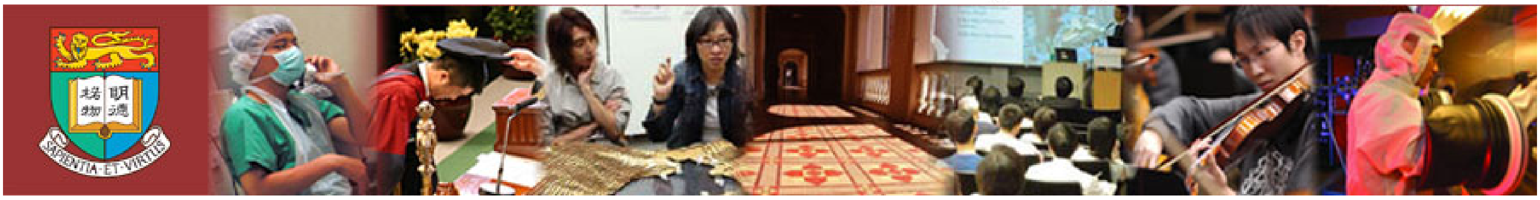

Title	Corrections to LRT on large-dimensional covariance matrix by RMT
Author(s)	Bai, Z; Jiang, D; Yao, JF; Zheng, S
Citation	Annals Of Statistics, 2009, v. 37 n. 6 B, p. 3822-3840
Issued Date	2009
URL	http://hdl.handle.net/10722/132605
Rights	Creative Commons: Attribution 3.0 Hong Kong License

The Annals of Statistics
2009, Vol. 37, No. 6B, 3822–3840
DOI: 10.1214/09-AOS694

CORRECTIONS TO LRT ON LARGE-DIMENSIONAL COVARIANCE MATRIX BY RMT[1]

BY ZHIDONG BAI[2], DANDAN JIANG, JIAN-FENG YAO AND SHURONG ZHENG[3]

Northeast Normal University, National University of Singapore, IRMAR and Université de Rennes 1

In this paper, we give an explanation to the failure of two likelihood ratio procedures for testing about covariance matrices from Gaussian populations when the dimension p is large compared to the sample size n. Next, using recent central limit theorems for linear spectral statistics of sample covariance matrices and of random F-matrices, we propose necessary corrections for these LR tests to cope with high-dimensional effects. The asymptotic distributions of these corrected tests under the null are given. Simulations demonstrate that the corrected LR tests yield a realized size close to nominal level for both moderate p (around 20) and high dimension, while the traditional LR tests with χ^2 approximation fails.

Another contribution from the paper is that for testing the equality between two covariance matrices, the proposed correction applies equally for non-Gaussian populations yielding a valid pseudo-likelihood ratio test.

1. Introduction. The rapid development and wide application of computer techniques permits to collect and store a huge amount data, where the number of measured variables is usually large. Such high-dimensional data occur in many modern scientific fields, such as micro-array data in biology, stock market analysis in finance and wireless communication networks. Traditional estimation or test tools are no more valid, or perform badly for such high-dimensional data, since they typically assume a large sample size n with respect to the number of variables p. A better approach in this high-dimensional data setting would be based on an asymptotic theory where both n and p approach infinity. To illustrate this purpose, let us mention the case of Hotelling's T^2-test. The failure of T^2-test for high-dimensional data has been mentioned as early as by Dempster [5]. As a remedy, Dempster proposed a so-called nonexact test. However, the theoretical

---

Received September 2008; revised February 2009.

[1]This version contains a selected set of proofs. A longer version of the paper containing all the proofs is to be found at arXiv:0902.0552.

[2]Supported by CNSF Grant 10871036 and NUS Grant R-155-000-079-112.

[3]Supported by CNSF Grant 0701021 and MENU Grant STC07001.

AMS 2000 subject classifications. Primary 62H15; secondary 62H10.

Key words and phrases. High-dimensional data, testing on covariance matrices, Marčenko–Pastur distributions, random F-matrices.

justification of Dempster's test arises much later in [1] inspired by modern random matrix theory (RMT). These authors have found necessary correction for the T^2-test to compensate effects due to high dimension.

In this paper, we consider two LR tests concerning covariance matrices. We first give a theoretical explanation for the fail of these tests in high-dimensional data context. Next, with the aid of random matrix theory, we provide necessary corrections to these LR tests to cope with the high-dimensional effects.

First, we consider the problem of one-sample covariance hypothesis test. Suppose that $\mathbf{x}$ follows a p-dimensional Gaussian distribution $N(\mu_p, \Sigma_p)$ and we want to test

$$H_0 : \Sigma_p = I_p, \tag{1.1}$$

where I_p denotes the p-dimensional identity matrix. Note that testing $\Sigma_p = A$ with an arbitrary covariance matrix A can always be reduced to the above null hypothesis by the transformation $A^{-1/2}\mathbf{x}$.

Let $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ be a sample from $\mathbf{x}$, where we assume $p < n$. The sample covariance matrix is

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

and set

$$L^* = \operatorname{tr}\mathbf{S} - \log|\mathbf{S}| - p. \tag{1.3}$$

The likelihood ratio test statistic is

$$T_n = n \cdot L^*. \tag{1.4}$$

Keeping p fixed while letting $n \to \infty$, then the classical theory states that T_n converges to the $\chi^2_{1/2p(p+1)}$ distribution under H_0.

However, as will be shown, this classical approximation leads to a test size much higher than the nominal test level in the case of high-dimensional data because T_n approaches infinity for large p. As seen from Table 1 in Section 3, for dimension and sample sizes $(p, n) = (50, 500)$, the realized size of the test is 22.5% instead of the nominal 5% level. The result is even worse for the case $(p, n) = (300, 500)$, with a 100% test size.

Based on a recent CLT for linear spectral statistics (LSS) of large-dimensional sample covariance matrices [3], we construct a corrected version of T_n in Section 3. As shown by the simulation results of Section 3.1, the corrected test performs much better in case of high dimensions. Moreover, it also performs correctly for moderate dimensions like $p = 10$ or 20. For dimension and sample sizes (p, n) cited above, the sizes of the corrected test are 5.9% and 5.2%, respectively, both close to the 5% nominal level.

The second test problem we consider is about the equality between two high-dimensional covariance matrices. Let $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{pi})^T, i = 1, \dots, n_1$ and $\mathbf{y}_j = (y_{1j}, y_{2j}, \dots, y_{pj})^T,\ j = 1, \dots, n_2$ be observations from two p-dimensional

TABLE 1

Sizes and powers of the traditional LRT and the corrected LRT, based on 10,000 independent replications with real Gaussian variables. Powers are estimated under the alternative $\Sigma_p = \mathrm{diag}(1, 0.05, 0.05, 0.05, \ldots)$

	CLRT			LRT	
(p, n)	**Size**	**Difference with 5%**	**Power**	**Size**	**Power**
(5, 500)	0.0803	0.0303	0.6013	0.0521	0.5233
(10, 500)	0.0690	0.0190	0.9517	0.0555	0.9417
(50, 500)	0.0594	0.0094	1	0.2252	1
(100, 500)	0.0537	0.0037	1	0.9757	1
(300, 500)	0.0515	0.0015	1	1	1

normal populations $N(\mu_k, \Sigma_k)$, $k = 1, 2$, respectively. We wish to test the null hypothesis

$$H_0 : \Sigma_1 = \Sigma_2. \tag{1.5}$$

The related sample covariance matrices are

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

where $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ are the respective sample means. Let

$$L_1 = \frac{|A|^{n_1/2} \cdot |B|^{n_2/2}}{|c_1 A + c_2 B|^{N/2}}, \tag{1.6}$$

where $N = n_1 + n_2$ and c_k denote $\frac{n_k}{N}$, $k = 1, 2$. The likelihood ratio test statistic is

$$T_N = -2\log L_1,$$

and when $n_1, n_2 \to \infty$, we get

$$T_N = -2\log L_1 \Rightarrow \chi^2_{1/2p(p+1)} \tag{1.7}$$

under H_0. Of course, in this limit scheme, the data dimension p is held fixed.

However, employing this χ^2 limit distribution for dimensions like 30 or 40, increases dramatically the size of the test. For instance, simulations in Section 4.1 show that, for dimension and sample sizes $(p, n_1, n_2) = (40, 800, 400)$, the test size equals 21.2% instead of the nominal 5% level. The result is worse for the case of $(p, n_1, n_2) = (80, 1600, 800)$, leading to a 49.5% test size. The reason for this failure of the classical LR test is the following. Modern RMT indicates that when both dimension and sample size are large, the likelihood ratio statistic T_N drifts to infinity almost surely. Therefore, the classical χ^2 approximation leads to many false rejections of H_0 in case of high-dimensional data.

Based on recent CLT for linear spectral statistics of F-matrices from RMT, we propose a correction to this LR test in Section 4. Although this corrected test is constructed under the asymptotic scheme $n_1 \wedge n_2 \to +\infty$, $y_{n_1} = p/n_1 \to y_1 \in (0, 1)$, $y_{n_2} = p/n_2 \to y_2 \in (0, 1)$, simulations demonstrate an overall correct behavior including small or moderate dimensions p. For example, for the above cited dimension and sample sizes (p, n_1, n_2), the sizes of the corrected test equal 5.6% and 5.2%, respectively, both close to the nominal 5% level.

Related work include Ledoit and Wolf [6], Schott [8] and Srivastava [9]. These authors propose several procedures in the high-dimensional setting for testing that (i) a covariance matrix is an identity matrix, proportional to an identity matrix (spherecity) and is a diagonal matrix or (ii) several covariance matrices are equal. These procedures have the following common feature: their construction involves some well-chosen distance function between the null and the alternative hypotheses and rely on the first two spectral moments, namely the statistics $\operatorname{tr} S_k$ and $\operatorname{tr} S_k^2$ from sample covariance matrices S_k. Therefore, the procedures proposed by these authors are different from the likelihood-based procedures we consider here. Another important difference concerns the Gaussian assumption on the random variables used in all these references. Actually, for testing the equality between two covariance matrices, the correction proposed in this paper applies equally for non-Gaussian and high-dimensional data leading to a valid pseudo-likelihood test.

The rest of the paper is organized as following. Preliminary and useful RMT results are recalled in Section 2. In Sections 3 and 4, we introduce our results for the two tests above. A selected set of proofs and technical derivations is postponed to the last section.

2. Useful results from the random matrix theory. We first recall several results from RMT, which will be useful for our corrections to tests. For any $p \times p$ square matrix M with real eigenvalues (λ_i^M), F_n^M denotes the empirical spectral distribution (ESD) of M, that is,

$$F_n^M(x) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}_{\lambda_i^M \le x}, \qquad x \in \mathbb{R}.$$

We will consider random matrix M whose ESD F_n^M converges (in a sense to be more precise) to a limiting spectral distribution (LSD) F^M. To make statistical inference about a parameter $\theta = \int f(x)\, dF^M(x)$, it is natural to use the estimator

$$\widehat{\theta} = \int f(x)\, dF_n^M(x) = \frac{1}{p} \sum_{i=1}^{p} f(\lambda_i^M),$$

which is a so-called linear spectral statistic (LSS) of the random matrix M.

2.1. *CLT for LSS of a high-dimensional sample covariance matrix.* Let $\{\xi_{ki} \in \mathbb{C}, i, k = 1, 2, \dots\}$ be a double array of i.i.d. complex variables with mean 0 and variance 1. Set $\xi_i = (\xi_{1i}, \xi_{2i}, \dots, \xi_{pi})^T$, the vectors $\xi_1, \dots, \xi_n$ are considered as an i.i.d. sample from some p-dimensional distribution with mean 0_p and covariance matrix I_p. Therefore, the sample covariance matrix is

$$S_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i \xi_i^*. \tag{2.1}$$

For $0 < \theta \le 1$, let $a(\theta) = (1-\sqrt{\theta})^2$ and $b(\theta) = (1+\sqrt{\theta})^2$. The Marčenko–Pastur distribution of index θ, denoted as F^θ, is the distribution on $[a(\theta), b(\theta)]$ with the following density function

$$g_\theta(x) = \frac{1}{2\pi\theta x}\sqrt{[b(\theta)-x][x-a(\theta)]}, \qquad a(\theta) \le x \le b(\theta).$$

Let

$$y_n = \frac{p}{n} \to y \in (0,1)$$

and F^y, F^{y_n} be the Marčenko–Pastur law of index y and y_n, respectively. Let $\mathcal{U}$ be an open set of the complex plane, including $[I_{(0,1)}(y)a(y), b(y)]$, and $\mathcal{A}$ be the set of analytic functions $f : \mathcal{U} \mapsto \mathbb{C}$. We consider the empirical process $G_n := \{G_n(f)\}$ indexed by $\mathcal{A}$,

$$G_n(f) = p \cdot \int_{-\infty}^{+\infty} f(x)[F_n - F^{y_n}](dx), \qquad f \in \mathcal{A}, \tag{2.2}$$

where F_n is the ESD of S_n. The following theorem will play a fundamental role in next derivations, which is a specialization of a general theorem from Bai and Silverstein [3] (Theorem 1.1).

THEOREM 2.1. *Assume that $f_1, \dots, f_k \in \mathcal{A}$, and $\{\xi_{ij}\}$ are i.i.d. random variables, such that $E\xi_{11} = 0$, $E|\xi_{11}|^2 = 1$, $E|\xi_{11}|^4 < \infty$. Moreover, $\frac{p}{n} \to y \in (0,1)$ as $n, p \to \infty$.*

We then get the following cases.

(i) Real case. *Assume $\{\xi_{ij}\}$ are real and $E(\xi_{11}^4) = 3$. Then the random vector $(G_n(f_1), \dots, G_n(f_k))$ weakly converges to a k-dimensional Gaussian vector with mean vector,*

$$\begin{aligned} m(f_j) = {} & \frac{f_j(a(y)) + f_j(b(y))}{4} \\ & - \frac{1}{2\pi}\int_{a(y)}^{b(y)} \frac{f_j(x)}{\sqrt{4y-(x-1-y)^2}}\,dx, \qquad j = 1, \dots, k, \end{aligned} \tag{2.3}$$

and covariance function

$$\begin{aligned} \upsilon(f_j, f_\ell) = -\frac{1}{2\pi^2}\oint\oint \frac{f_j(z_1) f_\ell(z_2)}{(\underline{m}(z_1)-\underline{m}(z_2))^2}\,d\underline{m}(z_1)\,d\underline{m}(z_2), \\ j, \ell \in \{1, \dots, k\}, \end{aligned} \tag{2.4}$$

where $\underline{m}(z) \equiv m_{\underline{F}^y}(z)$ is the Stieltjes Transform of $\underline{F}^y \equiv (1-y)I_{[0,\infty)} + yF^y$. The contours in (2.4) are nonoverlapping and both contain the support of F^y.

(ii) Complex case. *Assume* $\{\xi_{ij}\}$ *are complex and* $E\xi_{11}^2 = 0$, $E(|\xi_{11}|^4) = 2$. *Then the conclusion of* (i) *also holds*, *except the mean vector is zero and the covariance function is half of the function given in* (2.4).

It is worth noticing that Theorem 1.1 in Bai and Silverstein [3] covers more general sample covariance matrices of form $S_n' = T_n^{1/2} S_n T_n^{1/2}$ where (T_n) is a given sequence of positive-definite Hermitian matrices. In the "white" case, $T_n \equiv I$ as considered here, in a recent preprint Pastur and Lytova [7], the authors offer a new extension of the CLT where the constraints $E|\xi_{11}|^4 = 3$ or 2, as stated above, are removed.

2.2. *CLT for LSS of high-dimensional F matrix*. Let $\{\xi_{ki} \in \mathbb{C}, i, k = 1, 2, \ldots\}$ and $\{\eta_{kj} \in \mathbb{C}, j, k = 1, 2, \ldots\}$ are two independent double arrays of i.i.d. complex variables with mean 0 and variance 1. Write $\xi_i = (\xi_{1i}, \xi_{2i}, \ldots, \xi_{pi})^T$ and $\eta_j = (\eta_{1j}, \eta_{2j}, \ldots, \eta_{pj})^T$. Also, for any positive integers n_1, n_2, the vectors $(\xi_1, \ldots, \xi_{n_1})$ and $(\eta_1, \ldots, \eta_{n_2})$ can be thought as independent samples of size n_1 and n_2, respectively, from some p-dimensional distributions. Let S_1 and S_2 be the associated sample covariance matrices, that is,

$$S_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \xi_i \xi_i^* \quad \text{and} \quad S_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \eta_j \eta_j^*.$$

Then the following so-called F-matrix generalizes the classical Fisher statistics for the present p-dimensional case,

$$V_n = S_1 S_2^{-1}, \tag{2.5}$$

where $n_2 > p$. Here, we use the notation $n = (n_1, n_2)$.

Let

$$y_{n_1} = \frac{p}{n_1} \to y_1 \in (0, 1), \qquad y_{n_2} = \frac{p}{n_2} \to y_2 \in (0, 1). \tag{2.6}$$

Under suitable moment conditions, the ESD $F_n^{V_n}$ of V_n has a LSD F_{y_1, y_2}, which has a density (see page 72 of [4]), given by

$$\ell(x) = \begin{cases} \dfrac{(1 - y_2)\sqrt{(b - x)(x - a)}}{2\pi x(y_1 + y_2 x)}, & a \le x \le b, \\ 0, & \text{otherwise,} \end{cases} \tag{2.7}$$

where $a = (1 - y_2)^{-2}(1 - \sqrt{y_1 + y_2 - y_1 y_2})^2$ and $b = (1 - y_2)^{-2}(1 + \sqrt{y_1 + y_2 - y_1 y_2})^2$.

Similar to previously, let $\widetilde{\mathcal{U}}$ be an open set of the complex plane, including the interval

$$\left[I_{(0,1)}(y_1) \frac{(1 - \sqrt{y_1})^2}{(1 + \sqrt{y_2})^2}, \frac{(1 + \sqrt{y_1})^2}{(1 - \sqrt{y_2})^2} \right]$$

and $\widetilde{\mathcal{A}}$ be the set of analytic functions $f:\widetilde{\mathcal{U}} \mapsto \mathbb{C}$. Define the empirical process $\widetilde{G}_n := \{\widetilde{G}_n(f)\}$ indexed by $\widetilde{\mathcal{A}}$

$$\widetilde{G}_n(f) = p \cdot \int_{-\infty}^{+\infty} f(x)[F_n^{V_n} - F_{y_{n_1},y_{n_2}}](dx), \qquad f \in \widetilde{\mathcal{A}}. \tag{2.8}$$

Here, $F_{y_{n_1},y_{n_2}}$ is the limiting distribution in (2.7), but with y_{n_k} instead of y_k, $k = 1, 2$.

Recently, Zheng [10] establishes a general CLT for LSS of large-dimensional F matrix. The following theorem is a simplified one quoted from it, which will play an important role.

THEOREM 2.2. *Let $f_1, \dots, f_k \in \widetilde{\mathcal{A}}$, and assume the following: for each p, (ξ_{ij_1}) and (η_{ij_2}) variables are i.i.d., $1 \le i \le p, 1 \le j_1 \le n_1, 1 \le j_2 \le n_2$. $E\xi_{11} = E\eta_{11} = 0$, $E|\xi_{11}|^4 = E|\eta_{11}|^4 < \infty$, $y_{n_1} = \frac{p}{n_1} \to y_1 \in (0,1)$, $y_{n_2} = \frac{p}{n_2} \to y_2 \in (0,1)$. Then:*

(i) Real case. *Assume (ξ_{ij}) and (η_{ij}) are real, $E|\xi_{11}|^2 = E|\eta_{11}|^2 = 1$, then the random vector $(\widetilde{G}_n(f_1), \dots, \widetilde{G}_n(f_k))$ weakly converges to a k-dimensional Gaussian vector with the mean vector*

$$m(f_j) = \lim_{r \to 1_+} [(2.9) + (2.10) + (2.11)],$$

$$\frac{1}{4\pi i} \oint_{|\zeta|=1} f_j(z(\zeta)) \left[\frac{1}{\zeta - 1/r} + \frac{1}{\zeta + 1/r} - \frac{2}{\zeta + y_2/(hr)}\right] d\zeta \tag{2.9}$$

$$+ \frac{\beta \cdot y_1(1-y_2)^2}{2\pi i \cdot h^2} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{1}{(\zeta + y_2/(hr))^3} d\zeta \tag{2.10}$$

$$+ \frac{\beta \cdot y_2(1-y_2)}{2\pi i \cdot h} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{\zeta + 1/(hr)}{(\zeta + y_2/(hr))^3} d\zeta, \qquad j = 1, \dots, k, \tag{2.11}$$

where $z(\zeta) = (1-y_2)^{-2}[1 + h^2 + 2h\mathcal{R}(\zeta)]$, $h = \sqrt{y_1 + y_2 - y_1 y_2}$, $\beta = E|\xi_{11}|^4 - 3$, and the covariance function as $1 < r_1 < r_2 \downarrow 1$

$$\upsilon(f_j, f_\ell) = \lim_{1 < r_1 < r_2 \to 1_+} [(2.12) + (2.13)],$$

$$-\frac{1}{2\pi^2} \oint_{|\zeta_2|=1} \oint_{|\zeta_1|=1} \frac{f_j(z(r_1\zeta_1)) f_\ell(z(r_2\zeta_2)) r_1 r_2}{(r_2\zeta_2 - r_1\zeta_1)^2} d\zeta_1\, d\zeta_2, \tag{2.12}$$

$$-\frac{\beta \cdot (y_1 + y_2)(1-y_2)^2}{4\pi^2 h^2} \oint_{|\zeta_1|=1} \frac{f_j(z(\zeta_1))}{(\zeta_1 + y_2/(hr_1))^2} d\zeta_1 \times \oint_{|\zeta_2|=1} \frac{f_\ell(z(\zeta_2))}{(\zeta_2 + y_2/(hr_2))^2} d\zeta_2, \qquad j, \ell \in \{1, \dots, k\}. \tag{2.13}$$

(ii) Complex case. *Assume* (ξ_{ij}) *and* (η_{ij}) *are complex*, $E(\xi_{11}^2) = E(\eta_{11}^2) = 0$, *then the conclusion of* (i) *also holds*, *except the means are* $\lim_{r \to 1_+}[(2.10) + (2.11)]$ *and the covariance function is* $\lim_{1<r_1<r_2 \to 1_+}[\frac{1}{2} \cdot (2.12) + (2.13)]$, *where* $\beta = E|\xi_{11}|^4 - 2$.

We should point out that Zheng's CLT for F-matrices covers more general situations then those cited in Theorem 2.2. In particular, the fourth-moments $E|\xi_{11}|^4$ and $E|\eta_{11}|^4$ can be different.

The following lemma will be used in Section 4 for an application of Theorem 2.2 to obtain the formulas (4.5) and (4.6).

LEMMA 2.1. *For the function* $f(x) = \log(a + bx), x \in \mathbb{R}, a, b > 0$, *let* (c, d) *be the unique solution to the equations*

$$\begin{cases} c^2 + d^2 = a(1 - y_2)^2 + b(1 + h^2), \\ cd = bh, \\ 0 < d < c. \end{cases}$$

Analogously, let γ, η *be the constants similar to* (c, d) *but for the function* $g(x) = \log(\alpha + \beta x), \alpha > 0, \beta > 0$. *Then the mean and covariance functions in* (2.9) *and* (2.12) *equal to*

$$m(f) = \frac{1}{2} \log \frac{(c^2 - d^2)h^2}{(ch - y_2 d)^2},$$

$$\upsilon(f, g) = 2bhd^{-1}c^{-1} \log \frac{c\gamma}{c\gamma - d\eta}.$$

3. Testing the hypothesis that a high-dimensional covariance matrix is equal to a given matrix. To test the hypothesis $H_0 : \Sigma_p = I_p$, let be the sample covariance matrix $\mathbf{S}$ and likelihood ratio statistic T_n as defined in (1.2) and (1.4), respectively. For $\xi_i = \mathbf{x}_i - \mu_p$, the array $\{\xi_i\}_{i=1,\dots,n}$ contains p-dimensional standard normal variables under H_0. Let

$$\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i \xi_i^*$$

and

$$\widetilde{L}^* = \operatorname{tr} \mathbf{S}_n - \log|\mathbf{S}_n| - p.$$

THEOREM 3.1. *Assuming that the conditions of Theorem* 2.1 *hold*, L^* *is defined as* (1.3) *and* $g(x) = x - \log x - 1$. *Then under* H_0 *and when* $n \to \infty$

$$\widetilde{T}_n = \upsilon(g)^{-1/2}[L^* - p \cdot F^{y_n}(g) - m(g)] \Rightarrow N(0, 1), \tag{3.1}$$

where F^{y_n} *is the Marčenko–Pastur law of index* y_n.

PROOF. Because the difference between $\mathbf{S}$ and $\mathbf{S}_n$ is a rank-1 matrix, $\mathbf{S}$ and $\mathbf{S}_n$ have the same LSD. So, L^* and $\widetilde{L}^*$ have the same asymptotic distribution. We also have

$$\begin{aligned}\widetilde{L}^* &= \operatorname{tr}\mathbf{S}_n - \log|\mathbf{S}_n| - p\\ &= \sum_{i=1}^{p}(\lambda_i^{\mathbf{S}_n} - \log\lambda_i^{\mathbf{S}_n} - 1) = p\cdot\int(x - \log x - 1)\,dF_n(x)\\ &= p\cdot\int g(x)\,d\bigl(F_n(x) - F^{y_n}(x)\bigr) + p\cdot F^{y_n}(g),\end{aligned}$$

so that

$$G_n(g) = \widetilde{L}^* - p\cdot F^{y_n}(g). \tag{3.2}$$

By Theorem 2.1, $G_n(g)$ weakly converges to a Gaussian vector with the mean

$$m(g) = -\frac{\log(1-y)}{2} \tag{3.3}$$

and variance

$$\upsilon(g) = -2\log(1-y) - 2y \tag{3.4}$$

for the real case, which are calculated in Section 5. For the complex case, the mean $m(g)$ is zero and the variance is half of $\upsilon(g)$. Then by (3.2), we arrive at

$$\widetilde{L}^* - p\cdot F^{y_n}(g) \Rightarrow N(m(g), \upsilon(g)), \tag{3.5}$$

where

$$F^{y_n}(g) = 1 - \frac{y_n - 1}{y_n}\log(1 - y_n) \tag{3.6}$$

can be calculated by the density of LSD of sample covariance matrix in Section 5. Because $\widetilde{L}^*$ and L^* have the same asymptotic distribution and (3.5), finally we get

$$\widetilde{T}_n = \upsilon(g)^{-1/2}[L^* - p\cdot F^{y_n}(g) - m(g)] \Rightarrow N(0,1). \qquad \square$$

3.1. *Simulation study* I. For different values of (p,n), we compute the realized sizes of traditional likelihood ratio test (LRT) and the corrected likelihood ratio test (CLRT) proposed previously. The nominal test level is set to be $\alpha = 0.05$, and for each (p,n), we run 10,000 independent replications with real Gaussian variables. Results are given in Table 1 and Figure 1 below.

As seen in Table 1, the traditional LRT always rejects H_0 when p is large, like $p = 100$ or 300, while the sizes produced by the corrected LRT perfectly matches the nominal level. For moderate dimensions like $p = 50$, the corrected LRT still performs correctly while the traditional LRT has a size much higher than 5%.

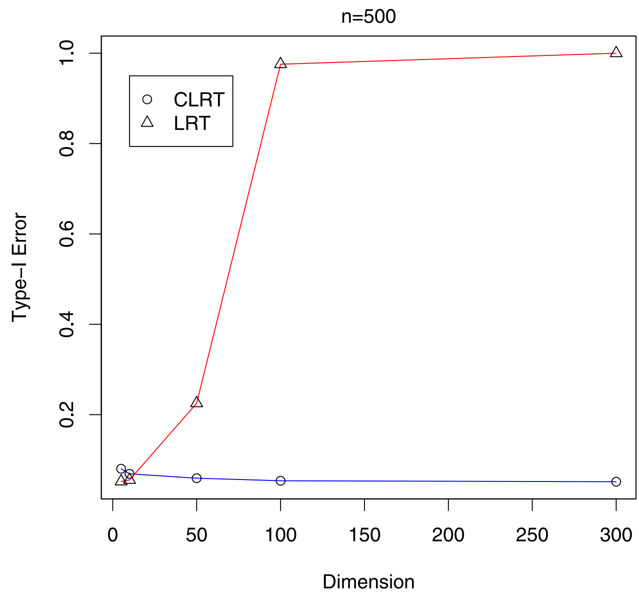

FIG. 1. *Realized sizes of the traditional LRT and the corrected LRT for different dimensions p with real Gaussian variables*. 10,000 *independent runs with* 5% *nominal level and sample size* $n = 500$.

4. Testing the equality of two high-dimensional covariance matrices. Let $(\mathbf{x}_i)$, $i = 1, \dots, n_1$ and $(\mathbf{y}_j)$, $j = 1, \dots, n_2$ be observations from two normal populations $N(\mu_k, \Sigma_k)$, $k = 1, 2$, respectively. We examine the test defined in (1.5) and (1.6). The aim is to find a good scaling of the LR statistic T_N, such that the scaled statistic weakly converges to some limiting distribution. Let

$$\xi_i = \Sigma^{-1/2}(\mathbf{x}_i - \mu_1), \qquad \eta_i = \Sigma^{-1/2}(\mathbf{y}_i - \mu_2),$$

where $\Sigma = \Sigma_1 = \Sigma_2$ denotes the common covariance matrix under H_0. Note that in a strict sense, the vectors $(\mathbf{x}_i)$, $(\mathbf{y}_i)$ and the matrices $\Sigma, \Sigma_1, \Sigma_2$ depend on p. However, we do not signify this dependence in notation for ease of statements. Due to Gaussian assumption, the arrays $(\xi_i)_{i=1,\dots,n_1}$ and $(\eta_j)_{j=1,\dots,n_2}$ contain i.i.d. $N(0, 1)$ variables, for which we can apply Theorem 2.2.

Let

$$S_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \xi_i \xi_i^* = \Sigma^{-1/2} C \Sigma^{-1/2},$$

$$S_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \eta_j \eta_j^* = \Sigma^{-1/2} D \Sigma^{-1/2},$$

where

$$C = \frac{1}{n_1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^*,$$

$$D = \frac{1}{n_2} \sum_{j=1}^{n_2} (\mathbf{y}_j - \mu_2)(\mathbf{y}_j - \mu_2)^*.$$

Note that

$$V_n = S_1 S_2^{-1}$$

forms a random F-matrix and we have

$$\widetilde{L}_1 = \frac{|S_1|^{n_1/2} \cdot |S_2|^{n_2/2}}{|c_1 S_1 + c_2 S_2|^{N/2}} = \frac{|C|^{n_1/2} \cdot |D|^{n_2/2}}{|c_1 C + c_2 D|^{N/2}}. \tag{4.1}$$

THEOREM 4.1. *Assuming that the conditions of Theorem 2.2 hold under H_0, L_1 as defined in (1.6) and*

$$f(x) = \log(y_{n_1} + y_{n_2} x) - \frac{y_{n_2}}{y_{n_1} + y_{n_2}} \log x - \log(y_{n_1} + y_{n_2}).$$

Then under H_0 and as $n_1 \wedge n_2 \to \infty$,

$$\widetilde{T}_N = \upsilon(f)^{-1/2} \left[-\frac{2 \log L_1}{N} - p \cdot F_{y_{n_1}, y_{n_2}}(f) - m(f) \right] \Rightarrow N(0, 1). \tag{4.2}$$

PROOF. As $A - C$ and $B - D$ are rank-1 random matrices, AB^{-1} and CD^{-1} have the same LSD. Also by (4.1), $\widetilde{L}_1$ and L_1 have the same asymptotic distribution. Because

$$\begin{aligned}
-\frac{2}{N} \log \widetilde{L}_1 &= -\frac{2}{N} \log \left(\frac{|S_1|^{n_1/2} \cdot |S_2|^{n_2/2}}{|c_1 S_1 + c_2 S_2|^{N/2}} \right) \\
&= \log |c_1 V_n^{-1} + c_2| - c_1 \cdot \log |V_n^{-1}| \\
&= \sum_{i=1}^{p} \log(c_1 \lambda_i^{V_n} + c_2) - c_1 \cdot \log(\lambda_i^{V_n}) \\
&= p \cdot \int [\log(c_1 x + c_2) - c_1 \cdot \log(x)] \, dF_n^{V_n}(x).
\end{aligned}$$

Define $f(x) = \log(c_1 x + c_2) - c_1 \cdot \log(x)$, with $c_1 = \frac{n_1}{N} = \frac{y_{n_2}}{y_{n_1} + y_{n_2}}$ and $c_2 = \frac{n_2}{N} = \frac{y_{n_1}}{y_{n_1} + y_{n_2}}$; $f(x)$ can also be written as

$$f(x) = \log(y_{n_1} + y_{n_2} x) - \frac{y_{n_2}}{y_{n_1} + y_{n_2}} \log x - \log(y_{n_1} + y_{n_2}). \tag{4.3}$$

From

$$\begin{aligned}
-\frac{2 \log \widetilde{L}_1}{N} &= p \cdot \int f(x) \, dF_n^{V_n}(x) \\
&= p \cdot \int f(x) \, d\big(F_n^{V_n}(x) - F_{y_{n_1}, y_{n_2}}(x)\big) + p \cdot F_{y_{n_1}, y_{n_2}}(f),
\end{aligned}$$

we get

$$\widetilde{G}_n(f) = -\frac{2\log \widetilde{L}_1}{N} - p \cdot F_{y_{n_1}, y_{n_2}}(f). \tag{4.4}$$

By Theorem 2.2, $\widetilde{G}_n(f)$ weakly converges to a Gaussian vector with mean

$$\begin{aligned} m(f) = \frac{1}{2}\biggl[\log\biggl(\frac{y_1 + y_2 - y_1 y_2}{y_1 + y_2}\biggr) - \frac{y_1}{y_1 + y_2}\log(1 - y_2) \\ - \frac{y_2}{y_1 + y_2}\log(1 - y_1)\biggr] \end{aligned} \tag{4.5}$$

and variance

$$\begin{aligned} \upsilon(f) = -\frac{2y_2^2}{(y_1 + y_2)^2}\log(1 - y_1) - \frac{2y_1^2}{(y_1 + y_2)^2}\log(1 - y_2) \\ - 2\log\frac{y_1 + y_2}{y_1 + y_2 - y_1 y_2} \end{aligned} \tag{4.6}$$

for the real case. For the complex case, the mean $m(f)$ is zero and the variance is half of $\upsilon(f)$. In other words,

$$-\frac{2\log \widetilde{L}_1}{N} - p \cdot F_{y_{n_1}, y_{n_2}}(f) \Rightarrow N(m(f), \upsilon(f)), \tag{4.7}$$

where

$$\begin{aligned} F_{y_{n_1}, y_{n_2}}(f) = \frac{-(y_{n_1} + y_{n_2} - y_{n_1} y_{n_2})}{y_{n_1} y_{n_2}} \log(y_{n_1} + y_{n_2} - y_{n_1} y_{n_2}) \\ + \frac{(y_{n_1} + y_{n_2} - y_{n_1} y_{n_2})}{y_{n_1} y_{n_2}} \log(y_{n_1} + y_{n_2}) \\ + \frac{y_{n_1}(1 - y_{n_2})}{y_{n_2}(y_{n_1} + y_{n_2})} \log(1 - y_{n_2}) \\ + \frac{y_{n_2}(1 - y_{n_1})}{y_{n_1}(y_{n_1} + y_{n_2})} \log(1 - y_{n_1}). \end{aligned}$$

Because $\widetilde{L}_1$ and L_1 have the same asymptotic distribution and by (4.7), we get by letting $n_1 \wedge n_2 \to \infty$,

$$\widetilde{T}_N = \upsilon(f)^{-1/2}\biggl[-\frac{2\log L_1}{N} - p \cdot F_{y_{n_1}, y_{n_2}}(f) - m(f)\biggr] \Rightarrow N(0, 1). \qquad \square$$

4.1. *Simulation study* II. For different values of (p, n_1, n_2), we compute the realized sizes of the traditional LRT and the corrected LRT with 10,000 independent replications. The nominal test level is $\alpha = 0.05$ and we use real Gaussian variables. Results are summarized in Table 2 and Figure 2.

TABLE 2

Sizes and powers of the traditional LRT and the corrected LRT based on 10,000 *independent replications using real Gaussian variables. Powers are estimated under the alternative* $\Sigma_1\Sigma_2^{-1} = \mathrm{diag}(3, 1, 1, 1, \ldots)$. *Upper*: $y_1 = y_2 = 0.05$. *Bottom*: $y_1 = 0.05$, $y_2 = 0.1$

	CLRT			LRT	
(p, n_1, n_2)	**Size**	**Difference with 5%**	**Power**	**Size**	**Power**
		$(y_1, y_2) = (0.05, 0.05)$			
(5, 100, 100)	0.0770	0.0270	1	0.0582	1
(10, 200, 200)	0.0680	0.0180	1	0.0684	1
(20, 400, 400)	0.0593	0.0093	1	0.0872	1
(40, 800, 800)	0.0526	0.0026	1	0.1339	1
(80, 1600, 1600)	0.0501	0.0001	1	0.2687	1
(160, 3200, 3200)	0.0491	−0.0009	1	0.6488	1
(320, 6400, 6400)	0.0447	−0.0053	0.9671	1	1
		$(y_1, y_2) = (0.05, 0.1)$			
(5, 100, 50)	0.0781	0.0281	0.9925	0.0640	0.9849
(10, 200, 100)	0.0617	0.0117	0.9847	0.0752	0.9904
(20, 400, 200)	0.0573	0.0073	0.9775	0.1104	0.9938
(40, 800, 400)	0.0561	0.0061	0.9765	0.2115	0.9975
(80, 1600, 800)	0.0521	0.0021	0.9702	0.4954	0.9998
(160, 3200, 1600)	0.0520	0.0020	0.9702	0.9433	1
(320, 6400, 3200)	0.0510	0.0010	1	0.9939	1

As we can see, when the dimension p increases, the traditional LRT leads to a dramatically high test size while the corrected LRT remains accurate. Furthermore, for moderate dimensions like $p = 20$ or 40, the sizes of the traditional LRT are much higher than 5%, whereas the ones of corrected LRT are very close. By a closer look at the column showing the difference with 5%, we note that this difference rapidly decreases as p increases for the corrected test. Figure 2 gives a vivid sight of these comparisons between the traditional LRT and the corrected LRT in term of test sizes.

4.2. *A pseudo-likelihood test for high-dimensional non-Gaussian data.* As said in the introduction, previous related works as Ledoit and Wolf [6], Srivastava [9] or Schott [8] all assume Gaussian variables. In contrast, Theorem 4.1 applies for general distributions having a fourth moment. For these non-Gaussian data, we consider the corrected LRT as generalized pseudo-likelihood ratio test (or Gaussian LRT).

Moreover, the methods proposed by these authors all rely on an appropriate normalization of the trace of squared difference between two sample covariances following the idea of Bai and Saranadasa [1]. We believe that their method would

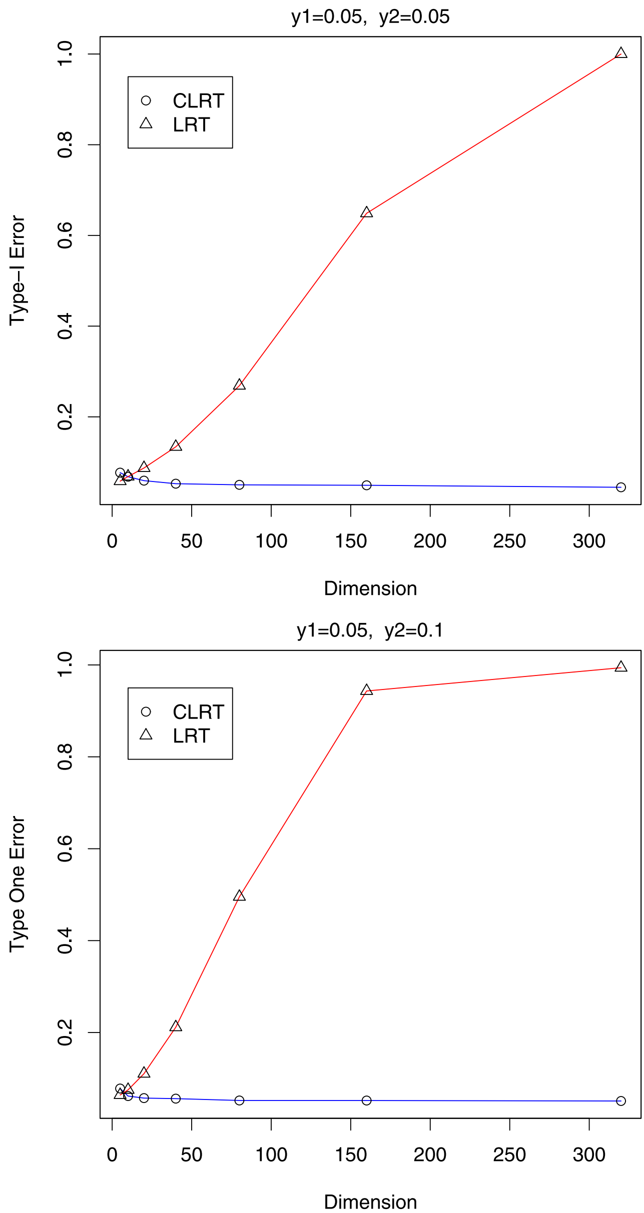

FIG. 2. *Sizes of the traditional LRT and the corrected LRT based on* 10,000 *independent replications using real Gaussian variables*. *Left*: $y_1 = y_2 = 0.05$. *Right*: $y_1 = 0.05$, $y_2 = 0.1$.

strongly depend on the normality assumption (what will be supported by simulation results below). On the other hand, based on general understanding, the LRT contains much higher information from data and its poor performance observed

TABLE 3
Sizes of the corrected pseudo-likelihood ration test and Schott's test for the case of $y_1 = 0.1$, $y_2 = 0.05$, based on 1000 independent replications with normalized t-distributed variables with 5 degrees of freedom

(p, n_1, n_2)	CLRT size	Schott's size
	$(y_1, y_2) = (0.05, 0.1)$	
(10, 100, 200)	0.067	0.517
(20, 200, 400)	0.065	0.603
(40, 400, 800)	0.054	0.703
(80, 800, 1600)	0.048	0.764
(160, 1600, 3200)	0.045	0.826
(320, 3200, 6400)	0.051	0.854

up to now is just caused by its large bias when dimension is large. Thus, from the intuitive understanding, we are confined ourselves to modify the LRT.

Let us develop an example in more detail. Assume that $\mathbf{x}$ follows a normalized t-distribution with 5 degree of freedom, that is, $\mathbf{x} = \sqrt{\frac{3}{5}}t(5)$, $\mathbf{x}$ and $\mathbf{y}$ are i.i.d., hence, $E\mathbf{x} = E\mathbf{y} = 0$, $E|\mathbf{x}|^2 = E|\mathbf{y}|^2 = 1$ and $E|\mathbf{x}|^4 = E|\mathbf{y}|^4 = 9$. We still employ the result in Theorem 4.1 for the test of equality between two covariance matrices, where

$$\begin{aligned} m_1(f) = \frac{1}{2}\biggl[&\log\left(\frac{y_1 + y_2 - y_1 y_2}{y_1 + y_2}\right) - \frac{y_1}{y_1 + y_2}\log(1 - y_2) \\ &- \frac{y_2}{y_1 + y_2}\log(1 - y_1) + \frac{6y_1^2 y_2}{(y_1 + y_2)^2} + \frac{6y_1 y_2^2}{(y_1 + y_2)^2}\biggr] \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} \upsilon_1(f) = &-\frac{2y_2^2}{(y_1 + y_2)^2}\log(1 - y_1) - \frac{2y_1^2}{(y_1 + y_2)^2}\log(1 - y_2) \\ &- 2\log\frac{y_1 + y_2}{y_1 + y_2 - y_1 y_2} \end{aligned} \tag{4.9}$$

instead of $m(f)$ and $\upsilon(f)$ for real case, respectively.

Table 3 summarizes a simulation study where we compare this corrected pseudo-LRT with the test proposed in Schott [8]. We use 1000 independent replications with the above t-distributed variables. Again, the nominal test level is $\alpha = 0.05$. As we can see, the corrected pseudo-LRT performs correctly while Schott's test is no more valid here since the variables are not Gaussian.

5. Selected proofs. To shorten the presentation of the paper, here we include only a selected set of proofs. The others, namely proofs of Lemma 2.1, (4.5) and (4.6), of the formula of $F_{y_{n_1},y_{n_2}}(f)$, (4.8) and (4.9) are to be found in a longer version of the paper at arXiv [2].

Proof of (3.3). By Theorem 2.1, for $g(x) = x - \log x - 1$, by using the variable change $x = 1 + y - 2\sqrt{y}\cos\theta$, $0 \le \theta \le \pi$, we have

$$\begin{aligned}
m(g) &= \frac{g(a(y)) + g(b(y))}{4} - \frac{1}{2\pi}\int_{a(y)}^{b(y)} \frac{g(x)}{\sqrt{4y - (x-1-y)^2}}\,dx \\
&= \frac{y - \log(1-y)}{2} \\
&\quad - \frac{1}{2\pi}\int_0^{\pi} \bigl[1 + y - 2\sqrt{y}\cos\theta - \log\bigl(1 + y - 2\sqrt{y}\cos\theta\bigr) - 1\bigr]\,d\theta \\
&= \frac{y - \log(1-y)}{2} - \frac{1}{4\pi}\int_0^{2\pi} \bigl[y - 2\sqrt{y}\cos\theta - \log\bigl|1 - \sqrt{y}e^{i\theta}\bigr|^2\bigr]\,d\theta \\
&= -\frac{\log(1-y)}{2},
\end{aligned}$$

where $\int_0^{2\pi} \log|1 - \sqrt{y}e^{i\theta}|^2\,d\theta = 0$ is calculated in [3].

Proof of (3.4). For $g(x) = x - \log x - 1$, by Theorem 2.1, we have

$$\upsilon(g) = -\frac{1}{2\pi^2}\oint\oint \frac{g(z_1)g(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2}\,d\underline{m}(z_1)\,d\underline{m}(z_2)$$

and

$$\begin{aligned}
g(z_1)g(z_2) = {}& z_1 z_2 - z_1\log z_2 - z_2\log z_1 + \log z_1\log z_2 \\
& - z_1 + \log z_1 - z_2 + \log z_2 + 1.
\end{aligned}$$

It is easy to see that $\upsilon(\mathbf{1},\mathbf{1}) = 0$, where $\mathbf{1}$ stands for the constant function equal to 1. For Stieltjes transform of F^y, the following equation is given in [3], for $z \in \mathbb{C}^+$:

$$z = -\frac{1}{\underline{m}(z)} + \frac{y}{1 + \underline{m}(z)}. \tag{5.1}$$

Let $m_i = \underline{m}(z_i)$, $i = 1, 2$. For fixed m_2, we have on a contour enclosing 1, $(y-1)^{-1}$ and -1, but not 0,

$$\begin{aligned}
\oint \frac{\log(z(m_1))}{(m_1 - m_2)^2}\,dm_1 &= \oint \frac{1/m_1^2 - y/(1+m_1)^2}{-1/m_1 + y/(1+m_1)}\frac{1}{(m_1 - m_2)}\,dm_1 \\
&= \oint \frac{(1+m_1)^2 - y m_1^2}{y m_1(m_1 - m_2)}\left(\frac{-1}{m_1+1} + \frac{1}{m_1 - 1/(y-1)}\right)dm_1 \\
&= 2\pi i \cdot \left(\frac{1}{m_2+1} - \frac{1}{m_2 - 1/(y-1)}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\oint \frac{-1/m_1 + y/(1+m_1)}{(m_1-m_2)^2}\,dm_1 \\
&\quad = y \oint \left(\frac{1}{1+m_1} + \frac{1-y}{y}\right) \cdot [1-(1+m_1)]^{-1} \cdot (m_2+1)^{-2} \\
&\qquad \times \left(1 - \frac{m_1+1}{m_2+1}\right)^{-2} dm_1 \\
&\quad = y \oint \left(\frac{1}{1+m_1} + \frac{1-y}{y}\right) \\
&\qquad \times \sum_{j=0}^{\infty} (1+m_1)^j (m_2+1)^{-2} \sum_{\ell=1}^{\infty} \ell \left(\frac{m_1+1}{m_2+1}\right)^{\ell-1} dm_1 \\
&\quad = 2\pi i \cdot \frac{y}{(m_2+1)^2}.
\end{aligned}$$

Then we also get $\upsilon(-z_1 + \log z_1, \mathbf{1}) = 0$. Similarly, $\upsilon(\mathbf{1}, -z_2 + \log z_2) = 0$. Furthermore,

$$\upsilon(z_1, z_2) = \frac{y^2}{\pi i} \oint \frac{1}{(m_2+1)^2} \left(\frac{1}{1+m_2} + \frac{1-y}{y}\right) \sum_{j=0}^{\infty} (1+m_2)^j \, dm_2 = 2y$$

and

$$\begin{aligned}
\upsilon(z_1, \log z_2) &= \frac{y}{\pi i} \oint \left(\frac{1}{m_2+1} - \frac{1}{m_2 - 1/(y-1)}\right)\left(\frac{1}{1+m_2} + \frac{1-y}{y}\right) \\
&\qquad \times [1-(1+m_2)]^{-1}\, dm_2 \\
&= \frac{y}{\pi i} \oint \left(\frac{1}{m_2+1} - \frac{1}{m_2 - 1/(y-1)}\right)\left(\frac{1}{1+m_2} + \frac{1-y}{y}\right) \\
&\qquad \times \sum_{j=0}^{\infty} (1+m_2)^j \, dm_2 = 2y.
\end{aligned}$$

By a computation in [3], we know that $\upsilon(\log z_1, \log z_2) = -2\log(1-y)$. Finally, we obtain

$$\begin{aligned}
\upsilon(g) &= \upsilon(z_1, z_2) + \upsilon(\log z_1, \log z_2) - 2\upsilon(z_1, \log z_2) \\
&\quad + \upsilon(-z_1 + \log z_1, \mathbf{1}) + \upsilon(\mathbf{1}, -z_2 + \log z_2) + \upsilon(\mathbf{1}, \mathbf{1}) \\
&= -2\log(1-y) - 2y.
\end{aligned}$$

Proof of (3.6). Since F^{y_n} is the Marčenko–Pastur law of index y_n, by using the variable change $x = 1 + y_n - 2\sqrt{y_n}\cos\theta, 0 \le \theta \le \pi$ we have

$$\begin{aligned} F^{y_n}(g) &= \int_{a(y_n)}^{b(y_n)} \frac{x - \log x - 1}{2\pi x y_n} \sqrt{(b(y_n) - x)(x - a(y_n))}\, dx \\ &= \frac{1}{2\pi y_n} \int_0^{\pi} \left[1 - \frac{\log(1 + y_n - 2\sqrt{y_n}\cos\theta) + 1}{1 + y_n - 2\sqrt{y_n}\cos\theta}\right] 4 y_n \sin^2\theta\, d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left[2\sin^2\theta - \frac{2\sin^2\theta}{1 + y_n - 2\sqrt{y_n}\cos\theta}(\log|1 - \sqrt{y_n} e^{i\theta}|^2 - 1)\right] d\theta \\ &= 1 - \frac{y_n - 1}{y_n}\log(1 - y_n), \end{aligned}$$

where

$$\begin{aligned} &\frac{1}{2\pi} \int_0^{2\pi} \frac{2\sin^2\theta}{1 + y_n - 2\sqrt{y_n}\cos\theta} \log|1 - \sqrt{y_n} e^{i\theta}|^2\, d\theta \\ &\quad = \frac{y_n - 1}{y_n}\log(1 - y_n) - 1 \end{aligned}$$

is calculated in [3].

REFERENCES

[1] BAI, Z. D. and SARANADASA, H. (1996). Effect of high dimension comparison of significance tests for a high-dimensional two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

[2] BAI, Z. D., JIANG, D., YAO, J.-F. and ZHENG, S. (2008). Corrections to LRT on large dimensional covariance matrix by RMT (full version). Preprint. Available at arXiv:0902.0552.

[3] BAI, Z. D. and SILVERSTEIN, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. MR2040792

[4] BAI, Z. D. and SILVERSTEIN, J. W. (2006). *Spectral Analysis of Large-Dimensional Random Matrices*, 1st ed. Science Press, Beijing.

[5] DEMPSTER, A. P. (1958). A high-dimensional two sample significance test. *Ann. Math. Statist.* **29** 995–1010. MR0112207

[6] LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. MR1926169

[7] PASTUR, L. and LYTOVA, A. (2008). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. Preprint. Available at arXiv:0809.4698v1. MR2461187

[8] SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample size. *Comput. Statist. Data Anal.* **51** 6535–6542. MR2408613

[9] SRIVASTAVA, M. S. (2005). Some tests concerning the covariance matrix in high-dimensional data. *J. Japan Statist. Soc.* **35** 251–272. MR2328427

[10] ZHENG, S. (2008). Central limit theorem for linear spectral statistics of large dimensional F-matrix. Preprint. Northeast Normal Univ., Changchun, China.

Z. BAI
D. JIANG
S. ZHENG
KLASMOE AND SCHOOL
OF MATHEMATICS AND STATISTICS
NORTHEAST NORMAL UNIVERSITY
5268 PEOPLE'S ROAD
130024 CHANGCHUN
CHINA
AND
DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
10, KENT RIDGE CRESCENT
SINGAPORE 119260
E-MAIL: stabaizd@nus.edu.sg
stajd@nus.edu.sg
zhengsr@nenu.edu.cn

J.-F. YAO
IRMAR AND UNIVERSITÉ DE RENNES 1
CAMPUS DE BEAULIEU
35042 RENNES CEDEX
FRANCE
E-MAIL: jian-feng.yao@univ-rennes1.fr