrac{1}{2}(p-1). \tag{3.24}$$

See, for example, Constantine (1963) and Muirhead (1982) for this fact.

Lemma 3.4. *Let T_N be as in* (1.7). *Assume $n_1 > p$ and $n_2 > p$. Then*

$$Ee^{tT_N} = C_{n_1,n_2} \cdot U_n(t) \cdot V_{1,n}(t)^{-1} \cdot V_{2,n}(t)^{-1},$$

for all $t < (1/2)(1-(p/(n_1 \wedge n_2)))$, where

$$C_{n_1,n_2} = \frac{n_1^{n_1pt}n_2^{n_2pt}}{(n_1+n_2)^{(n_1+n_2)pt}}, \quad U_n(t) = \prod_{i=N-p-1}^{N-2} \frac{\Gamma(\frac{1}{2}i)}{\Gamma(\frac{1}{2}i-Nt)},$$

$$V_{1,n}(t) = \prod_{i=n_1-p}^{n_1-1} \frac{\Gamma(\frac{1}{2}i)}{\Gamma(\frac{1}{2}i-n_1t)} \quad \textit{and} \quad V_{2,n}(t) = \prod_{i=n_2-p}^{n_2-1} \frac{\Gamma(\frac{1}{2}i)}{\Gamma(\frac{1}{2}i-n_2t)}. \tag{3.25}$$

Proof. From (1.7), $e^{tT_N} = (L_1)^{-2t}$ for any $t \in \mathbb{R}$. Therefore, by Lemma 3.3,

$$Ee^{tT_N} = C_{n_1,n_2} \cdot E(|\mathbf{C}|^{-n_1t} \cdot |\mathbf{I}-\mathbf{C}|^{-n_2t}) = C_{n_1,n_2} \cdot E\left(\prod_{j=1}^{p} \lambda_j^{-n_1t}(1-\lambda_j)^{-n_2t}\right),$$

where $\lambda_1,\ldots,\lambda_p$ are the eigenvalues of $\mathbf{C}$ in (3.13). Write $c_J^{a_1,a_2} = c_J^{1,a_1,a_2}$. By (3.22) and (3.24),

$$Ee^{tT_N} = C_{n_1,n_2} \cdot c_J^{a_1,a_2} \cdot \int_{[0,1]^p} \prod_{j=1}^{p} \lambda_j^{a_1-n_1t-q}(1-\lambda_j)^{a_2-n_2t-q} \cdot \prod_{1 \le i < j \le p} |\lambda_i-\lambda_j|\, d\lambda_1 \cdots \lambda_p$$

$$= C_{n_1,n_2} \cdot \frac{c_J^{a_1,a_2}}{c_J^{a_1-n_1t,a_2-n_2t}}, \tag{3.26}$$

since $f(\lambda_1,\ldots,\lambda_p)$ is a probability density function. Of course, recalling $a_i=\frac{1}{2}(n_i-1)$ for $i=1, 2$ and the assumption that $t<\frac{1}{2}(1-p/(n_1\wedge n_2))$, we know

$$a_1-n_1t>\tfrac{1}{2}(p-1) \quad \text{and} \quad a_2-n_2t>\tfrac{1}{2}(p-1),$$

which are required in (3.21). From (3.26), we see

$$\begin{aligned} Ee^{tT_N} &= C_{n_1,n_2}\cdot\prod_{j=1}^{p}\frac{\Gamma(a_1+a_2-\frac{1}{2}(p-j))}{\Gamma(a_1+a_2-Nt-\frac{1}{2}(p-j))}\cdot\left[\prod_{j=1}^{p}\frac{\Gamma\left(a_1-\frac{1}{2}(p-j)\right)}{\Gamma\left(a_1-n_1t-\frac{1}{2}(p-j)\right)}\right]^{-1}\cdot\left[\prod_{j=1}^{p}\frac{\Gamma\left(a_2-\frac{1}{2}(p-j)\right)}{\Gamma\left(a_2-n_2t-\frac{1}{2}(p-j)\right)}\right]^{-1}\\ &=: C_{n_1,n_2}\cdot\tilde{U}_n(t)\cdot\tilde{V}_{1,n}(t)^{-1}\cdot\tilde{V}_{2,n}(t)^{-1}. \end{aligned} \tag{3.27}$$

Now, use $a_i=\frac{1}{2}(n_i-1)$ for $i=1, 2$ again to have

$$a_1-\tfrac{1}{2}(p-j)=\tfrac{1}{2}(n_1-p+j-1);\quad a_2-\tfrac{1}{2}(p-j)=\tfrac{1}{2}(n_2-p+j-1);\quad a_1+a_2-\tfrac{1}{2}(p-j)=\tfrac{1}{2}(N-p+j-2).$$

Thus, by setting $i=N-p+j-2$ for $j=1,2,\ldots,p$, we have

$$\tilde{U}_n(t)=\prod_{j=1}^{p}\frac{\Gamma(a_1+a_2-\frac{1}{2}(p-j))}{\Gamma(a_1+a_2-Nt-\frac{1}{2}(p-j))}=\prod_{i=N-p-1}^{N-2}\frac{\Gamma(\frac{1}{2}i)}{\Gamma(\frac{1}{2}i-Nt)}=U_n(t).$$

Similarly, $\tilde{V}_{i,n}(t)=V_{i,n}(t)$ for $i=1, 2$. These combining with (3.27) yield the desired result. □

Lemma 3.5. *Let T_N be as in (1.7). Assume $n_i>p$ and $p/n_i\to y_i\in(0,1]$ for $i=1, 2$. Recall σ_n^2 in (1.8). Then, $0<\sigma_n<\infty$ for all $n_1\geq 2,n_2\geq 2$, and $E\exp\{(T_N/(N\sigma_n))t\}<\infty$ for all $t\in\mathbb{R}$ as n_1 and n_2 are sufficiently large.*

Proof. First, we claim that

$$\sigma^2 := 2[\log(1-y)-\gamma_1^2\log(1-y_1)-\gamma_2^2\log(1-y_2)]>0, \tag{3.28}$$

for all $y_1,y_2\in(0,1)$, where $\gamma_1=y_2(y_1+y_2)^{-1}$, $\gamma_2=y_1(y_1+y_2)^{-1}$ and $y=y_1y_2(y_1+y_2)^{-1}$.

In fact, consider $h(x)=-\log(1-x)$ for $x<1$. Then, $h''(x)=(1-x)^{-2}>0$ for $x<1$. That is, $h(x)$ is a convex function. Take $\gamma_3=2y_1y_2/(y_1+y_2)^2$. Then, $\gamma_1^2+\gamma_2^2+\gamma_3=1$. Hence, by the convexity,

$$\begin{aligned} -\gamma_1^2\log(1-y_1)-\gamma_2^2\log(1-y_2) &= -\gamma_1^2\log(1-y_1)-\gamma_2^2\log(1-y_2)-\gamma_3\log(1-0)\\ &< -\log(1-(\gamma_1^2y_1+\gamma_2^2y_2+\gamma_3\cdot 0))=-\log(1-y), \end{aligned}$$

where the strict inequality comes since $y_1\neq 0$ and $y_2\neq 0$.

Now, taking $y_i=p/n_i\in(0,1)$ for $i=1, 2$ in (3.28), we get

$$\gamma_1=\frac{y_2}{y_1+y_2}=\frac{n_1}{N},\quad \gamma_2=\frac{y_1}{y_1+y_2}=\frac{n_2}{N}\quad\text{and}\quad y=\frac{y_1y_2}{y_1+y_2}=\frac{p}{N}.$$

Evidently, $n_1/N,n_2/N,p/N\in(0,1)$. Then, by (3.28), we know $0<\sigma_n<\infty$ for all $n_1\geq 2,n_2\geq 2$.

Second, noting that

$$\left\{t:\frac{t}{N\sigma_n}<\frac{1}{2}\left(1-\frac{p}{n_1\wedge n_2}\right)\right\}=\left(-\infty,\frac{1}{2}\left(1-\frac{p}{n_1\wedge n_2}\right)N\sigma_n\right),$$

to prove the second part, it suffices to show from Lemma 3.4 that

$$\lim_{n_1,n_2\to\infty}\left(1-\frac{p}{n_1\wedge n_2}\right)N\sigma_n=+\infty. \tag{3.29}$$

Case 1: $y_1<1, y_2<1$. Recall σ^2 in (3.28). Evidently, $\sigma_n^2\to\sigma^2\in(0,\infty)$ as $n_1,n_2\to+\infty$. Hence, (3.29) follows since $1-(p/(n_1\wedge n_2))\to 1-y_1\vee y_2>0$.

Case 2: $\max\{y_1,y_2\}=1$. This implies $\sigma_n^2\to+\infty$ as $n_1,n_2\to\infty$ because $\log(1-(p/N))\to\log y\in(-\infty,0)$ and the sum of the last two terms on the right hand side of (1.8) goes to $+\infty$. Further, the given conditions say that $n_i-1\geq p$, and hence, $1-(p/n_i)\geq 1/n_i\geq 1/N$ for $i=1, 2$. Thus,

$$\left(1-\frac{p}{n_1\wedge n_2}\right)N\sigma_n=\min\left\{1-\frac{p}{n_1},1-\frac{p}{n_2}\right\}\cdot N\sigma_n\geq\sigma_n\to+\infty,$$

as $n_1,n_2\to\infty$. We get (3.29). The proof is completed. □

Proof of Theorem 2. From Lemma 3.5, we assume, without loss of generality, that $E\exp\{(T_N/(N\sigma_n))t\}<\infty$ for all $n_1\geq 2,n_2\geq 2$ and $t\in\mathbb{R}$. Fix $t\in\mathbb{R}$. Set $t_n=t_{n_1,n_2}=t/\sigma_n$ for $n_1,n_2\geq 2$. From the condition $p/n_i\to y_i$ for $i=1,2$ as $p\wedge n_1\wedge n_2=p\to\infty$ by the assumption $n_1>p$ and $n_2>p$ (we will simply say "$p\to\infty$" in similar situations later), we know

σ_n^2 has a positive limit (possibly $+\infty$) as $p\to\infty$. It follows that $\{t_n;\ n_1,n_2\geq 2\}$ is bounded. By Lemma 3.4,

$$\log E\exp\left\{\frac{T_N}{N\sigma_n}t\right\}=-\log V_{1,n}\left(\frac{t_n}{N}\right)-\log V_{2,n}\left(\frac{t_n}{N}\right)+\log U_n\left(\frac{t_n}{N}\right)+\frac{pt_n}{N}(n_1\log n_1+n_2\log n_2-N\log N). \tag{3.30}$$

Set $\gamma_1=y_2(y_1+y_2)^{-1}$, $\gamma_2=y_1(y_1+y_2)^{-1}$ and $y=y_1y_2(y_1+y_2)^{-1}$. Easily,

$$\frac{n_i}{N}\to\gamma_i\in(0,1),\quad \frac{p}{N-1}\to y\in(0,1)\quad\text{and}\quad 2\log\left(1-\frac{p}{N}\right)\to 2\log(1-y)\in(-\infty,0),$$

as $p\to\infty$. Then, from (1.8) we know that

$$\frac{n_i}{N}t_n\sim\gamma_i t\cdot\frac{1}{\sigma_n}=\mathrm{O}\left(\left(-\log\left(1-\frac{p}{n_i}\right)\right)^{-1/2}\right)\quad\text{and}\quad t_n=\mathrm{O}\left(\left(-\log\left(1-\frac{p}{N-1}\right)\right)^{-1/2}\right), \tag{3.31}$$

for $i=1,2$ as $p\to\infty$. Replacing "t" in Proposition 2.1 with "n_1t_n/N", we have

$$-\log V_{1,n}\left(\frac{t_n}{N}\right)=\log\prod_{i=n_1-p}^{n_1-1}\frac{\Gamma\left(\frac{i}{2}-\frac{n_1}{N}t_n\right)}{\Gamma\left(\frac{i}{2}\right)}=\frac{n_1pt_n}{N}(1+\log 2)-\frac{n_1pt_n}{N}\log n_1+r_{n,1}^2\left(\frac{n_1^2}{N^2}t_n^2+(p-n_1+1.5)\frac{n_1}{N}t_n\right)+\mathrm{o}(1), \tag{3.32}$$

as $p\to\infty$, where

$$r_{n,i}:=\left(-\log\left(1-\frac{p}{n_i}\right)\right)^{1/2},\quad i=1,2. \tag{3.33}$$

Similarly,

$$-\log V_{2,n}\left(\frac{t_n}{N}\right)=\log\prod_{i=n_2-p}^{n_2-1}\frac{\Gamma\left(\frac{i}{2}-\frac{n_2}{N}t_n\right)}{\Gamma\left(\frac{i}{2}\right)}=\frac{n_2pt_n}{N}(1+\log 2)-\frac{n_2pt_n}{N}\log n_2+r_{n,2}^2\left(\frac{n_2^2}{N^2}t_n^2+(p-n_2+1.5)\frac{n_2}{N}t_n\right)+\mathrm{o}(1), \tag{3.34}$$

as $p\to\infty$. By the same argument, by using (3.31) we see

$$-\log U_n\left(\frac{t_n}{N}\right)=\log\prod_{i=(N-1)-p}^{(N-1)-1}\frac{\Gamma\left(\frac{i}{2}-t_n\right)}{\Gamma\left(\frac{i}{2}\right)}=pt_n(1+\log 2)-pt_n\log(N-1)+R_n^2(t_n^2+(p-N+2.5)t_n)+\mathrm{o}(1), \tag{3.35}$$

as $p\to\infty$, where

$$R_n=\left(-\log\left(1-\frac{p}{N-1}\right)\right)^{1/2}. \tag{3.36}$$

From (3.32) and (3.34),

$$\begin{aligned}-\log V_{i,n}\left(\frac{t_n}{N}\right)+\frac{pt_n}{N}n_i\log n_i&=\frac{n_ipt_n}{N}(1+\log 2)+r_{n,i}^2\left(\frac{n_i^2}{N^2}t_n^2+(p-n_i+1.5)\frac{n_i}{N}t_n\right)+\mathrm{o}(1)\\&=\frac{n_ipt_n}{N}(1+\log 2)+\frac{n_i^2r_{n,i}^2}{N^2}t_n^2+\frac{(p-n_i+1.5)n_ir_{n,i}^2}{N}t_n+\mathrm{o}(1),\end{aligned} \tag{3.37}$$

as $p\to\infty$ for $i=1,2$. Since $\{t_n\}$ is bounded, use $\log(1+x)=x+\mathrm{O}(x^2)$ as $x\to 0$ to see

$$pt_n\log N-pt_n\log(N-1)=pt_n\log\left(1+\frac{1}{N-1}\right)=yt_n+\mathrm{o}(1),$$

as $p\to\infty$, where $\lim p/(N-1)=y_1y_2/(y_1+y_2)=y<1$. Therefore, by (3.35) and the fact $N=n_1+n_2$,

$$\begin{aligned}-\log U_n\left(\frac{t_n}{N}\right)+pt_n\log N&=pt_n(1+\log 2)+yt_n+R_n^2(t_n^2+(p-N+2.5)t_n)+\mathrm{o}(1)\\&=\frac{n_1pt_n+n_2pt_n}{N}(1+\log 2)+R_n^2t_n^2+(y+(p-N+2.5)R_n^2)t_n+\mathrm{o}(1),\end{aligned} \tag{3.38}$$

as $p\to\infty$. Joining (3.30) with (3.37) and (3.38), we obtain

$$\log Ee^{t_nT_N/N}=\left(\frac{n_1^2}{N^2}r_{n,1}^2+\frac{n_2^2}{N^2}r_{n,2}^2-R_n^2\right)t_n^2+\rho_nt_n+\mathrm{o}(1), \tag{3.39}$$

as $p\to\infty$, where

$$\rho_n=\frac{1}{N}((p-n_1+1.5)n_1r_{n,1}^2+(p-n_2+1.5)n_2r_{n,2}^2)-(p-N+2.5)R_n^2-y. \tag{3.40}$$

By using the fact $\log(1+x)=x+o(x^2)$ again, we have that

$$\log\left(\frac{N-1}{N}\cdot\frac{N-p}{N-p-1}\right)=\log\left(1-\frac{1}{N}\right)-\log\left(1-\frac{1}{N-p}\right)=\frac{p}{N(N-p)}+O\left(\frac{1}{N^2}\right),$$

as $p\to\infty$. Reviewing (3.36), we have

$$R_n^2=-\log\left(1-\frac{p}{N-1}\right)=-\log\left(1-\frac{p}{N}\right)+\log\left(\frac{N-1}{N}\cdot\frac{N-p}{N-p-1}\right)=r_n^2+\frac{p}{N(N-p)}+O\left(\frac{1}{N^2}\right), \tag{3.41}$$

as $p\to\infty$, where

$$r_n=\left(-\log\left(1-\frac{p}{N}\right)\right)^{1/2}.$$

In particular, since $\{t_n\}$ is bounded,

$$R_n^2t_n^2=r_n^2t_n^2+o(1), \tag{3.42}$$

as $p\to\infty$. By (3.41), recalling $p/N\to y$, we get

$$(p-N+2.5)R_n^2=(p-N+2.5)r_n^2-\frac{p}{N}+o(1)=(p-N+2.5)r_n^2-y+o(1),$$

as $p\to\infty$. Plug this into (3.40) to have that

$$\rho_n=\frac{1}{N}((p-n_1+1.5)n_1r_{n,1}^2+(p-n_2+1.5)n_2r_{n,2}^2)-(p-N+2.5)r_n^2+o(1), \tag{3.43}$$

as $p\to\infty$. Now plug the above and (3.42) into (3.39), since $\{t_n\}$ is bounded, we have

$$\log Ee^{t_nT_N/N}=\left(\frac{n_1^2}{N^2}r_{n,1}^2+\frac{n_2^2}{N^2}r_{n,2}^2-r_n^2\right)t_n^2+\mu_nt_n+o(1), \tag{3.44}$$

as $p\to\infty$ with

$$\mu_n=\frac{1}{N}((p-n_1+1.5)n_1r_{n,1}^2+(p-n_2+1.5)n_2r_{n,2}^2)-(p-N+2.5)r_n^2.$$

Using $t_n=t/\sigma_n$ and the definition of σ_n, we get

$$\left(\frac{n_1^2}{N^2}r_{n,1}^2+\frac{n_2^2}{N^2}r_{n,2}^2-r_n^2\right)t_n^2=t_n^2\left(\log\left(1-\frac{p}{N}\right)-\frac{n_1^2}{N^2}\log\left(1-\frac{p}{n_1}\right)-\frac{n_2^2}{N^2}\log\left(1-\frac{p}{n_2}\right)\right)\to\frac{t^2}{2},$$

as $p\to\infty$. This and (3.44) conclude that

$$\log E\exp\left\{\frac{T_n-N\mu_n}{N}t_n\right\}=\log Ee^{t_nT_N/N}-\mu_nt_n\to\frac{t^2}{2},$$

as $p\to\infty$, which is equivalent to that

$$E\exp\left\{\frac{1}{\sigma_n}\left(\frac{T_n}{N}-\mu_n\right)t\right\}\to e^{t^2/2}=Ee^{tN(0,1)},$$

as $p\to\infty$ for any $t\in\mathbb{R}$. The proof is completed by using (3.7). □

Acknowledgment

We thank Danning Li very much for her check of our proofs and many good suggestions. We also thank an anonymous referee for very helpful comments for revision.

References

Ahlfors, L.V., 1979. Complex Analysis, 3rd ed. McGraw-Hill, Inc.
Bai, Z., Saranadasa, H., 1996. Effect of high dimension comparison of significance tests for a high-dimensional two sample problem. Statistica Sinica 6, 311–329.
Bai, Z., Silverstein, J., 2004. CLT for linear spectral statistics of large dimensional sample covariance matrices. Annals of Probability 32, 553–605.
Bai, Z., Jiang, D., Yao, J., Zheng, S., 2009. Corrections to LRT on large-dimensional covariance matrix by RMT. Annals of Statistics 37 (6B), 3822–3840.
Billingsley, P., 1986. Probability and Measure. Wiley Series in Probability and Mathematical Statistics, 2nd ed.
Constantine, A., 1963. Some non-central distribution problems in multivariate analysis. Annals of Mathematical Statistics 34, 1270–1285.
Dempster, A., 1958. A high-dimensional two sample significance test. Annals of Mathematical Statistics 29, 995–1010.
Dumitriu, I., Edelman, A., 2002. Matrix models for beta-ensembles. Journal of Mathematical Physics 43 (11), 5830–5847.

Forrester, P., Warnaar, S., 2008. The importance of the Selberg integral. Bulletin of the American Mathematical Society 45 (4), 489–534.
Freitag, E., Busam, R., 2005. Complex Analysis. Springer.
Gamelin, T.W., 2001. Complex Analysis, 1st ed. Springer.
Jiang, T. Limit theorems on Beta-Jacobi ensembles 〈http://arXiv.org/abs/0911.2262〉.
Mehta, M.L., 2004. Random Matrices, Pure and Applied Mathematics (Amsterdam), 3rd ed. vol. 142. Elsevier, Academic Press, Amsterdam.
Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. Wiley, New York.
Zheng, S., 2008. Central limit theorem for linear spectral statistics of large dimensional F-matrix. Preprint. Northeast Normal University, Changchun, China.