CORRECTIONS TO LRT ON LARGE-DIMENSIONAL COVARIANCE MATRIX BY RMT

WEIHAO LI

1. Introduction

Classical statistical method rely on large sample asymptotic theory, generally we assume dimension p is fixed and sample size n goes to infinity under this regime. However, with the advent of better computational tools, high dimension data become popular in many different areas, the classical method may fails when dimension p become large. Basically, researcher investigate the high dimension data under two framework, one is random matrix theory which assume both p, n go to infinity but the ratio of p and n converge to some positive number, another framework is to assume true parameters enjoy some low dimensional structures(such as sparsity, low rank, etc) and derive some convex optimization methods to recover the truth from samples ([PA14]).

Covariance matrix is important in multivariate statistics and many different statistical applications. Testing for equality of covariance matrix and identity matrix is especially crucial since many statistical models rely on the independence assumption between covariates. However, classical method such as likelihood ratio test may fail in high dimensional regime. Many researchers were devoted to developing different methods to conduct hypothesis testing for special covariance structure. Testing the bandedness of the covariance matrix of a high-dimensional Gaussian distribution was investigated in [CJ$^+$11], two samples covariance matrix test with application in high dimensional genomic studies was proposed in [LC12]. Above listed two method can be applied to large p small n scenario but they are not remedies to classical method, which means that they made different assumptions compared with classical method.

In this project, we focus on a remedy for classical likelihood ratio test by random matrix theory([BJY$^+$09]), in which new limiting distribution of same test statistic is derived. The organization of this paper is following: first I will introduce the problem setting for testing the equality of covariance matrix and then introduce likelihood ratio test and do a simulation to show the result become very bad when dimension p increase. Second, I will talk about how to derive an alternative test using random matrix theory and do a simulation to show the result is always good as dimension p goes up. Code is attached in the appendix.

2. Classical Likelihood ratio test

We focus on the problem of one-sample covariance hypothesis test. Suppose we have $X_1, \cdots, X_n$ follow from p-dimension($p < n$) normal distribution $N(\mu_p, \Sigma_p)$, we want to test $H_0 : \Sigma_p = I_p$ based on sample covariance matrix S:

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

The sample covariance S is maximum likelihood estimator of Σ_p, which is consistent estimator for Σ_p as $n \to \infty, p$ is fixed. But in order to do inference and hypothesis testing, consistency is not enough, we need to derive some asymptotic distribution. Likelihhod ratio asymptotic provide a limiting Chi-square distribution, first we need to compute the maximum likelihood under null space and unconstrained space respectively.

$$\text{Null:} \quad L_N = \prod_{i=1}^{n}\left[(2\pi)^{-p/2}\exp(-\frac{1}{2}(X_i - \bar{X})^T(X_i - \bar{X}))\right]$$

$$\text{Unconstrained:} \quad L_U = \prod_{i=1}^{n}\left[(2\pi)^{-p/2}|S|^{\frac{1}{2}}\exp(-\frac{1}{2}(X_i - \bar{X})^T S_n^{-1}(X_i - \bar{X}))\right]$$

Using relation $Tr\left[(X_i - \bar{X})^T S_n^{-1}(X_i - \bar{X}))\right] = Tr\left[S_n^{-1}(X_i - \bar{X})(X_i - \bar{X})^T)\right]$, the likelihood ratio test statistic is give by:

$$T_n = -2\log(\frac{L_N}{L_U}) = n\left(Tr(S_n) - \log|S_n| - p\right) \tag{2.1}$$

From large sample asymptotic theory, we know that as n goes to infinity with p fixed, T_n will converge to $\chi^2_{p(p+1)/2}$ under H_0. Hence we can derive a hypothesis test with size α as following:

1. Compute S_n, T_n based on $X_1, \cdots, X_n$.
2. Compute $1-\alpha$ quantile of $\chi^2_{p(p+1)/2}$, denoted as C_α
3. If $T_n > C_\alpha$, reject H_0, otherwise accept.

The simulation was conducted with $n = 500$ and different p. For each (n, p) 1000 independent experiments were done and in each experiment I simulated $X_1, \cdots, X_n \sim N(0, I_p)$. Simulation result follows

Table 1. Simulation: χ^2-test based on Likelihood ratio,$\alpha = 0.05$

(p.n)	(10,500)	(50,500)	(100,500)	(200,500)
proportion of rejection	0.042	0.255	0.982	1.000

Theoretically, we are expected to make 5% mistakes(rejections), but from above table we can see the performance of LR test become worse as p become larger. This phenomenon suggests that the large sample asymptotic theory may not be appropriate when p is large or p and n are roughly same order.

3. Remedy by random matrix theory

Given celebrated Marčenko-Pastur law of normalized Wishart matrix ([MP67]), the behavior of large dimension covariance matrix is expected to be better understand. Let $Z_i = X_i - \mu_p \sim N(0, I_p), \tilde{S}_n = \frac{1}{n}\sum Z_i Z_i^T$, the empirical spectral distribution of $\tilde{S}_n$ denote as F_n. As $p, n \to \infty, \frac{p}{n} \to y \in (0, 1)$, F_n will converge to Marčenko-Pastur law F^y. Our focus is to derive a new hypothesis test, parallel to classical hypothesis testing, we are going to find some limiting distribution under H_0, and

then use specific quantile of limiting distribution and test statistic to conduct the hypothesis test.
Since the limiting distribution is based on $\tilde{S}_n$, we first argue that S_n and $\tilde{S}_n$ are equivalent asymptotically.

$$\begin{aligned} S_n = \tilde{S}_n &+ \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_p)(\mu_p-\bar{X})^T + \frac{1}{n}\sum_{i=1}^{n}(\mu_p-\bar{X})(X_i-\mu_p)^T \\ &+ (\mu_p-\bar{X})(\mu_p-\bar{X})^T \end{aligned}$$

Last three terms will vanish as $n \to \infty$, so S_n and $\tilde{S}_n$ have same limiting spectral distribution. It suffices to focus on $\tilde{S}_n$.
Recall the format of T_n and define $\tilde{L}^* := Tr(\tilde{S}_n) - \log(|\tilde{S}_n|) - p$, it can be shown that $\tilde{L}^*$ can be written as functional of empirical spectral distribution. Parallel to classical central limit theorem which produce the asymptotic distribution of $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$, in random matrix theory it is also possible to derive some asymptotic Gaussian distribution for functional of empirical spectral distribution. It was shown that $\tilde{L}^*$ converge to Gaussian distribution in [BS04], I will discuss derivation of asymptotic mean and variance in subsequent section.

$$\begin{aligned} \tilde{L}^* &= Tr(\tilde{S}_n) - \log(|\tilde{S}_n|) - p \\ &= \sum_{i=1}^{p}\left(\lambda_i^{\tilde{S}_n)} - \log\lambda_i^{\tilde{S}_n)} - 1\right) = p\cdot\int(x-\log x-1)dF_n(x) \\ &= p\cdot\int g(x)d\left(F_n(x)-F^{y_n}(x)\right) + p\cdot F^{y_n}(g) \end{aligned}$$

where $g(x) = x - \log(x) - 1$. First term is an empirical process denoted as $G_n(g)$, second term is an deterministic term which we will compute later. Here we take following theorem in [BS04] as granted and only find mean and variance.

Theorem 3.1. *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 $n, p \to \infty$. Assume $\{\xi_{ij}\}$ are real and $E\left(\xi_{11}^4\right) = 3$. Then the random vector $\left(G_n(f_1), \ldots, G_n(f_k)\right)$ weakly converges to a k-dimensional Gaussian vector with mean vector:*
(3.1)

$$m\left(f_j\right) = \frac{f_j(a(y)) + f_j(b(y))}{4} - \frac{1}{2\pi}\int_{a(y)}^{b(y)}\frac{f_j(x)}{\sqrt{4y-(x-1-y)^2}}dx, \quad j = 1, \ldots, k$$

and covariance function

$$v\left(f_j, f_\ell\right) = -\frac{1}{2\pi^2}\oint\oint\frac{f_j\left(z_1\right)f_\ell\left(z_2\right)}{\left(\underline{m}\left(z_1\right)-\underline{m}\left(z_2\right)\right)^2}d\underline{m}\left(z_1\right)d\underline{m}\left(z_2\right) \quad j, \ell \in \{1, \ldots, k\} \tag{3.2}$$

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 (3.2) are nonoverlapping and both contain the support of F^y.

By Theorem 3.1 $G_n(g)$ weakly converges to a Gaussian random variable with the mean

$$m(g) = \frac{g(a(y)) + g(b(y))}{4} - \frac{1}{2\pi}\int_{a(y)}^{b(y)}\frac{g(x)}{\sqrt{4y-(x-1-y)^2}}dx \tag{3.3}$$

Where $a(y) = (1-\sqrt{y})^2, b(y) = (1+\sqrt{y})^2$. Using change of variable $x = 1 + y - 2\sqrt{y}\cos\theta, 0 \leq \theta \leq \pi, dx = 2\sqrt{y}\sin(\theta)d\theta$ we have

$$\begin{aligned}
m(g) &= \frac{y-\log(1-y)}{2} - \frac{1}{2\pi}\int_0^\pi \frac{y-2\sqrt{y}\cos(\theta)-\log(1+y-2\sqrt{y}\cos(\theta))}{\sqrt{4y\sin^2(\theta)}} 2\sqrt{y}\sin(\theta)d\theta \\
&= \frac{y-\log(1-y)}{2} - \frac{1}{2\pi}\int_0^\pi y-2\sqrt{y}\cos(\theta)-\log(1+y-2\sqrt{y}\cos(\theta))d\theta \\
&= \frac{-\log(1-y)}{2} - \frac{1}{2\pi}\frac{1}{2}\int_0^{2\pi} -\log(1+y-2\sqrt{y}\cos(\theta))d\theta \\
&= \frac{-\log(1-y)}{2} + \frac{1}{4\pi}\int_0^{2\pi} \log(1+y\sin^2(\theta)+y\cos^2(\theta)-2\sqrt{y}\cos(\theta))d\theta \\
&= \frac{-\log(1-y)}{2} + \frac{1}{4\pi}\int_0^{2\pi} \log[(1-\sqrt{y}\cos(\theta))^2+y\sin^2(\theta)]d\theta \\
&= \frac{-\log(1-y)}{2} + \frac{1}{4\pi}\int_0^{2\pi} \log|1-\sqrt{y}e^{i\theta}|^2 d\theta
\end{aligned}$$

We need to compute $\frac{1}{2\pi}\int_0^{2\pi}\log|1-\sqrt{y}e^{i\theta}|^2 d\theta$, the format of this integral remind us to use Gauss mean value theorem in complex integral. Gauss's mean-value theorem says that if a function u is analytic on and inside a circle of radius r centered at a point z_0, then

$$u(z_0) = \frac{1}{2\pi}\int_0^{2\pi} u\left(z_0 + re^{i\theta}\right) d\theta \tag{3.4}$$

If we take $u(z) = \log|1-\sqrt{y}z|^2, z_0 = 0, r = 1$, we can get that

$$0 = \log|1| = \frac{1}{2\pi}\int_0^{2\pi} \log|1-\sqrt{y}e^{i\theta}|^2 d\theta \tag{3.5}$$

Hence it suffices to check function $u(z) = \log|1-\sqrt{y}z|^2$ is analytic on the unit circle in $\mathbb{C}$. Since $\sqrt{y} < 1$, the singularity point of derivative $u^{(k)}(z), k \in \mathbb{N}$ all lie outside the unit circle, we can approximate the $u(z)$ with Taylor series centered at any point inside the complex unit circle. That is to say, our function $u(z)$ is analytic on the complex unit circle. In original paper [BS04], instead of mean value theorem they apply Poisson's integral formula, which is given by

$$u(z) = \frac{1}{2\pi}\int_0^{2\pi} u\left(e^{i\theta}\right) \frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)} d\theta \tag{3.6}$$

where u is harmonic on the unit disk in $\mathbb{C}$, and $z = re^{i\phi}$ with $r \in [0,1)$. Take $u(z) = \log|1-\sqrt{y}z|^2$ and $r = 0$, we get (3.5) again. Let $z = a + ib$ it suffices to check $u(a,b)$ is a harmonic function. We can check this function u satisfies Laplace's equation. Let $u(a,b) = \log|1-\sqrt{y}a-\sqrt{y}bi|^2 = \log[(1-\sqrt{y}a)^2 + yb^2]$. It reduces to check $\frac{\partial^2 u}{\partial a^2} + \frac{\partial^2 u}{\partial b^2} = 0$.

$$\frac{\partial u}{\partial a} = \frac{-2\sqrt{y}(1-\sqrt{y}a)}{(1-\sqrt{y}a)^2 + yb^2}, \frac{\partial u}{\partial b} = \frac{2yb}{(1-\sqrt{y}a)^2 + yb^2}$$
$$\frac{\partial y^2}{\partial a^2} = \frac{2y^2b^2 - 2y(1-\sqrt{y}a)^2}{(1-\sqrt{y}a)^2 + yb^2}, \frac{\partial y^2}{\partial b^2} = \frac{2y(1-\sqrt{y}a)^2 - 2y^2b^2}{(1-\sqrt{y}a)^2 + yb^2}$$
$$\Longrightarrow \frac{\partial^2 u}{\partial a^2} + \frac{\partial^2 u}{\partial b^2} = 0$$

The integral $\frac{1}{2\pi}\int_0^{2\pi} \log|1-\sqrt{y}e^{i\theta}|^2 d\theta$ vanish. We conclude that $m(g) = \frac{-\log(1-y)}{2}$. The approach by Poisson's integral formula is an alternative way to compute the mean $m(g)$, later I will use it to find another integral, for that integral I only know how to use Poisson's integral formula to solve the problem.
Next we are going to find the variance $v(g,g)$ for $g(x) = x - \log(x) - 1$.

$$v(g,g) = -\frac{1}{2\pi^2} \oint \oint \frac{g(z_1)\, g(z_2)}{\left(\underline{m}(z_1) - \underline{m}(z_2)\right)^2} d\underline{m}(z_1)\, d\underline{m}(z_2) \tag{3.7}$$

By proposition 3.6 in [YZB15], we can rewrite the variance $v(g,g)$ as following:

$$\begin{aligned} \frac{1}{2}v(g,g) &= \lim_{r\downarrow 1} -\frac{1}{4\pi^2} \oint_{|\xi_1|=1} \oint_{|\xi_2|=1} \frac{g\left(\left|1+\sqrt{y}\xi_1\right|^2\right) g\left(\left|1+\sqrt{y}\xi_2\right|^2\right)}{\{\xi_1 - r\xi_2\}^2} d\xi_1 d\xi_2 \\ &:= \lim_{r\downarrow 1} J(g,g,r) \end{aligned} \tag{3.8}$$

Expand the numerator we can get many pieces of integrals, which will be slightly easier to compute them one by one:

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

Set $f_1(x) = x, f_2(x) = \log(x), f_3(x) = 1$, we can further decompose $J(g,g,r)$ as

$$\begin{aligned} J(g,g,r) = & J(f_1,f_1,r) - 2J(f_1,f_2,r) + J(f_2,f_2,r) \\ & - 2J(f_1,f_3,r) + 2J(f_2,f_3,r) + J(f_3,f_3,r) \end{aligned}$$

Detailed but lengthy argument in section 3.2.1 of [YZB15] show that

$$\begin{aligned} & J(f_1,f_1,r) = \frac{y}{r^2}, J(f_1,f_2,r) = \frac{y}{r^2}, J(f_2,f_2,r) = -\frac{1}{r}\log(1-\frac{y}{r}) \\ & J(f_1,f_3,r) = J(f_2,f_3,r) = J(f_3,f_3,r) = 0 \end{aligned} \tag{3.9}$$

Combine with (3.8) we know that

$$\begin{aligned} & \frac{1}{2}v(g,g) = \lim_{r\downarrow 1} -\frac{y}{r^2} - \frac{1}{r}\log(1-\frac{y}{r}) = -y - \log(1-y) \\ & \Longrightarrow v(g,g) = -2y - 2\log(1-y) \end{aligned}$$

Based on Theorem (3.1) and the computation of $m(g), v(g,g)$, we arrive at following theorem:

Theorem 3.2. *Given* $m(g) = -\frac{\log(1-y)}{2}, v(g,g) = -2\log(1-y) - 2y$

$$G_n(g) = \tilde{L}^* - p \cdot F^{y_n}(g) \Rightarrow N(m(g), v(g,g)) \tag{3.10}$$

To construct the hypothesis test, it remains to find the value of $F^{y_n}(g)$.

$$\begin{aligned}
F^{y_n}(g) = & \int_{a(y_n)}^{b(y_n)} \frac{x-\log x-1}{2\pi x y_n}\sqrt{\left(b\left(y_n\right)-x\right)\left(x-a\left(y_n\right)\right)}dx \\
& \text{Let } x = 1+y-2\sqrt{y}\cos\theta, 0\leq\theta\leq\pi, dx = 2\sqrt{y}\sin(\theta)d\theta \\
& = \frac{1}{2\pi y_n}\int_0^{\pi}\left[1-\frac{\ln\left(1+y_n-2\sqrt{y_n}\cos\theta\right)+1}{1+y_n-2\sqrt{y_n}\cos\theta}\right]4y_n\sin^2\theta d\theta \\
& = \frac{1}{2\pi}\int_0^{2\pi}\left[2\sin^2\theta-\frac{2\sin^2\theta}{1+y_n-2\sqrt{y_n}\cos\theta}\left(\ln\left|1-\sqrt{y_n}e^{i\theta}\right|^2-1\right)\right]d\theta \\
& = -\frac{1}{2\pi}\int_0^{2\pi}\left[\frac{2\sin^2\theta}{1+y_n-2\sqrt{y_n}\cos\theta}\left(\ln\left|1-\sqrt{y_n}e^{i\theta}\right|^2\right)\right]d\theta
\end{aligned}$$

This time the final integral become hard, from Poisson's integral formula(3.6) we realize $y_n = r^2, \phi = 0$, $u(e^{i\theta})$ should be equal to $2\sin^2\theta\ln\left|1-\sqrt{y_n}e^{i\theta}\right|^2$. However, what should be the general form of harmonic function $u(z)$?
A slight complicated function $f(z)$ defined in [BS04] as following:

$$f(z) \equiv -\left(z-z^{-1}\right)^2\left(\log(1-\sqrt{y_n}z)+\sqrt{y_n}z\right)-\sqrt{y_n}\left(z-z^3\right) \tag{3.11}$$

The reason why they construct this function is because analytic functions have harmonic pieces(We can assume $f(z)$ is an analytic function,explain later), the real part and imaginary part are both harmonic function. Furthermore, the real part of $f(e^{i\theta})$ is given by (Here we recall $\log(z) = \ln|z| + i(Arg(z)+2k\pi)$)

$$\Re f\left(e^{i\theta}\right) = 2\sin^2\theta\ln\left|1-\sqrt{y_n}e^{i\theta}\right|^2 \tag{3.12}$$

Which is exactly what we want. Plug in $\Re f(e^{i\theta})$ to Poisson's integral formula with $r = \sqrt{y_n}, \phi = 0, z = re^{i\phi} = \sqrt{y_n}$, we can get:

$$\begin{aligned}
\Re f(\sqrt{y_n}) &= \frac{1}{2\pi}\int_0^{2\pi}\left[\frac{1-\sqrt{y_n}}{1+y_n-2\sqrt{y_n}\cos\theta}2\sin^2\theta\left(\ln\left|1-\sqrt{y_n}e^{i\theta}\right|^2\right)\right]d\theta \\
\Longrightarrow F^{y_n}(g) &= -\frac{\Re f(\sqrt{y_n})}{1-\sqrt{y_n}} = 1-\frac{y_n-1}{y_n}\log\left(1-y_n\right)
\end{aligned}$$

Now let's briefly explain why $f(z)$ is an analytic function. As we can see $f(z)$ is a combination of some fundamental functions such as z, z^3, which is analytic. But we also notice that non analytic function(on unit disk) $\frac{1}{z}, \log(z)$ is also in our expression. Intuitively, $f(z)$ can not be analytic because it contains some non analytic function, however if we apply Taylor expansion $\log(1-z)+z = -\frac{z^2}{2}+O\left(z^3\right)$, we will notice that some cancellation happens between $(z-\frac{1}{z})^2$ and $\log(1-z)+z$ which force the whole term together be infinitely differentiable, thus analytic on complex unit disk.
Besides above explanation, I also check $\Re f(z)$ is actually a harmonic function by verifying $\Re f(z)$ satisfies Laplace's equation, i.e

$$z = x+iy, f(z) \equiv f(x,y) = f_1(x,y)+if_2(x,y), \quad \frac{\partial^2 f_1}{\partial x^2}+\frac{\partial^2 f_1}{\partial y^2} = 0$$

However, the computation is too lengthy I do not show it here. Up to now, we arrive following Theorem:

Theorem 3.3. *Assume* $\frac{p}{n} \to y \in (0,1)$, *under null hypothesis* $H_0 : \Sigma_p = I_p$

$$\tilde{L}^* - p\left(1 - \frac{y_n - 1}{y_n}\log(1 - y_n)\right) \Rightarrow N\left(-\frac{\log(1-y)}{2}, -2\log(1-y) - 2y\right)$$

In simulation and real data, we just assume $y_n = y$ in order to use Theorem 3.3. The testing procedure for $H_0 : \Sigma_p = I_p$ follows:

(1) Compute $L^* = Tr(S_n) - \log|S_n| - p, m_n = \frac{\log(1-y_n)}{2}, v_n = -2\log(1-y_n) - 2y_n$ and define

$$T := v_n^{-1/2}\left[L^* - m_n - p\left(1 - \frac{y_n - 1}{y_n}\log(1 - y_n)\right)\right]$$

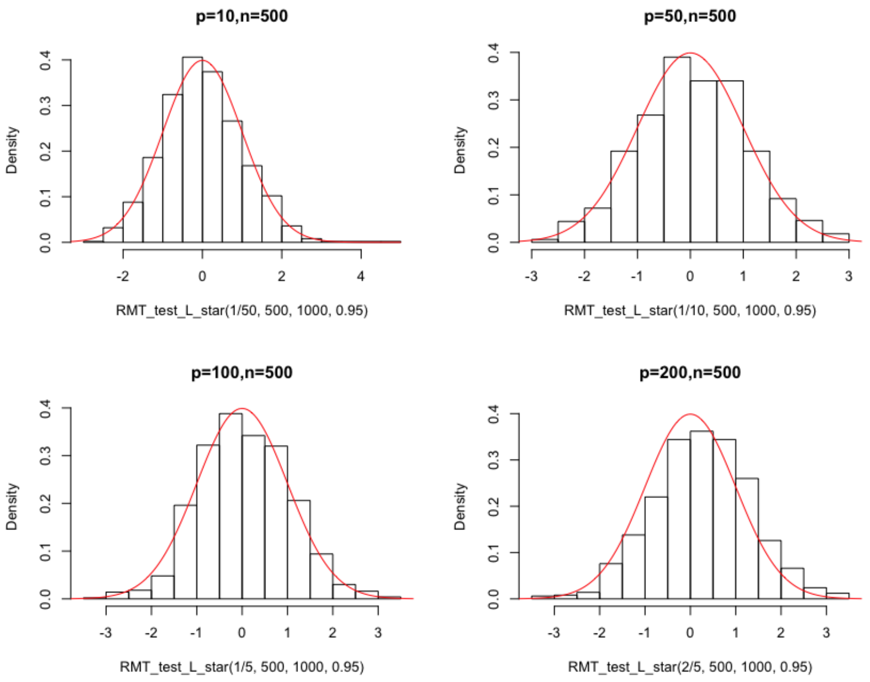
(2) Compute critical value $C_\alpha = -z_{\alpha/2}$, where $z_{\alpha/2}$ is $\alpha/2$ quantile of standard Gaussian distribution.

(3) If $T > C_\alpha$, reject H_0, otherwise accept.

Table 2. Simulation: Correction based on Random matrix theory,$\alpha = 0.05$

(p.n)	(10,500)	(50,500)	(100,500)	(200,500)
proportion of rejection	0.061	0.043	0.067	0.063

After we modify the asymptotic distribution of likelihood ratio statistic via random matrix theory, the simulation result behave well as we expected, we make around 5% mistakes no matter how large the dimension p is. Also we can see the distribution of L^* is closed to standard Gaussian(red curve).



4. Conclusion and Discussion

Compared to classical asymptotic Chi-square distribution, the modification via random matrix theory to a limiting Gaussian distribution seems to be more reasonable in high dimensional case.

The limiting ratio between p and n converge to $y \in (0,1)$ in this project. The critical case $y = 1$ was investigated in [JJY12] through Selberg integral. Another interesting topic covered in [ZBY^{+}15] is called substitution for unbiased sample covariance estimator. The sample covariance matrix S_n defined in this project is not an unibased estimator, the unbiased sample covariance matrix is defined to be

$$\bar{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

It seems that there is no difference between $n-1$ and n, but as we will multiply p to get the limiting distribution, and p, n are same order, which will make a difference. We can see from below argument:

Define $\mu_{S_n}(g) := \int g(x) dF_n^{S_n}(x)$, we have shown

$$p\left\{\mu_{S_n}(g) - F^{y_n}(g)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}(m(g), v(g)) \tag{4.1}$$

Also we have following decomposition:

$$\mu_{\bar{S}_n}(g) - F^{y_n}(g) = \left(\mu_{S_n}(g) - F^{y_n}(g)\right) + \left(\mu_{S_n}(g) - \mu_{\bar{S}_n}(g)\right)$$

Since $S_n = (1 - \frac{1}{n})\bar{S}_n$:

$$p \cdot \left(\mu_{S_n}(g) - \mu_{\bar{S}_n}(g)\right) = \sum_{i=1}^{p}\left\{g\left((1-\frac{1}{n})\lambda_i\right) - g\left(\lambda_i\right)\right\} \to -yF^y\left(\lambda g'(\lambda)\right)$$

Hence there is an non-varnishing term $yF^y\left(\lambda g'(\lambda)\right)$ show up in the asymptotic mean. We need to be careful when we do analysis and code programming.

References

[BJY$^+$09] Zhidong Bai, Dandan Jiang, Jian-Feng Yao, Shurong Zheng, et al. Corrections to lrt on large-dimensional covariance matrix by rmt. *The Annals of Statistics*, 37(6B):3822–3840, 2009.

[BS04] ZD Bai and Jack W Silverstein. Clt for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability*, 32(1A):553–605, 2004.

[CJ$^+$11] T Tony Cai, Tiefeng Jiang, et al. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39(3):1496–1525, 2011.

[JJY12] Dandan Jiang, Tiefeng Jiang, and Fan Yang. Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference*, 142(8):2241–2256, 2012.

[LC12] Jun Li and Song Xi Chen. Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940, 2012.

[MP67] VA Marchenko and LA Pastur. The distribution of eigenvalues in certain sets of random matrices math. *Sbornik*, 72:507–536, 1967.

[PA14] Debashis Paul and Alexander Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

[YZB15] Jianfeng Yao, Shurong Zheng, and ZD Bai. *Sample covariance matrices and high-dimensional data analysis*. Cambridge University Press Cambridge, 2015.

[ZBY$^+$15] Shurong Zheng, Zhidong Bai, Jianfeng Yao, et al. Substitution principle for clt of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *Annals of Statistics*, 43(2):546–591, 2015.

5. Appendix

Listing 1. Simulation code

```
# package: compute the trace of matrix
library(psych)
LR_test<-function(n,p,replication_number,alpha_size){
  test_result<-rep(0,replication_number)
  degree_freedom=p*(p+1)/2
  for(i in 1:replication_number){
    D_xT=matrix(rnorm(n*p),nc=p)
    cov_S=cov(D_xT)*(n-1)/n
    log_det=determinant(cov_S,logarithm = TRUE)$modulus
    log_det=as.numeric(log_det)
    T_n=n*tr(cov_S)-n*log_det-n*p
    if(T_n>qchisq(alpha_size,degree_freedom))
      test_result[i]=1
  }
  return(mean(test_result))
}

RMT_test<-function(p_over_n,n,replication_number,alpha_size){
  p=n*p_over_n
  test_result<-rep(0,replication_number)
  for(i in 1:replication_number){
    D_xT=matrix(rnorm(n*p),nc=p)
    cov_S=cov(D_xT)*(n-1)/n
    log_det=determinant(cov_S, logarithm = TRUE)$modulus
    log_det=as.numeric(log_det)
    L_tilde_star=tr(cov_S)-log_det-p
    mean_g=-log(1-p_over_n)/2
    v_g=-2*log(1-p_over_n)-2*p_over_n
    temp=(p_over_n -1)/p_over_n*log(1-p_over_n)
    test_stat=(L_tilde_star -p*(1-temp)-mean_g)/sqrt(v_g)
    if(abs(test_stat/qnorm(1-(1-alpha_size)/2))>1)
      test_result[i]=1
  }
  return(mean(test_result))
}

set.seed(1)
LR_test(500,10,1000,0.95)
LR_test(500,50,1000,0.95)
LR_test(500,100,1000,0.95)
RMT_test(1/50,500,1000,0.95)
RMT_test(1/10,500,1000,0.95)
RMT_test(1/5,500,1000,0.95)
RMT_test_L_star<-function(p_over_n,n,replication_number,alpha_size){
  p=n*p_over_n
  test_result<-rep(0,replication_number)
  test_stat_L_star<-NULL
  for(i in 1:replication_number){
    D_xT=matrix(rnorm(n*p),nc=p)
    cov_S=cov(D_xT)*(n-1)/n
    log_det=determinant(cov_S, logarithm = TRUE)$modulus
    log_det=as.numeric(log_det)
    L_tilde_star=tr(cov_S)-log_det-p
    mean_g=-log(1-p_over_n)/2
    v_g=-2*log(1-p_over_n)-2*p_over_n
    temp=(p_over_n -1)/p_over_n*log(1-p_over_n)
    test_stat=(L_tilde_star -p*(1-temp)-mean_g)/sqrt(v_g)
    test_stat_L_star<-c(test_stat_L_star, test_stat)
  }
  return(test_stat_L_star)
}
set.seed(1)
x=seq(-5,5,length.out = 200);y=dnorm(x)
par(mfrow=c(2,2))
hist(RMT_test_L_star(1/50,500,1000,0.95),freq=FALSE,main = ”p=10,n=500”)
lines(x,y,col=2)
hist(RMT_test_L_star(1/10,500,1000,0.95),freq=FALSE,main = ”p=50,n=500”)
lines(x,y,col=2)
```

```
hist(RMT_test_L_star(1/5,500,1000,0.95),freq=FALSE,main = ”p=100,n=500”)
lines(x,y,col=2)
hist(RMT_test_L_star(2/5,500,1000,0.95),freq=FALSE,main = ”p=200,n=500”)
lines(x,y,col=2)
```