<table>
<thead>
<tr>
<th>Title</th>
<th>Corrections to LRT on large-dimensional covariance matrix by RMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Bai, Z; Jiang, D; Yao, JF; Zheng, S</td>
</tr>
<tr>
<td>Citation</td>
<td>Annals Of Statistics, 2009, v. 37 n. 6 B, p. 3822-3840</td>
</tr>
<tr>
<td>Issued Date</td>
<td>2009</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10722/132605">http://hdl.handle.net/10722/132605</a></td>
</tr>
<tr>
<td>Rights</td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
CORRECTIONS TO LRT ON LARGE-DIMENSIONAL COVARIANCE MATRIX BY RMT\textsuperscript{1}

BY ZHIDONG BAI\textsuperscript{2}, DANDAN JIANG, JIAN-FENG YAO AND SHURONG ZHENG\textsuperscript{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 $\chi^2$ 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
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 $x$ follows a $p$-dimensional Gaussian distribution $N(\mu_p, \Sigma_1p)$ and we want to test

$$H_0: \Sigma_1p = I_p,$$

where $I_p$ denotes the $p$-dimensional identity matrix. Note that testing $\Sigma_1p = A$ with an arbitrary covariance matrix $A$ can always be reduced to the above null hypothesis by the transformation $A^{-1/2}x$.

Let $(x_1, \ldots, x_n)$ be a sample from $x$, where we assume $p < n$. The sample covariance matrix is

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^*$$

and set

$$L^* = \text{tr}S - \log |S| - p.$$

The likelihood ratio test statistic is

$$T_n = n \cdot L^*.$$

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 $x_i = (x_{i1}, x_{i2}, \ldots, x_{pi})^T, i = 1, \ldots, n_1$ and $y_j = (y_{j1}, y_{j2}, \ldots, y_{pj})^T, j = 1, \ldots, n_2$ be observations from two $p$-dimensional
normal populations $N(\mu_k, \Sigma_k), k = 1, 2$, respectively. We wish to test the null hypothesis

$$H_0: \Sigma_1 = \Sigma_2.$$  

The related sample covariance matrices are

$$A = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \bar{x})(x_i - \bar{x})^*, \quad B = \frac{1}{n_2} \sum_{i=1}^{n_2} (y_i - \bar{y})(y_i - \bar{y})^*,$$

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

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

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

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

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

However, employing this $\chi^2$ 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 $\chi^2$ 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 con-

### 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 $\Sigma_p = \text{diag}(1, 0.05, 0.05, 0.05, \ldots)$

<table>
<thead>
<tr>
<th>$(p, n)$</th>
<th>Size</th>
<th>Difference with 5%</th>
<th>Power</th>
<th>CLRT Size</th>
<th>CLRT Power</th>
<th>LRT Size</th>
<th>LRT Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5, 500)</td>
<td>0.0803</td>
<td>0.0303</td>
<td>0.6013</td>
<td>0.0521</td>
<td>0.5233</td>
<td>0.0555</td>
<td>0.9417</td>
</tr>
<tr>
<td>(10, 500)</td>
<td>0.0690</td>
<td>0.0190</td>
<td>0.9517</td>
<td>0.0555</td>
<td>0.9417</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(50, 500)</td>
<td>0.0594</td>
<td>0.0094</td>
<td>1</td>
<td>0.2252</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(100, 500)</td>
<td>0.0537</td>
<td>0.0037</td>
<td>1</td>
<td>0.9757</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(300, 500)</td>
<td>0.0515</td>
<td>0.0015</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
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 (sphericity) 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 \( \text{tr} \ S_k \) and \( \text{tr} \ S_k^2 \) 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 \((\lambda^M_i)\), \( F^M_n \) denotes the empirical spectral distribution (ESD) of \( M \), that is,

\[
F^M_n(x) = \frac{1}{p} \sum_{i=1}^{p} 1_{\lambda^M_i \leq x}, \quad x \in \mathbb{R}.
\]

We will consider random matrix \( M \) whose ESD \( F^M_n \) 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

\[
\hat{\theta} = \int f(x) dF^M_n(x) = \frac{1}{p} \sum_{i=1}^{p} f(\lambda^M_i),
\]

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 \( \{ \xi_{ki} \in \mathbb{C}, i, k = 1, 2, \ldots \} \) be a double array of i.i.d. complex variables with mean 0 and variance 1. Set \( \xi_i = (\xi_{1i}, \xi_{2i}, \ldots, \xi_{pi})^T \), the vectors \( \xi_1, \ldots, \xi_n \) 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
\begin{equation}
S_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i \xi_i^*.
\end{equation}

For $0 < \theta \leq 1$, let $a(\theta) = (1 - \sqrt{\theta})^2$ and $b(\theta) = (1 + \sqrt{\theta})^2$. The Marchenko–Pastur distribution of index $\theta$, denoted as $F^\theta$, is the distribution on $[a(\theta), b(\theta)]$ with the following density function
\begin{equation}
g_\theta(x) = \frac{1}{2\pi \theta x} \sqrt{[b(\theta) - x][x - a(\theta)]}, \quad a(\theta) \leq x \leq b(\theta).
\end{equation}

Let $y_n \to y \in (0, 1)$ and $F^y, F^{y_n}$ be the Marchenko–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 $\mathcal{A}$,
\begin{equation}
G_n(f) = p \cdot \int_{-\infty}^{+\infty} f(x)[F_n - F^{y_n}](dx), \quad f \in \mathcal{A},
\end{equation}
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 $f_1, \ldots, f_k \in \mathcal{A}$, 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 $(G_n(f_1), \ldots, G_n(f_k))$ weakly converges to a $k$-dimensional Gaussian vector with mean vector
\begin{equation}
m(f_j) = \frac{f_j(a(y)) + f_j(b(y))}{4}
\end{equation}
\begin{equation}
- \frac{1}{2\pi} \int_{a(y)}^{b(y)} \frac{f_j(x)}{\sqrt{4y - (x - 1 - y)^2}} dx, \quad j = 1, \ldots, k,
\end{equation}
and covariance function
\begin{equation}
u(f_j, f_\ell) = -\frac{1}{2\pi^2} \iint \frac{f_j(z_1)f_\ell(z_2)}{(m(z_1) - m(z_2))^2} dm(z_1) dm(z_2),
\end{equation}
\begin{equation}
j, \ell \in \{1, \ldots, k\},
\end{equation}
where $m(z) \equiv m_{F^y}(z)$ is the Stieltjes Transform of $F^y \equiv (1 - y)I_{(0,\infty)} + yF^y$. 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 \( S_n' = T_n^{1/2}S_nT_n^{1/2} \) 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 \( (\xi_{11}, \ldots, \xi_{n_1}) \) and \( (\eta_{11}, \ldots, \eta_{n_2}) \) 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,
\[
V_n = S_1 S_2^{-1},
\]
where \( n_2 > p. \) Here, we use the notation \( n = (n_1, n_2). \)

Let \( y_{n_1} = \frac{p}{n_1} \to y_1 \in (0, 1), \quad y_{n_2} = \frac{p}{n_2} \to y_2 \in (0, 1). \)

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
\[
\ell(x) = \begin{cases} 
\frac{(1 - y_2) \sqrt{(b - x)(x - a)}}{2\pi x(y_1 + y_2 x)}, & a \leq x \leq b, \\
0, & \text{otherwise},
\end{cases}
\]
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 \( \hat{U} \) 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 $\tilde{A}$ be the set of analytic functions $f : \mathbb{C} \mapsto \mathbb{C}$. Define the empirical process $\tilde{G}_n := \{\tilde{G}_n(f)\}$ indexed by $\tilde{A}$

\begin{equation}
\tilde{G}_n(f) = p \cdot \int_{-\infty}^{+\infty} f(x)[F_n^{V_n} - F_{y_{n_1}, y_{n_2}}](dx), \quad f \in \tilde{A}.
\end{equation}

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 $f_1, \ldots, f_k \in \tilde{A}$, and assume the following: for each $p$, $(\xi_{ij})$ and $(\eta_{ij})$ variables are i.i.d., $1 \leq i \leq p, 1 \leq j_1 \leq n_1, 1 \leq j_2 \leq n_2$. $E\xi_{11} = E\eta_{11} = 0$, $E|\xi_{11}|^4 = E|\eta_{11}|^4 < \infty$, $y_{n_1} = \frac{p}{n_1} \rightarrow y_1 \in (0, 1)$, $y_{n_2} = \frac{p}{n_2} \rightarrow y_2 \in (0, 1)$. Then:

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

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

\begin{equation}
\frac{1}{4\pi i} \oint_{|\xi|=1} f_j(z(\xi)) \left[ \frac{1}{\xi - 1/r} + \frac{1}{\xi + 1/r} - \frac{2}{\xi + y_2/(hr)} \right] d\xi
\end{equation}

\begin{equation}
+ \frac{\beta \cdot y_1(1-y_2)^2}{2\pi i \cdot h^2} \oint_{|\xi|=1} f_j(z(\xi)) \frac{1}{(\xi + y_2/(hr))^3} d\xi
\end{equation}

\begin{equation}
+ \frac{\beta \cdot y_2(1-y_2)}{2\pi i \cdot h} \oint_{|\xi|=1} f_j(z(\xi)) \frac{\xi + 1/(hr)}{(\xi + y_2/(hr))^3} d\xi,
\end{equation}

\begin{equation}
\text{where } z(\xi) = (1-y_2)^{-2}[1+h^2+2hR(\xi)], h = \sqrt{y_1 + y_2 - y_1 y_2}, \beta = E|\xi_{11}|^4 - 3, \text{ and the covariance function as } 1 < r_1 < r_2 \downarrow 1
\end{equation}

\begin{equation}
u(f_j, f_\ell) = \lim_{1 < r_1 < r_2 \rightarrow 1_+} [(2.12) + (2.13)]
\end{equation}

\begin{equation}
- \frac{1}{2\pi^2} \oint_{|\xi_2|=1} \oint_{|\xi_1|=1} \frac{f_j(z(r_1 \xi_1)) f_\ell(z(r_2 \xi_2)) r_1 r_2}{(r_2 \xi_2 - r_1 \xi_1)^2} d\xi_1 d\xi_2,
\end{equation}

\begin{equation}
- \frac{\beta \cdot (y_1 + y_2)(1-y_2)^2}{4\pi^2 h^2} \oint_{|\xi_1|=1} \frac{f_j(z(\xi_1))}{(\xi_1 + y_2/(hr_1))^2} d\xi_1
\end{equation}

\begin{equation}
\times \oint_{|\xi_2|=1} \frac{f_\ell(z(\xi_2))}{(\xi_2 + y_2/(hr_2))^2} d\xi_2, \quad j, \ell \in \{1, \ldots, k\}.
\end{equation}
(ii) Complex case. Assume \((\xi_{ij})\) and \((\eta_{ij})\) are complex, 
\[E(\xi_{11}^2) = E(\eta_{11}^2) = 0,\] then the conclusion of (i) also holds, except the means are 
\[\lim_{r \to 1+} \left( \frac{1}{2} \cdot (2.12) + (2.13) \right),\] and the covariance function is 
\[\lim_{1 < r_1 < r_2 \to 1+} \left[ \frac{1}{2} \cdot (2.12) + (2.13) \right],\] where \(\beta = E|\xi_{11}|^4 - 2.\)

We should point out that Zheng’s CLT for \(F\)-matrices covers more general situations than 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 
\[
\begin{align*}
  c^2 + d^2 &= a(1 - y_2)^2 + b(1 + h^2), \\
  cd &= bh, \\
  0 < d < c.
\end{align*}
\]

Analogously, let \(\gamma, \eta\) 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 
\[
\begin{align*}
  m(f) &= \frac{1}{2} \log \frac{(c^2 - d^2)h^2}{(ch - y_2d)^2}, \\
  \nu(f, g) &= 2bdh^{-1}c^{-1} \log \frac{c\gamma}{c\gamma - d\eta}.
\end{align*}
\]

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 \(S\) and likelihood ratio statistic \(T_n\) as defined in (1.2) and (1.4), respectively. For \(\xi_i = x_i - \mu_p,\) the array \(\{\xi_i\}_{i=1,\ldots,n}\) contains \(p\)-dimensional standard normal variables under \(H_0.\) Let 
\[
S_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i \xi_i^* \]
and
\[
\tilde{L}^* = \text{tr} S_n - \log |S_n| - p.
\]

**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 \(n \to \infty\)
\[
T_n = \nu(g)^{-1/2} [L^* - p \cdot F^\gamma_n (g) - m(g)] \Rightarrow N(0, 1),
\]
where \(F^\gamma_n\) is the Marčenko–Pastur law of index \(\gamma_n.\)
PROOF. Because the difference between $S$ and $S_n$ is a rank-1 matrix, $S$ and $S_n$ have the same LSD. So, $L^*$ and $\tilde{L}^*$ have the same asymptotic distribution. We also have

$$\tilde{L}^* = \text{tr} S_n - \log |S_n| - p$$

$$= \sum_{i=1}^{p} (\lambda_i^{S_n} - \log \lambda_i^{S_n} - 1) = p \cdot \int (x - \log x - 1) dF_n(x)$$

$$= p \cdot \int g(x) d(F_n(x) - F^y_n(x)) + p \cdot F^y_n(g),$$

so that

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

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

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

and variance

(3.4) $$\nu(g) = -2 \log (1 - y) - 2y$$

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 $\nu(g)$. Then by (3.2), we arrive at

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

where

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

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

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

$\square$

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%.
4. Testing the equality of two high-dimensional covariance matrices. Let $(x_i), i = 1, \ldots, n_1$ and $(y_j), j = 1, \ldots, n_2$ 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

$$\xi_i = \Sigma^{-1/2}(x_i - \mu_1), \quad \eta_i = \Sigma^{-1/2}(y_i - \mu_2),$$

where $\Sigma = \Sigma_1 = \Sigma_2$ denotes the common covariance matrix under $H_0$. Note that in a strict sense, the vectors $(x_i), (y_i)$ 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 $(\xi_i)_{i=1, \ldots, n_1}$ and $(\eta_j)_{j=1, \ldots, n_2}$ 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

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

$$D = \frac{1}{n_2} \sum_{j=1}^{n_2} (y_j - \mu_2)(y_j - \mu_2)^*. $$
\[ D = \frac{1}{n_2} \sum_{j=1}^{n_2} (y_j - \mu_2)(y_j - \mu_2)^* . \]

Note that
\[ V_n = S_1 S_2^{-1} \]
forms a random $F$-matrix and we have
\[ \tilde{L}_1 = \left| S_1^{n_1/2} \cdot S_2^{n_2/2} \right| \left| c_1 S_1 + c_2 S_2 \right|^{N/2} \]
\[ = \frac{\left| S_1^{n_1/2} \cdot S_2^{n_2/2} \right|}{\left| c_1 C + c_2 D \right|^{N/2}} . \]

**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 $n_1 \land n_2 \to \infty$,
\[ \tilde{T}_N = \nu(f)^{-1/2} \left[ -\frac{2 \log L_1}{N} - p \cdot F_{y_{n_1}, y_{n_2}}(f) - m(f) \right] \Rightarrow N(0, 1) . \]

**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), $\tilde{L}_1$ and $L_1$ have the same asymptotic distribution. Because
\[ -\frac{2}{N} \log \tilde{L}_1 = -\frac{2}{N} \log \left( \left| S_1^{n_1/2} \cdot S_2^{n_2/2} \right| \right) \]
\[ = \log |c_1 V_n^{-1} + c_2| - c_1 \cdot \log |V_n^{n_1}| \]
\[ = \sum_{i=1}^{p} \log(c_1 \lambda_i V_n + c_2) - c_1 \cdot \log(\lambda_i V_n) \]
\[ = p \cdot \int [\log(c_1 x + c_2) - c_1 \cdot \log(x)] dF_n^{V_n}(x) . \]

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

From
\[ -\frac{2}{N} \log \tilde{L}_1 = p \cdot \int f(x) dF_n^{V_n}(x) \]
\[ = p \cdot \int f(x) d(F_n^{V_n}(x) - F_{y_{n_1}, y_{n_2}}(x)) + p \cdot F_{y_{n_1}, y_{n_2}}(f) , \]
we get

\[ \tilde{G}_n(f) = -\frac{2\log \tilde{L}_1}{N} - p \cdot F_{y_{n_1}, y_{n_2}}(f). \]

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

\[ m(f) = \frac{1}{2} \left[ \log \left( \frac{y_1 + y_2 - y_1 y_2}{y_1 + y_2} \right) - \frac{y_1}{y_1 + y_2} \log(1 - y_2) \right. \]

\[ \left. - \frac{y_2}{y_1 + y_2} \log(1 - y_1) \right] \]

and variance

\[ \nu(f) = -\frac{2y_2^2}{(y_1 + y_2)^2} \log(1 - y_1) - \frac{2y_1^2}{(y_1 + y_2)^2} \log(1 - y_2) \]

\[ -2\log \frac{y_1 + y_2}{y_1 + y_2 - y_1 y_2} \]

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

\[ -\frac{2\log \tilde{L}_1}{N} - p \cdot F_{y_{n_1}, y_{n_2}}(f) \Rightarrow N(m(f), \nu(f)), \]

where

\[ F_{y_{n_1}, y_{n_2}}(f) = -\left( \frac{y_{n_1} + y_{n_2} - y_{n_1} y_{n_2}}{y_{n_1} y_{n_2}} \right) \log \left( \frac{y_{n_1} + y_{n_2} - y_{n_1} y_{n_2}}{y_{n_1} y_{n_2}} \right) \]

\[ + \left( \frac{y_{n_1} + y_{n_2} - y_{n_1} y_{n_2}}{y_{n_1} y_{n_2}} \right) \log \left( \frac{y_{n_1} + y_{n_2}}{y_{n_1} + y_{n_2}} \right) \]

\[ + \left( \frac{y_{n_1} y_{n_2} (1 - y_{n_2})}{y_{n_1} y_{n_2} (1 - y_{n_2})} \right) \log (1 - y_{n_2}) \]

\[ + \left( \frac{y_{n_1} y_{n_2} (1 - y_{n_1})}{y_{n_1} y_{n_2} (1 - y_{n_1})} \right) \log (1 - y_{n_1}). \]

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

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

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.
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
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
Sizes of the corrected pseudo-likelihood ratio 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

<table>
<thead>
<tr>
<th>$(p, n_1, n_2)$</th>
<th>CLRT size</th>
<th>Schott’s size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(0.05, 0.1)$</td>
<td>0.067</td>
<td>0.517</td>
</tr>
<tr>
<td>$(10, 100, 200)$</td>
<td>0.065</td>
<td>0.603</td>
</tr>
<tr>
<td>$(20, 200, 400)$</td>
<td>0.054</td>
<td>0.703</td>
</tr>
<tr>
<td>$(40, 400, 800)$</td>
<td>0.048</td>
<td>0.764</td>
</tr>
<tr>
<td>$(80, 800, 1600)$</td>
<td>0.045</td>
<td>0.826</td>
</tr>
<tr>
<td>$(160, 1600, 3200)$</td>
<td>0.051</td>
<td>0.854</td>
</tr>
<tr>
<td>$(320, 3200, 6400)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

$$m_1(f) = \frac{1}{2} \left[ \log \left( \frac{y_1 + y_2 - y_1 y_2}{y_1 + y_2} \right) - \frac{y_1}{y_1 + y_2} \log(1 - y_2) \right. \right.$$

(4.8)

$$- \left. \frac{y_2}{y_1 + y_2} \log(1 - y_1) + \frac{6y_1^2 y_2}{(y_1 + y_2)^2} + \frac{6y_1 y_2^2}{(y_1 + y_2)^2} \right]$$

and

$$\nu_1(f) = - \frac{2y_2^2}{(y_1 + y_2)^2} \log(1 - y_1) - \frac{2y_1^2}{(y_1 + y_2)^2} \log(1 - y_2)$$

(4.9)

$$- 2 \log \frac{y_1 + y_2}{y_1 + y_2 - y_1 y_2}$$

instead of $m(f)$ and $\nu(f)$ 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 $F_{\nu_1,\nu_2}(f)$, (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 $x = 1 + y - 2\sqrt{y} \cos \theta$, $0 \leq \theta \leq \pi$, we have

$$m(g) = \frac{g(a(y)) + g(b(y))}{4} - \frac{1}{2\pi} \int_{a(y)}^{b(y)} \frac{g(x)}{\sqrt{4y - (x - 1 - y)^2}} dx$$

$$= \frac{y - \log(1 - y)}{2} - \frac{1}{2\pi} \int_{0}^{\pi} \left[ 1 + y - 2\sqrt{y} \cos \theta - \log(1 + y - 2\sqrt{y} \cos \theta) - 1 \right] d\theta$$

$$= \frac{y - \log(1 - y)}{2} - \frac{1}{4\pi} \int_{0}^{2\pi} \left[ y - 2\sqrt{y} \cos \theta - \log|1 - \sqrt{y} e^{i\theta}|^2 \right] d\theta$$

$$= -\frac{\log(1 - y)}{2},$$

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

$$\nu(g) = -\frac{1}{2\pi^2} \oint \oint \frac{g(z_1)g(z_2)}{(m(z_1) - m(z_2))^2} \frac{dm(z_1) dm(z_2)}{z_1 z_2 - z_1 \log z_2 - z_2 \log z_1 + \log z_1 \log z_2 - z_1 + \log z_1 - z_2 + \log z_2 + 1}$$

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

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

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

$$\oint \frac{\log(z_1(z_2))}{(m_1 - m_2)^2} dm_1 = \oint \frac{1}{m_1^2 - y/(1 + m_1)^2} \frac{1}{-1/m_1 + y/(1 + m_1) (m_1 - m_2)} dm_1$$

$$= \oint \frac{(1 + m_1)^2 - y m_1^2}{ym_1 (m_1 - m_2)} \left( \frac{-1}{m_1 + 1} + \frac{1}{m_1 - 1/(y - 1)} \right) dm_1$$

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

Then we also get \( \nu(-z_1 + \log z_1, 1) = 0 \). Similarly, \( \nu(1, -z_2 + \log z_2) = 0 \). Furthermore,
\[
\nu(z_1, z_2) = \frac{y^2}{\pi i} \oint \frac{1}{(m_2 + 1)^2} \left( \frac{1}{1 + m_2} + \frac{1 - y}{y} \right) \sum_{j=0}^{\infty} (1 + m_2)^j \, dm_2 = 2y
\]
and
\[
\nu(z_1, \log z_2) = \frac{y}{\pi i} \oint \left( \frac{1}{m_2 + 1} - \frac{1}{m_2 - 1/(y - 1)} \right) \left( \frac{1}{1 + m_2} + \frac{1 - y}{y} \right) \times [1 - (1 + m_2)]^{-1} \, dm_2 = \frac{y}{\pi i} \oint \left( \frac{1}{m_2 + 1} - \frac{1}{m_2 - 1/(y - 1)} \right) \left( \frac{1}{1 + m_2} + \frac{1 - y}{y} \right) \times \sum_{j=0}^{\infty} (1 + m_2)^j \, dm_2 = 2y.
\]

By a computation in [3], we know that \( \nu(\log z_1, \log z_2) = -2\log(1 - y) \). Finally, we obtain
\[
\nu(g) = \nu(z_1, z_2) + \nu(\log z_1, \log z_2) - 2\nu(z_1, \log z_2) + \nu(-z_1 + \log z_1, 1) + \nu(1, -z_2 + \log z_2) + \nu(1, 1)
= -2\log(1 - y) - 2y.
\]
Proof of (3.6). Since $F^y_n$ is the Marčenko–Pastur law of index $y_n$, by using the variable change $x = 1 + y_n - 2\sqrt{y_n} \cos \theta$, $0 \leq \theta \leq \pi$ we have

$$F^y_n(g) = \int_{a(y_n)}^{b(y_n)} \frac{x - \log x - 1}{2\pi x y_n} \sqrt{(b(y_n) - x)(x - a(y_n))} \, dx$$

$$= \frac{1}{2\pi y_n} \int_0^\pi \left[ 1 - \frac{\log(1 + y_n - 2\sqrt{y_n} \cos \theta) + 1}{1 + y_n - 2\sqrt{y_n} \cos \theta} \right] 4y_n \sin^2 \theta \, d\theta$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \left[ 2\sin^2 \theta - \frac{2\sin^2 \theta}{1 + y_n - 2\sqrt{y_n} \cos \theta} \left( \log|1 - \sqrt{y_n e^{i\theta}}|^2 - 1 \right) \right] \, d\theta$$

$$= 1 - \frac{y_n - 1}{y_n} \log(1 - y_n),$$

where

$$\frac{1}{2\pi} \int_0^{2\pi} \frac{2\sin^2 \theta}{1 + y_n - 2\sqrt{y_n} \cos \theta} \log|1 - \sqrt{y_n e^{i\theta}}|^2 \, d\theta$$

$$= \frac{y_n - 1}{y_n} \log(1 - y_n) - 1$$

is calculated in [3].

REFERENCES


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: stabaidz@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