Identity Tests for High Dimensional Data Using RMT

Cheng Wang[a,b,c], Jing Yang[b], Baiqi Miao[a], Longbing Cao[b]

[a] *Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China*
[b] *Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia*
[c] *wwcc@mail.ustc.edu.cn*

Abstract

In this work, we redefined two important statistics, the CLRT test (Bai et.al., Ann. Stat. **37** (2009) 3822-3840) and the LW test (Ledoit and Wolf, Ann. Stat. **30** (2002) 1081-1102) on identity tests for high dimensional data using random matrix theories. Compared with existing CLRT and LW tests, the new tests can accommodate data which has unknown means and non-Gaussian distributions. Simulations demonstrate that the new tests have good properties in terms of size and power. What's more, even for Gaussian data, our new tests perform favorably in comparison to existing tests. Finally, we find the CLRT is more sensitive to eigenvalues less than 1 while the LW test has more advantages in relation to detecting eigenvalues larger than 1.

Keywords: High dimensional data, Identity test, Random Matrix Theory(RMT)
2000 MSC: 62H15, 62H10

1. Introduction

Suppose $X_1, \cdots, X_n$ are independent and identically distributed (i.i.d.) p-dimensional random vectors with population covariance matrix Σ_p and our interest is to test

$$H_0 : \ \Sigma_p = I_p \ vs. \ H_1 : \ \Sigma_p \neq I_p, \tag{1.1}$$

where I_p denotes the $p-$dimensional identity matrix. Note that the identity matrix in (1.1) can be replaced by any other positive definite matrix Σ_0 through multiplying the data by $\Sigma_0^{-1/2}$.

In this work, we assume $y_n = p/n \to y \in (0, \infty)$. For canonical statistical analysis where the sample size n tends to infinity while the dimension p remains fixed, one can refer to Anderson (2003). When $y < 1$, Bai et al. (2009) proposed a correction to the classic likelihood ration test (CLRT) and derived the central limit theorem(CLT) using random matrix theories(RMT). When $y \geq 1$, CLRT is degenerate since the sample covariance is no longer invertible with probability one and Ledoit and Wolf (2002) gave a new statistics (LW test) which could accommodate situations for any $y > 0$. We note that the LW test has received much attention in relevant literature including Schott (2006) who considered the test for part of the eigenvalues of Σ_p are equal and Birke and Dette (2005) who extended the LW test to cases $y = 0$ and ∞. There has also been a substantial body of research motivated by the LW test such as Fisher et al. (2010), Srivastava (2005) and Lin and Xiang (2008).

However, most of these results were derived under Gaussian assumptions or equivalent conditions such as the fourth moment equals three. The difficulty in relaxing Gaussian assumptions is due to the central limit theorems for linear spectral statistics defined by eigenvalues. More details can be found in Bai and Silverstein (2004) who built the CLT for linear spectral statistics of large-dimensional sample covariance matrices under the assumption of fourth moments and Pan and Zhou (2008) who improved the results for general finite fourth moments. In Bai and Silverstein (2004) and Pan and Zhou (2008), the authors proposed a simplified version of classic sample covariance matrices where the means of the data must be known. Recently, Pan (2012) derived the CLT for linear spectral statistics of classic large-dimensional sample covariance matrices. Some other important results include Bai et al. (2010), Lytova and Pastur (2009) and so on.

CLRT in Bai et al. (2009) is only applicable to Gaussian data with known means and the LW test in Ledoit and Wolf (2002) can only be applied to Gaussian data. Since the two tests are too narrow for use in applications, in this work, we will redefine the CLRT and LW tests using classic sample covariance matrices. The CLTs of the two new tests are derived in general conditions which can accommodate data with unknown means and non-Gaussian distributions. Simulations demonstrate that the proposed tests have good properties in terms of size and power. What's more, even for Gaussian data, our new tests perform favorably in comparison to existing tests. We also study the features of each test. That is, compared with the LW test, the CLRT has its own advantages on detecting the eigenvalues of Σ_p near zero which means the CLRT is more sensitive to eigenvalues less than 1

while the LW test has more advantages on detecting eigenvalues larger than 1. In the existing literature, there is also some work which is not based on sample covariance matrices such as Chen et al. (2010) who proposes a test by constructing estimators from the data directly. We also conduct simulations to compare our proposed tests with the one in Chen et al. (2010).

The paper is structured as follows: Section 2 introduces the basic data structure and establishes the asymptotic normality of the new CLRT and new LW tests while Section 3 reports simulation studies. All the technical details include proofs and the preliminary results in RMT are presented in the Appendix.

2. Main Results

We assume the observations $X_1, \cdots, X_n$ satisfies a multivariate model (Bai and Saranadasa (1996))

$$X_i = \Sigma_p^{1/2} Y_i + \mu, \ for\ i = 1, \cdots, n, \tag{2.2}$$

where μ is a p-dimensional constant vector and the entries of $\mathcal{Y}_n = (Y_{ij})_{p\times n} = (Y_1, \cdots, Y_n)$ are i.i.d. with $EY_{ij} = 0$, $EY_{ij}^2 = 1$ and $EY_{ij}^4 = 3 + \Delta$. Here we introduce two versions of the sample covariance matrices. The classic one is defined as

$$S_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})',$$

where $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ and a simplified version takes the form

$$B_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)(X_k - \mu)'.$$

We refer to Bai and Silverstein (2010) and Pan (2012) for the differences between S_n and B_n in RMT. Then we can introduce CLRT in Bai et al. (2009)

$$\hat{L}_n = \frac{1}{p}tr(B_n) - \frac{1}{p}\log|B_n| - 1, \tag{2.3}$$

where tr denotes the trace.

When $X_i \sim N_p(0, I_p)$ (or $X_i \sim N_p(\mu, I_p)$ where μ is known), Bai et al. (2009) derived the CLT of CLRT

$$p(\hat{L}_n - (1 + (1/y_n - 1)\log(1 - y_n))) \xrightarrow{D} \mathbb{N}(-\log(1-y)/2, -2y - 2\log(1-y)),$$

where $\xrightarrow{D}$ denotes convergence in distribution and $\mathbb{N}$ the normal distribution.

When p is larger than sample size n, since the sample covariance matrix S_n is singular, Ledoit and Wolf (2002) proposed the LW test which is defined as

$$\hat{W}_n = \frac{1}{p}tr(S_n - I_p)^2 - \frac{p}{n-1}(\frac{1}{p}tr(S_n))^2, \tag{2.4}$$

If $X_i \sim N_p(\mu, I_p)$ and under some other assumptions, Ledoit and Wolf (2002) has proven

$$n\hat{W}_n \xrightarrow{D} \mathbb{N}(1, 4). \tag{2.5}$$

In our work, we will redefine the CLRT as

$$L_n = \frac{1}{p}tr(S_n) - \frac{1}{p}\log|S_n| - 1, \tag{2.6}$$

and the LW test as

$$W_n = \frac{1}{p}tr(S_n - I_p)^2 - \frac{p}{n-1}(\frac{1}{p}tr(S_n))^2 - \frac{(1+\Delta)(n-2)(n-1)-2}{n(n-1)^2}. \tag{2.7}$$

In (2.6) and (2.7), noting that the statistics are defined by S_n which is invariant under the shift transformation $X_i = X_i + c$, we can assume $\mu = 0$ in (2.2) without loss of generality. By Bai and Silverstein (2004), we know almost surely

$$\frac{1}{p}\log|S_n| - \frac{1}{p}\log|\Sigma_p| \xrightarrow{a.s.} -1 - (1/y - 1)\log(1-y)) \equiv d(y),$$

which means $L_n + d(y)$ is a estimator of $(tr(\Sigma_p) - \log|\Sigma_p| - p)/p$. Similarly, W_n is a estimator of $tr(\Sigma_p - I_p)^2/p$. Noting that

$$\Sigma_p = I_p \Leftrightarrow (tr(\Sigma_p) - \log|\Sigma_p| - p)/p = 0 \Leftrightarrow tr(\Sigma_p - I_p)^2/p = 0,$$

therefore, L_n and W_n can act as the statistics to test (1.1) theoretically. Next, we will establish the asymptotic normalities of L_n and W_n.

Theorem 2.1. *When* $\Sigma_p = I_p$ *and* $p/y \to y \in (0,1)$,

$$p(L_n - (1 + (1/y_n - 1)\log(1 - y_n))) \xrightarrow{D} \mathbb{N}(m, v),$$

where $m = y(\Delta/2 - 1) - 3\log(1-y)/2$ *and* $v = -2y - 2\log(1-y)$.

Compared with the CLRT in Bai et al. (2009) which is only applicable to Gaussian data with known means, our new CLRT can be applied to general data with unknown means. Further, if the population mean is unknown, the CLRT in Bai et al. (2009) will behave poorly and the new one will still be applicable.

Theorem 2.2. *Under* H_0 *and* $p/n \to y \in (0, \infty)$,

$$pW_n \xrightarrow{D} \mathbb{N}(0, 4y^2).$$

When $EY_{ij} \neq 0$, Theorem 2.1 and 2.2 are still applicable under new assumptions $E(Y_{ij} - EY_{ij})^2 = 1$ and $E(Y_{ij} - EY_{ij})^4 = 3 + \Delta$ by Pan (2012). In Theorem 2.2, when $X_i \sim N(\mu, I_p)$ that is $\Delta = 0$, we can get

$$p(\hat{W}_n - \frac{(n-2)(n-1)-2}{n(n-1)^2}) \xrightarrow{D} \mathbb{N}(0, 4y^2)$$

which is in accordance with (2.5) by Slutsky's theorem. Further, direct calculations can show that when $\Sigma = I_p$, the new LW test is the unique best unbiased estimator of $\frac{1}{p}tr(\Sigma_p - I_p)^2$ by Lehmann-Scheffé theorem when the LW test always has a $O(\frac{1}{n^2})$ bias. Therefore, our new LW test behaves better than the LW test for Gaussian data especially when the sample size n is small. Moreover, the new LW test can be applied to data with general distributions while the existing LW test is only applicable to Gaussian data.

Here we also mention the result of Chen et al. (2010) (CZZ test) which is also an unbiased estimator of $\frac{1}{p}tr(\Sigma_p - I_p)^2$ and based on $\{X_1, \cdots, X_n\}$ directly. Compared with the CZZ test which does not depend on Δ, our new LW test has several advantages. Firstly, for Gaussian data where $\Delta = 0$, the new LW test behaves better due to it is the unique best unbiased estimator of $\frac{1}{p}tr(\Sigma_p - I_p)^2$. Secondly, the new LW test has a more simple formula. For example, to calculate the new LW test, we only need the sample covariance matrix S_n when the CZZ test consists of five parts. Lastly, for general distributions, simulations show that the new LW test has a better

size when $\Delta < 0$. In addition, the fourth moment Δ is a regular condition in the CLT of statistics based on a high-dimensional sample covariance matrix. Δ appears, for example, in the work of Bai and Silverstein (2004), Pan and Zhou (2008), Lytova and Pastur (2009), Bai et al. (2010) and Pan (2012).

3. Simulations

We report results from simulation studies which were designed to evaluate the performance of the proposed identity tests. Here, the ratio y could be estimated by $y_n = p/n$. To evaluate the power of the tests, two different population covariance matrices will be considered in the simulations. We set $\Sigma_p^{(1)} = diag(1.5\ I_{[0.2p]}, I_{p-[0.2p]})$ and $\Sigma_p^{(2)} = diag(0.5\ I_{[0.2p]}, I_{p-[0.2p]})$, where $[x]$ denoted the integer truncation of x. The diagonal covariance Σ_p has respectively 20% of its diagonal elements being 1.5 or 0.5 whereas the rest are 1.

3.1. CLRT and New CLRT

In this part, we will study our new CLRT and the existing CLRT. Since the existing CLRT in Bai et al. (2009) can only deal with Gaussian variables with known means, we only consider Gaussian variables with zero means in our simulations. Table 1 shows the empirical sizes and powers of our redefined CLRT and the existing CLRT for Gaussian variables $Y_{ij} \sim \mathbb{N}(0, 1)$. The nominal test level is set at 5% and all results are based on 10^3 replications.

From Table 1, we know even for Gaussian variables with known means, our new CLRT is comparable to one in Bai et al. (2009). As p and n both have increased, the sizes or powers of the two tests are quite close to 5% or 1 and make not much difference. Further, for the same sample size n, when y gets smaller, the sizes are closer to 5% and the powers are closer to 1. The explanation is that if we only have n samples, when p gets smaller (that is y gets smaller), our redefined CLRT or existing CLRT, as the estimator of $(tr(\Sigma_p) - \log|\Sigma_p| - p)/p$, will become more accurate.

If the true mean μ is not zero and we still use the CLRT in Bai et al. (2009), the results of the last part in Table 1 is the experiment for $Y_{ij} \sim \mathbb{N}(1/4, 1)$. It can be found that the existing CLRT behaves very poorly and our CLRT is still applicable.

Table 1: Performances of the redefined CLRT and existing CLRT

	Redefined CLRT			Existing CLRT		
n	$y=0.25$	$y=0.5$	$y=0.75$	$y=0.25$	$y=0.5$	$y=0.75$
$\Sigma_p = I_p,\ Y_{ij} \sim \mathbb{N}(0,1)$						
40	0.077	0.072	0.076	0.071	0.061	0.062
80	0.062	0.061	0.062	0.060	0.056	0.055
160	0.054	0.053	0.054	0.053	0.052	0.053
$\Sigma_p = \Sigma_p^{(1)},\ \ Y_{ij} \sim \mathbb{N}(0,1)$						
40	0.220	0.182	0.177	0.214	0.168	0.162
80	0.397	0.342	0.281	0.397	0.337	0.275
160	0.819	0.769	0.632	0.816	0.762	0.625
200	0.926	0.889	0.815	0.925	0.895	0.816
$\Sigma_p = \Sigma_p^{(2)},\ \ Y_{ij} \sim \mathbb{N}(0,1)$						
40	0.371	0.369	0.272	0.370	0.367	0.278
80	0.841	0.749	0.618	0.840	0.762	0.641
160	1	1	0.986	1	0.999	0.990
$\Sigma_p = I_p,\ \ Y_{ij} \sim \mathbb{N}(1/4,1)$						
100	0.066	0.051	0.059	0.689	0.837	0.834

3.2. LW, New LW and CZZ tests

From the definitions of the LW and the new LW tests, we know $n(W_n - \hat{W}_n) = O(\frac{1}{n})$ which means the two statistics are quite similar for normal distributions. Therefore, our first experiment is to investigate the empirical sizes of the LW, the new LW and the CZZ tests on Gaussian data with a small sample size. Results based on 10^4 replications are reported in Table 2.

Table 2: Sizes of the new LW, LW and CZZ tests on Gaussian data.

	New LW test			LW test			CZZ test		
p	$n=5$	10	50	$n=5$	10	50	$n=5$	10	50
5	0.100	0.086	0.067	0.106	0.088	0.067	0.163	0.112	0.073
10	0.107	0.086	0.068	0.115	0.087	0.068	0.167	0.098	0.069
50	0.108	0.082	0.059	0.114	0.083	0.059	0.158	0.095	0.063
100	0.114	0.085	0.057	0.122	0.086	0.057	0.157	0.100	0.060

We observe from Table 2 that the sizes of the LW and the new LW tests are always better than the CZZ test for normal distributions. The reason

is due to the fact that the LW and the new LW tests are based on the sample covariance matrix S_n which is a completely sufficient statistic for Σ_p for Gaussian data. When the sample size n is small such as $n = 5$ in the experiments, the new LW test has the better sizes than the LW test and when n is large, the performances of the LW and the new LW tests are quite similar. This is because the new LW test is always the unique best unbiased estimator of $\frac{1}{p}tr(\Sigma_p - I_p)^2$ while the LW test has a $O(\frac{1}{n^2})$ bias.

For non-Gaussian data, from Theorem 2.2, we know that the CLT of the LW test depends on Δ and it is not reasonable that Chen et al. (2010) assumed $\Delta = 0$ even for gamma random vectors. From Theorem 2.2, when $\Sigma_p = I_p$ and $Y_{i,j} \sim Gamma[4, 0.5]$, by Slutsky's theorem we know

$$(n\hat{W}_n - 1)/2 \xrightarrow{D} \mathbb{N}(0.75, 1).$$

However, in Chen et al. (2010), the authors still thought $(n\hat{W}_n - 1)/2 \xrightarrow{D} \mathbb{N}(0, 1)$. Moreover, since $P(Z > \Phi^{-1}(0.95)) = 0.185$ where $Z \sim \mathbb{N}(0.75, 1)$ and Φ is the distribution function of standard norm variables, this explains why the size of the LW test in Chen et al. (2010) is near 0.185 not 5%.

Here we will repeat part of the simulations in Chen et al. (2010) using the new LW test. Two scenarios are considered
(I) $Y_{i,j}$ $i.i.d.$ $\sim Gamma[4, 0.5]$ where $\Delta = 1.5$;
(II) $Y_{i,j}$ $i.i.d.$ $\sim Uniform[0, 2\sqrt{3}]$ where $\Delta = -1.2$.
Simulation results are reported in Table 3 where the performances are based on 10^4 replications. It is noted that Table 1 in Chen et al. (2010) has the results for the sphericity test, not the identity test. Since the authors claimed the simulation results for the identity test followed very similar patterns to those of the sphericity test, here for comparison purposes, we will still use Table 1 in Chen et al. (2010) for the identity test.

From Table 3, we can see that the new LW test is not as bad as shown in Table 1 of Chen et al. (2010). The sizes of the new LW test and the CZZ test are comparable and when p and n increase, the sizes both tend to the nominal 5% level. Specially, for Gamma data ($\Delta = 1.5$), the CZZ test has a better size while the sizes of the new LW test are closer to the nominal level for Uniform variables ($\Delta = -1.2$). From Table 3, it seems like that the LW test has a better size when $\Delta < 0$ compared with the CZZ test. To verify this point, we designed another experiment to investigate the differences of the sizes between the new LW test and the CZZ test. For the new simulations,

Table 3: Performances of the new LW test and the CZZ test on non-Gaussian data

	LW test				CZZ test			
p	$n=20$	$n=40$	$y=60$	$n=80$	$n=20$	$n=40$	$n=60$	$n=80$
Gamma random vectors								
38	0.1165	0.0923	0.0833	0.0807	0.0861	0.0767	0.0707	0.0704
55	0.1110	0.0828	0.0822	0.0734	0.0810	0.0697	0.0694	0.0658
89	0.1042	0.0840	0.0704	0.0655	0.0792	0.0685	0.0591	0.0590
159	0.0998	0.0795	0.0649	0.0626	0.0754	0.0653	0.0583	0.0588
Uniform random vectors								
38	0.0574	0.0517	0.0490	0.0501	0.0678	0.0565	0.0527	0.0530
55	0.0614	0.0592	0.0539	0.0543	0.0728	0.0659	0.0563	0.0563
89	0.0548	0.0503	0.0530	0.0563	0.0650	0.0548	0.0570	0.0582
159	0.0592	0.0556	0.0546	0.0518	0.0718	0.0607	0.0563	0.0535

we set

$$P(Y_{ij} = -\sqrt{\frac{1-\gamma}{\gamma}}) = 1 - P(Y_{ij} = \sqrt{\frac{\gamma}{1-\gamma}}) = \gamma \in (0,1),$$

then it is easy to show

$$EY_{ij} = 0,\ EY_{ij}^2 = 1,\ \Delta = EY_{ij}^4 - 3 = \frac{(1-\gamma)^2}{\gamma} + \frac{\gamma^2}{1-\gamma} - 3.$$

In the experiment, by adjusting γ in $(0, 0.5)$ or $(0.5, 1)$, we can get the results for different Δ. The results based on 10^5 replications are reported in Figure 1 where $p = 50,\ n = 100$.

We observe from Figure 1 that when $\Delta < 0$, the new LW test has a better size and the CZZ test is better for $\Delta > 0$ which is consistent with results in Table 3. Here we can see the performances of the new LW test and the CZZ test are similar being around $\Delta = 0$ which is a little different from the results for Gaussian data ($\Delta = 0$). Another interesting result is that when Δ increases, the sizes of the new LW test and the CZZ test become worse although the CZZ test does not depend on Δ. We hope these questions can be addressed in future studies.

3.3. The powers of the new LW and CLRT tests

Results based on 10^3 replications are reported in Figure 2, and these correspond with $\Sigma_p = \Sigma_p^{(1)}$ or $\Sigma_p = \Sigma_p^{(2)}$ for three different distributions.

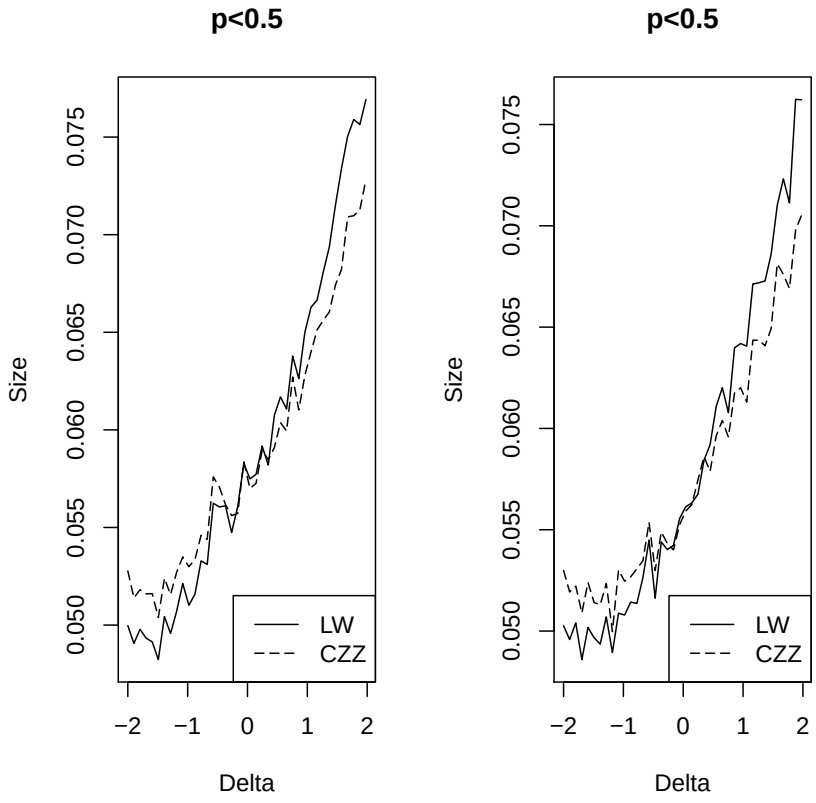

Figure 1: Realized sizes of the LW test and the CZZ test for different fourth moment $3 + \Delta$.

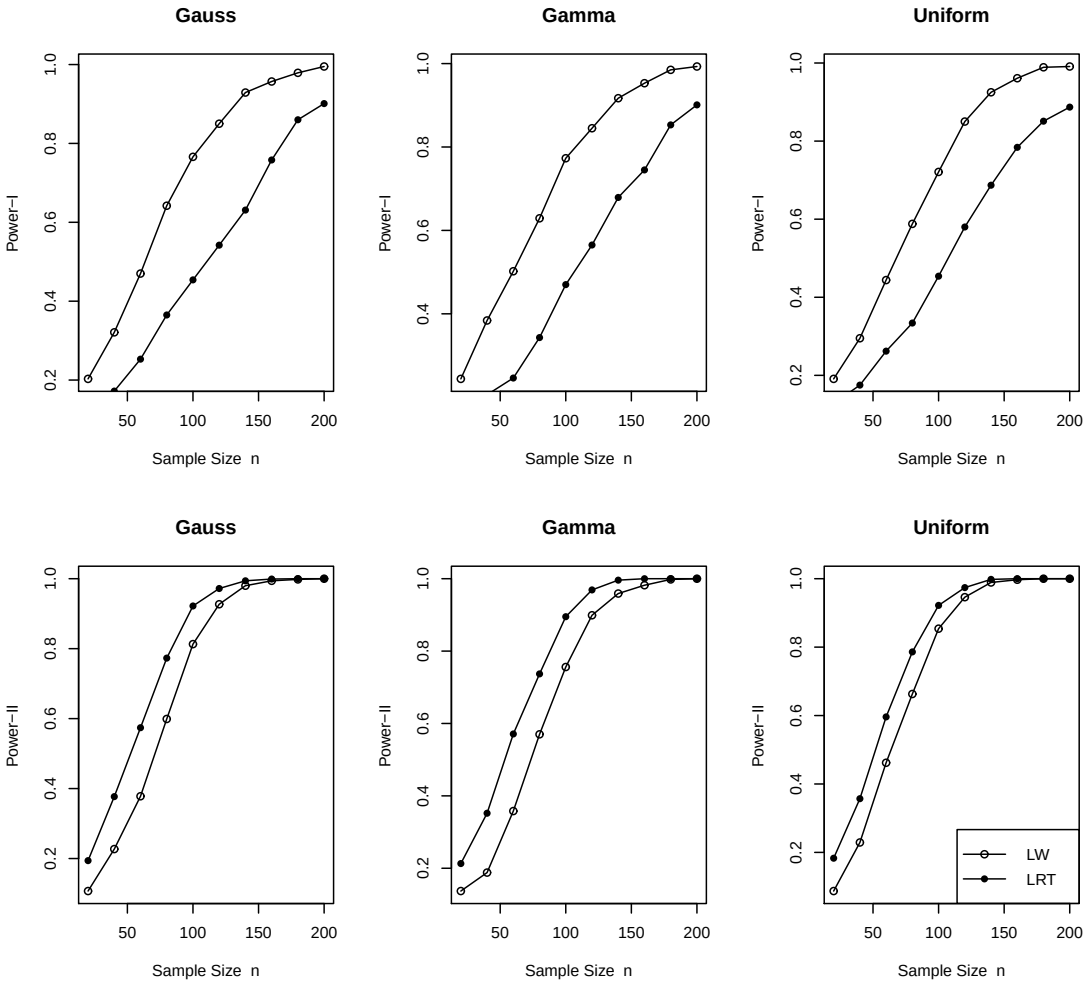

Figure 2: Realized powers of the LW test and the CLRT for different sample size n and $p/n = 0.5$. The top three are for $\Sigma_p = \Sigma_p^{(1)}$ and the bottom three are for $\Sigma_p = \Sigma_p^{(2)}$.

The power results in Figure 2 show that the new CLRT and the new LW tests both approach to 1 when n is increased. For $\Sigma_p^{(1)}$, when part of the eigenvalues of Σ_p are larger than 1, the power of the new CLRT is worse than the one of the new LW test and for $\Sigma_p^{(2)}$, when part of eigenvalues are less than 1, the new CLRT behaves better than the new LW test. The reason is due to the differences between $tr(\Sigma_p) - \log|\Sigma_p| - p$ and $tr(\Sigma_p - I_p)^2$ where the former is more sensitive to small eigenvalues and the latter has more advantages in terms of detecting large eigenvalues.

4. Conclusions

In this work, we modified two identity tests CLRT and LW test for high dimensional data. Compared with the existing CLRT, the new CLRT and LW test can accommodate data with unknown means and non-Gaussian distributions. Even for Gaussian data, our new tests perform favorably in comparison to existing tests. In this paper, we also studied the features of each test which show that the CLRT is more sensitive to eigenvalues less than 1 while the LW test has more advantages in relation to detecting eigenvalues larger than 1.

From simulations, we found the new LW test has a better size when $\Delta < 0$ and the CZZ test is better for $\Delta > 0$. The performances of the new LW test and the CZZ test are similar around $\Delta = 0$ which is a little different compared to the results for Gaussian data ($\Delta = 0$). Another interesting result is that when Δ increases, the sizes of the new LW test and the CZZ test become worse although the CZZ test does not depend on Δ.

In addition, from the simulations, we found the CLRT is not the best one for Gaussian variables although the CLRT came from the likelihood functions of normal distributions. Finally, the powers of the tests (including CZZ test) depend on the population covariance matrix. We hope these questions can be addressed in future studies and an accurate estimator for Δ can be derived.

Acknowledgements

The authors would like to thank the associate editor and an anonymous referee for their helpful comments. Cheng Wang's research was supported by NSF of China Grands (No. 11101397, 71001095 and 11271347). Longbing Cao's research was supported by Australian Research Council Discovery

Grants (DP1096218 and DP1301691) and Australian Research Council Linkage Grant (LP100200774).

5. Appendix

5.1. Preliminary results in RMT

Suppose A_n is an $n \times n$ Hermitian matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$. Define the empirical spectral distribution (ESD) of A_n as

$$F^{A_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(\lambda_i \leq x).$$

The limit distribution of F^{A_n} is called the limiting spectral distribution (LSD) of the sequence $\{A_n\}$.

And the Stieltjes transform of F^{A_n} is given by

$$m^{F^{A_n}}(z) = \int \frac{1}{x-z} dF^{A_n}(x) = \frac{1}{n} tr(A_n - zI_n)^{-1},$$

where $z = \mu + i\nu \in \mathcal{C}^+$. By the inverse formula,

$$F^{A_n}\{[a,b]\} = \lim_{v \to 0^+} \frac{1}{\pi} \int_a^b Im(m^{F^{A_n}}(x+iv))dx. \tag{5.8}$$

Here we need another sample covariance matrix which is defined as

$$\mathfrak{S}_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})' = \frac{n-1}{n} S_n. \tag{5.9}$$

If the spectral norm of Σ_p is bounded by a positive constant and F^{Σ_p} converges weakly to a non-random distribution H as $p \to \infty$, by Silverstein and Bai (1995) or Pan (2010), with probability 1, $F^{\mathfrak{S}_n}$ and F^{B_n} tend to the same probability distribution $F_{y,H}$, whose Stieltjes transform $m = m(z)$ $(z \in \mathcal{C}^+)$ satisfies

$$m = \int \frac{1}{t(1-y-yzm)-z} dH(t). \tag{5.10}$$

Denoting $G_n(x) = p(F^{\mathfrak{S}_n}(x) - F_{y_n,H_n}(x))$, for any analytic function f, $\int f(x)dG_n(x)$ converges weakly to a Gaussian variable X_f under some assumptions on Σ_p by Pan (2012).

When $\Sigma_p = I_p$, $F_{y,H}$ is standard MP law F_y whose density function is

$$g_y(x) = \frac{1}{2\pi yx}\sqrt{((1+\sqrt{y})^2 - x)(x - (1-\sqrt{y})^2))},\ (1-\sqrt{y})^2 \le x \le (1+\sqrt{y})^2,$$

and from (5.10), we know

$$m = \frac{1}{1 - y - yzm - z}.$$

Writing $\underline{m} = ym - \frac{1-y}{z}$, by Pan (2012), we have

$$\begin{aligned} EX_f &= -\frac{1}{2\pi i}\int \frac{y\underline{m}f(-\frac{1}{\underline{m}} + \frac{y}{1+\underline{m}})}{(1+\underline{m})((1+\underline{m})^2 - c\underline{m}^2)}d\underline{m} - \frac{\Delta}{2\pi i}\int \frac{y\underline{m}f(-\frac{1}{\underline{m}} + \frac{y}{1+\underline{m}})}{(1+\underline{m})^3}d\underline{m} \\ &\quad + \frac{y}{2\pi i}\int \frac{f(-\frac{1}{\underline{m}} + \frac{y}{1+\underline{m}})}{(1+\underline{m})(y\underline{m} - 1 - \underline{m})}d\underline{m}, \end{aligned} \tag{5.11}$$

and

$$\begin{aligned} Var(X_f) &= -\frac{1}{2\pi^2}\int\int f(z_1)f(z_2)\frac{1}{(\underline{m}(z_1) - \underline{m}(z_2))^2}d\underline{m}(z_1)d\underline{m}(z_2) \\ &\quad - \frac{y\Delta}{4\pi^2}(\int \frac{f(z)}{(1+\underline{m}(z))^2}d\underline{m}(z))^2. \end{aligned} \tag{5.12}$$

The contours in (5.11) and (5.12) are both contained in the analytic region for the function f and both enclose the support of $F_{y_n,H_n}(x)$ for large n. Moreover, the contours in (5.12) are disjoint.

5.2. Proofs of Theorem 2.1

Writing $f(x) = x - \log(x) - 1$, we have

$$L_n = \int f(x)dF^{S_n}(x),$$

and

$$
\begin{aligned}
&p(L_n - \int f(x)dF_{y_n}(x)) \\
&= p(\int f(x)dF^{\mathfrak{S}_n}(x) - \int f(x)dF_{y_n}(x)) + p(\int f(x)dF^{S_n}(x) - \int f(x)dF^{\mathfrak{S}_n}(x)) \\
&= p(\int f(x)dF^{\mathfrak{S}_n}(x) - \int f(x)dF_{y_n}(x)) + (tr(S_n) - log|S_n| - tr(\mathfrak{S}_n) + \log|\mathfrak{S}_n|) \\
&= p(\int f(x)dF^{\mathfrak{S}_n}(x) - \int f(x)dF_{y_n}(x)) + \frac{1}{n-1}tr(\mathfrak{S}_n) - p\log(\frac{n}{n-1}) \\
&= p(\int f(x)dF^{\mathfrak{S}_n}(x) - \int f(x)dF_{y_n}(x)) + o(1).
\end{aligned}
$$

From Pan (2012), $p(\int f(x)dF^{\mathfrak{S}_n}(x) - \int f(x)dF_{y_n}(x))$ converges weakly to the Gaussian variable X_f. Next we calculate the mean and variance of X_f. The following results have been given in Bai et.al.(2008)

$$\int f(x)dF_{y_n}(x) = 1 - \frac{y_n - 1}{y_n}\log(1 - y_n),$$

$$-\frac{1}{2\pi i}\int \frac{y\underline{m}f(-\frac{1}{\underline{m}} + \frac{y}{1+\underline{m}})}{(1+\underline{m})((1+\underline{m})^2 - c\underline{m}^2)}d\underline{m} = -\frac{1}{2}\log(1-y), \tag{5.13}$$

and

$$-\frac{1}{2\pi^2}\int\int \frac{f(z_1)f(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2}d\underline{m}(z_1)d\underline{m}(z_2) = -2y - 2\log(1-y). \tag{5.14}$$

Also, from Pan and Zhou (2008), we know

$$\frac{1}{2\pi i}\oint(-\frac{1}{\underline{m}} + \frac{y}{1+\underline{m}} - 1)\frac{y\underline{m}}{(1+\underline{m})^3}d\underline{m} = 0.$$

Therefore, to get EX_g, we still need to calculate

$$\begin{aligned}
& \frac{y}{2\pi i}\oint \log(-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}})\frac{\underline{m}}{(1+\underline{m})^3}d\underline{m} \\
= & \frac{y}{2\pi i}\oint \log(-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}})d(-\frac{1}{1+\underline{m}}+\frac{1}{2(1+\underline{m})^2}) \\
= & \frac{y}{2\pi i}\oint \frac{\frac{1}{\underline{m}^2}-\frac{y}{(1+\underline{m})^2}}{-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}}}(\frac{1}{1+\underline{m}}-\frac{1}{2(1+\underline{m})^2})d\underline{m} \\
= & \frac{y}{2\pi i}\oint (\frac{1+\underline{m}}{\underline{m}(y\underline{m}-\underline{m}-1)}-\frac{y\underline{m}}{(\underline{m}+1)(y\underline{m}-\underline{m}-1)})(\frac{1}{1+\underline{m}}-\frac{1}{2(1+\underline{m})^2})d\underline{m} \\
= & \frac{y}{2\pi i}\oint (\frac{1}{\underline{m}(y\underline{m}-\underline{m}-1)}-\frac{1}{2\underline{m}(\underline{m}+1)(y\underline{m}-\underline{m}-1)})d\underline{m} \\
= & \frac{y}{2}, \qquad (5.15)
\end{aligned}$$

and

$$\begin{aligned}
& \frac{y}{2\pi i}\int \frac{f(-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}})}{(1+\underline{m})(y\underline{m}-1-\underline{m})}d\underline{m} \\
= & \frac{y}{2\pi i}\int \frac{-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}}-1-\log(-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}})}{(1+\underline{m})(y\underline{m}-1-\underline{m})}d\underline{m} \\
= & -y+\frac{1}{2\pi i}\int \log(-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}})(\frac{1}{1+\underline{m}}-\frac{1}{\underline{m}-\frac{1}{y-1}})d\underline{m} \\
= & -y-\log(1-y), \qquad (5.16)
\end{aligned}$$

where we used the following results

$$\oint \frac{1}{(\underline{m}+1)^k(y\underline{m}-\underline{m}-1)}d\underline{m}=0,\ k=1,2,3,$$

and one equality in Bai and Silverstein (2004)

$$\frac{1}{\pi i}\int \log(-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}})(\frac{1}{1+\underline{m}}-\frac{1}{\underline{m}-\frac{1}{y-1}})d\underline{m}=-2\log(1-y).$$

By (5.13), (5.15) and (5.16), $EX_f=y(\Delta/2-1)-3\log(1-y)/2$.

Through a routine calculation, we have

$$\frac{1}{2\pi i}\oint f(-\frac{1}{\underline{m}}+\frac{y}{1+\underline{m}})\frac{1}{(1+\underline{m})^2}d\underline{m}=0. \qquad (5.17)$$

By (5.14) and (5.17), we have $Var(X_f) = -2y - 2\log(1-y)$.

The proof of Theorem 2.1 is complete.

5.3. Proofs of Theorem 2.2

Noticing that $p(\frac{1}{p}tr(\mathfrak{S}_n) - 1)$ satisfies CLT and $\frac{1}{p}tr(\mathfrak{S}_n) \to 1, a.s.$, we have

$$p((\frac{1}{p}tr(\mathfrak{S}_n))^2 - 2\frac{1}{p}tr(\mathfrak{S}_n) + 1) = p([\frac{1}{p}tr(\mathfrak{S}_n)] - 1)^2 = o(1).$$

By (5.9), we have

$$\begin{aligned}
&p[\hat{W}_n - (\frac{1}{p}tr(\mathfrak{S}_n - I_p)^2 - \frac{2}{n}tr(\mathfrak{S}_n) + \frac{p}{n}))\\
&= tr(\frac{n}{n-1}\mathfrak{S}_n - I_p)^2 - tr(\mathfrak{S}_n - I_p)^2 + (\frac{1}{n} - \frac{n^2}{(n-1)^3})(tr(\mathfrak{S}_n))^2 + o(1)\\
&= (\frac{n^2}{(n-1)^2} - 1)tr(\mathfrak{S}_n^2) - 2(\frac{n}{n-1} - 1)tr(\mathfrak{S}_n) + (\frac{1}{n} - \frac{n^2}{(n-1)^3})(tr(\mathfrak{S}_n))^2\\
&= -y^2 + o(1),
\end{aligned}$$

and

$$\begin{aligned}
&p(\frac{1}{p}tr(\mathfrak{S}_n - I_p)^2 - \frac{2}{n}tr(\mathfrak{S}_n) + \frac{p}{n}) - \int((x-1)^2 - 2yx + y)dp(F^{\mathfrak{S}_n}(x) - F_{y_n}(x))\\
&= (y - y_n)(2tr(\mathfrak{S}_n) - p) + p\int((x-1)^2 - 2yx + y)dF_{y_n}(x)\\
&= 2(y - y_n)(tr(\mathfrak{S}_n) - p)\\
&= o(1).
\end{aligned}$$

Writing $g(x) = (x-1)^2 - 2yx + y$, by Pan (2012), $\int g(x)dp(F^{\mathfrak{S}_n}(x) - F_{y_n}(x))$ converges weakly to gaussian variable X_g and through a routine calculation

$$EX_g = y + y\Delta + y^2 = y(1 + \Delta + y).$$

To get the variance $Var(X_g)$, we need

$$\begin{aligned}
&\oint \frac{g(z(m_1))}{(m_1 - m_2)^2}dm_1\\
=\ &\oint \frac{1}{(m_1 - m_2)^2}[(-\frac{1}{m_1} + \frac{y}{1+m_1} - 1)^2 + 2y(-\frac{1}{m_1} + \frac{y}{1+m_1} + 1)]dm_1\\
=\ &\frac{4\pi i y^2}{(m_2+1)^3} - \frac{4\pi i y^2}{(m_2+1)^2} = -\frac{4\pi i y^2 m_2}{(m_2+1)^3},
\end{aligned}$$

$$-\frac{2y^2}{\pi i}\oint g(z(m))\frac{m}{(m+1)^3}dm = 4y^2,$$

and

$$\frac{1}{2\pi i}\int g(-\frac{1}{m}+\frac{y}{1+m})\frac{1}{(1+m)^2}dm = 0.$$

Then

$$Var(X_g) = 4y^2.$$

Above all, we can get

$$p\hat{W}_n \xrightarrow{D} \mathbb{N}(y, 4y^2).$$

By Slutsky's theorem, the proof of Theorem 2.2 is complete.

References

Anderson, T., 2003. An introduction to multivariate statistical analysis. Hoboken, NJ:Wiley.

Bai, Z., Jiang, D., Yao, J., Zheng, S., 2009. Corrections to LRT on large-dimensional covariance matrix by RMT. The Annals of Statistics 37 (6B), 3822–3840.

Bai, Z., Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. Statistica Sinica 6, 311–330.

Bai, Z., Silverstein, J., 2004. CLT for linear spectral statistics of large-dimensional sample covariance matrices. The Annals of Probability 32 (1A), 553–605.

Bai, Z., Silverstein, J., 2010. Spectral analysis of large dimensional random matrices. Springer Verlag.

Bai, Z., Wang, X., Zhou, W., 2010. Functional CLT for sample covariance matrices. Bernoulli 16 (4), 1086–1113.

Birke, M., Dette, H., 2005. A note on testing the covariance matrix for large dimension. Statistics & Probability letters 74 (3), 281–289.

Chen, S., Zhang, L., Zhong, P., 2010. Tests for high-dimensional covariance matrices. Journal of the American Statistical Association 105 (490), 810–819.

Fisher, T., Sun, X., Gallagher, C., 2010. A new test for sphericity of the covariance matrix for high dimensional data. Journal of Multivariate Analysis 101 (10), 2554–2570.

Ledoit, O., Wolf, M., 2002. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. The Annals of Statistics, 1081–1102.

Lin, Z., Xiang, Y., 2008. A hypothesis test for independence of sets of variates in high dimensions. Statistics & Probability Letters 78 (17), 2939–2946.

Lytova, A., Pastur, L., 2009. Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. The Annals of Probability 37 (5), 1778–1840.

Pan, G., 2010. Strong convergence of the empirical distribution of eigenvalues of sample covariance matrices with a perturbation matrix. Journal of Multivariate Analysis 101 (6), 1330–1338.

Pan, G., 2012. Comparison between two types of sample covariance matrices. Accpeted by Annales de l'Institut Henri Poincaré.

Pan, G., Zhou, W., 2008. Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. The Annals of Applied Probability 18 (3), 1232–1270.

Schott, J., 2006. A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. Journal of Multivariate Analysis 97 (4), 827–843.

Silverstein, J., Bai, Z., 1995. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. Journal of Multivariate analysis 54 (2), 175–192.

Srivastava, M., 2005. Some tests concerning the covariance matrix in high dimensional data. J. Japan Statist. Soc 35 (2), 251–272.