Central Limit Theorem for Testing the Equality of Covariance Matrices

A PROJECT
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA DULUTH
BY

Wenchuan Guo

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisor: Yongcheng Qi

May, 2014

Acknowledgements

I would first like to thank my advisor, Professor Yong-cheng Qi. I am deeply indebted to him for his support and commitment, his enthusiastic guidance, his insight throughout the writing process. Furthermore, I would like to thank him for supporting me as research assistant for this project in the second year. In addition, I would like to thank Dr. Barry James and Dr. Marshall Hampton for willing to serve on my committee, taking the time to read this paper, and hearing the defense. Last but not least, I would like to profusely thank the faculty, staff, and fellow students in the math department whose contributions were always helpful and illuminating to me.

Abstract

In this project, we are interested in testing for the equality of k covariance matrices of p-dimensional multivariate normal distributions where the likelihood ratio test is used. The asymptotic distribution of the test statistics under different conditions is needed. In traditional multivariate analysis, the dimension p is considered to be a fixed constant that is unchanged as n approaches infinity. However, in practice, the dimension may be proportional to sample size n. In this project, we treat the dimension p as a function of sample size n and studied the hypothesis test under the general conditions that $\min_{1\leq i\leq k} n_i/p \geq c > 1$. We also assume the number of populations k to be a variable that changes with sample size n_i under the condition $k/n_i \to 0$. In this case, the limiting distribution of the log likelihood ratio test statistics when sample size n goes to infinity is no longer chi-square distribution. We derive the central limit theorem when p, k are variables. Numerical simulations for the two different approximations including histograms of simulated values and estimation of power and size are presented at the end of this project.

Keywords: Likelihood ratio test, central limit theorem, multi- variate normal distribution, multivariate Gamma function.

Contents

Acknowledgements **i**

Abstract **ii**

1 Introduction **1**

1.1 Likelihood ratio test 1

1.2 Main result 4

2 Preliminary results **5**

2.1 Basic definitions 5

2.2 Preliminary proposition and lemmas 6

3 Proof of the main result **16**

4 Numerical simulations **21**

4.1 Estimation of size 22

4.2 Estimation of power 23

References **24**

Appendix **25**

1 Introduction

High-dimensional data are increasingly common in many applications of statistics, encountered particularly in biological, meteorological, and financial studies. If a sample of size n comes from a population with dimension p, it is natural to set up an $n \times p$ data matrix. The analysis of such matrices includes hypothesis testing, interval estimation and modeling, and is elaborated in many classical textbooks of multivariate analysis such as Anderson [2], Muirhead [13], and Eaton [7]. In these books, the properties of these tests, estimations or models are based on the assumption that the data dimension p is a fixed constant or is negligible compared with the sample size n. However, due to the explosive development and improvement of computing power and detection methods in biology, the data dimension often is high. In this case, the former assumption of small p large n is no longer valid because the dimension p can be proportionally large compared with the sample size n. Some significant results have been achieved for this situation as reviewed in Bai *et al.* [3], Ledoit and Wolf [12], Jiang *et al.* [9], Jiang and Yang [11], Jiang and Qi [10]. In this project, we are interested in testing the equality of k covariance matrices of multivariate normal distributions under the assumption that $p \to \infty$ and $k \to \infty$.

1.1 Likelihood ratio test

In this subsection, we give the formal descriptions of the problem and introduce some results related to this topic.

Let $k \geq 2$ be an integer. We denote $N_p(\mu, \Sigma)$ as the multivariate normal distribution with dimension p, where μ is the mean vector and Σ is the covariance matrix. Consider the following hypothesis test:

$$H_0 : \Sigma_1 = \cdots = \Sigma_k = \Sigma \quad \text{vs} \quad H_a : H_0 \text{ is not true.} \tag{1.1}$$

Let $x_{i1}, \cdots, x_{in_i}$, for $1 \leq i \leq k$, be independent random samples from $N_p(\mu_i, \Sigma_i)$, where the mean vectors μ_i and covariance matrices Σ_i are unknown. Define

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \text{ and } A_i = \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)', \ 1 \leq i \leq k, \tag{1.2}$$

and

$$A = A_1 + \cdots + A_k \quad \text{and} \quad n = n_1 + \cdots + n_k. \tag{1.3}$$

We derive the likelihood ratio test for (1.1) first. The likelihood function for the k independent samples is given by

$$L(\mu_1,\ldots,\mu_k,\Sigma_1,\cdots,\Sigma_k)=\prod_{i=1}^{k}L(\mu_1,\Sigma_1)=\prod_{i=1}^{k}\left\{|\Sigma_i|^{-n_i/2}\cdot\mathrm{etr}\left(-\frac{1}{2}\Sigma_i^{-1}A_i\right)\right.$$
$$\left.\cdot\exp\left(-\frac{1}{2}n_i(\overline{x}_i-\mu_i)'\Sigma_i^{-1}(\overline{x}_i-\mu_i)\right)\right\}, \tag{1.4}$$

where etr is the exponential trace function, i.e., for a square matrix X, $\mathrm{etr}(X)=\exp(\mathrm{tr}(X))$. The likelihood ratio test is given as follows:

$$\Lambda_n=\frac{\sup_{\mu_1,\ldots,\mu_k,\Sigma}L(\mu_1,\ldots,\mu_k,\Sigma,\ldots,\Sigma)}{\sup_{\mu_1,\ldots,\mu_k,\Sigma_1,\ldots,\Sigma_k}L(\mu_1,\ldots,\mu_k,\Sigma_1,\ldots,\Sigma_k)}. \tag{1.5}$$

Observe that maximizing the likelihood amounts to maximizing each of the likelihoods in the product (1.4). Therefore, the denominator in (1.5) is

$$L(\overline{x}_1,\ldots,\overline{x}_k,\hat{\Sigma}_1,\ldots,\hat{\Sigma}_k)=e^{-pn/2}\prod_{i=1}^{k}\left(n_i^{-pn_i/2}|A_i|^{-n_i/2}\right), \tag{1.6}$$

where $\hat{\Sigma}_i=n_i^{-1}A_i$. When the null hypothesis $H_0:\ \Sigma_1=\cdots=\Sigma_k=\Sigma$ is true, we write the likelihood function in (1.4) as

$$L(\mu_1,\ldots,\mu_k,\Sigma,\cdots,\Sigma)=|\Sigma|^{-n/2}etr\left(-\frac{1}{2}\Sigma^{-1}A\right)$$
$$\cdot\prod_{i=1}^{k}\exp\left(-\frac{1}{2}n_i(\overline{x}_i-\mu_i)'\Sigma^{-1}(\overline{x}_i-\mu_i)\right), \tag{1.7}$$

which is maximized when $\mu_i=\overline{x}_i$ and $\Sigma=n^{-1}A$. Thus the numerator in (1.5) becomes

$$L(\overline{x}_1,\ldots,\overline{x}_k,n^{-1}A,\ldots,n^{-1}A)=n^{pn/2}e^{-pn/2}|A|^{-n/2}. \tag{1.8}$$

Combining (1.6) and (1.8) in (1.5) gives

$$\Lambda_n=\frac{\prod_{i=1}^{k}(|A_i|)^{n_i/2}}{(|A|)^{n/2}}\cdot\frac{n^{np/2}}{\prod_{i=1}^{k}n_i^{n_ip/2}}. \tag{1.9}$$

The likelihood ratio test (1.9) was first obtained by Wilks [17]. The details can be found in Section 8.2.2 of Muirhead [13].

By the result aforementioned, we reject the null hypothesis H_0 (1.1) as $\Lambda_n\le c_\alpha$, where c_α is chosen so that the test has the significance level of α. The test statistic Λ_n is non-degenerate

only if every A_i has full ranks which means that the condition $p < n_i$ for all $i = 1, \ldots, k$ is required by the fact that only with this condition, the determinants are nonzero. However, the drawback of this likelihood-ratio test is that when the sample sizes $n_1, \ldots, n_k$ are not all equal, it is biased, which implies that when H_0 is true, the probability of rejecting H_0 (1.1) can be larger than the probability of rejecting H_0 with H_0 being false. The bias was first noticed by Brown [6] with $p = 1$. By this fact, we use instead the following modified likelihood ratio test statistic Λ_n^* suggested by Bartlett [4]:

$$\Lambda_n^* = \frac{\prod_{i=1}^k (|A_i|)^{(n_i-1)/2}}{(|A|)^{(n-k)/2}} \cdot \frac{(n-k)^{(n-k)p/2}}{\prod_{i=1}^k (n_i-1)^{(n_i-1)p/2}}. \tag{1.10}$$

Λ_n^* is derived by substituting the sample size n_i by the degree of freedom $n_i - 1$ and sum of n_i by $n - k$. It is proved by Sugiura and Nagao [15] that this modified likelihood ratio test is unbiased when $k = 2$. A generalization to general integer k is made in Perlman [14].

When p is a fixed number, a lot of previous research has developed large sample approximation for the limit distribution of Λ_n^*. Under the null hypothesis (1.1), Box [5] proved that $-2\rho \log \Lambda_n^*$ has a chi-square approximation:

$$-2\rho \log \Lambda_n^* \xrightarrow{d} \chi_f^2 \tag{1.11}$$

as $\min_{1 \le i \le k} n_i \to \infty$, where

$$\begin{aligned} f &= \frac{1}{2} p(p+1)(k-1), \\ \rho &= 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)(n-k)} \left(\sum_{i=1}^k \frac{n-k}{n_i - 1} - 1 \right). \end{aligned} \tag{1.12}$$

Actually, the cumulative distribution of $-2\rho \log \Lambda_n^*$ can be expanded as

$$P(-2\rho \log \Lambda_n^* \le x) = P(\chi_f^2 \le x) + \frac{\gamma}{M^2}[P(\chi_{f+4}^2) \le x - P(\chi_f^2 \le x)] + O(M^{-3}), \tag{1.13}$$

as $n \to \infty$, where $M = \rho(n-k)$ and

$$\gamma = \frac{p(p+1)}{48} \left\{ (p+2)(p-1) \left[\sum_{i=1}^k \frac{(n-k)^2}{(n_i-1)^2} - 1 \right] - 6(k-1)(n-k)^2(1-\rho)^2 \right\}. \tag{1.14}$$

The likelihood ratio test of size α based on the asymptotic distribution of $-2\rho \log \Lambda_n^*$ is to reject H_0 (1.1) if $-2\rho \log \Lambda_n^* > c_\alpha$, where c_α is the $100\alpha\%$ upper quantile of χ_f^2 distribution. The details can be found in [13, p. 309].

In recent studies, the assumption on the fixed dimension p is relaxed to be changing with the sample size n_i so that the ratio p/n_i converges to a constant between 0 and 1, i.e.,

$$\lim_{n\to\infty} \frac{p}{n_i} = y_i \in (0,1). \tag{1.15}$$

For instance, using the Random Matrix Theory method, the correction to the traditional likelihood ratio test for testing the equality of two covariance matrices is studied in Bai *et al.* [3] under assumption (1.15). Denote the sample covariance matrices for two populations as A and B. Let

$$L = \frac{|A|^{n_1/2}\cdot|B|^{n_2/2}}{|n_1A/n + n_2B/n|^{n/2}}. \tag{1.16}$$

Then the chi-square approximation of the test statistic is represented as

$$-2\log L \xrightarrow{d} \chi^2_{\frac{1}{2}p(p+1)}. \tag{1.17}$$

A special case with $y_i = 1$ is studied in Jiang and Yang [11]. The following central limit theorem (CLT)

$$\frac{\log\Lambda_n^* - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1), \tag{1.18}$$

where μ_n and σ_n are functions of n and p (see, e.g., [11, 10]), is obtained in Jiang and Yang [11], Jiang and Qi [10] under the assumption that the dimension p can be proportionally large to the sample size n_i or much smaller than n_i, that is, $p/n_i \to y_i \in [0,1]$ and $n_i - 1 \geq p$ for $1 \leq i \leq k$.

1.2 Main result

Throughout this project, we assume that both the dimension p and number of the populations k change along with the sample size n_i, but k is much smaller than n_i, that is, $k/n_i \to 0$. In addition, the condition $\min_{1\leq i\leq k} n_i > p$ is always assumed. We have the following Theorem 1.1.

Theorem 1.1. *Assume $n_i = n_i(p,k)$ for all $1 \leq i \leq k$ such that $p > 2$, $\min_{1\leq i\leq k} n_i/p \geq c > 1$, $k = o(\min_{1\leq i\leq k} n_i)$, where $\min_{1\leq i\leq k} n_i \to \infty$. Let Λ_n^* be as in (1.10). Then, under H_0 in (1.1), $(\log\Lambda_n^* - \mu_n)/\sigma_n$ converges to $N(0,1)$ as $p\to\infty$, $k\to\infty$, where*

$$\begin{aligned}\mu_n =& \frac{1}{2}\left[\sum_{i=1}^{k}\left(p - n_i + \frac{3}{2}\right)\log\left(1-\frac{p}{n_i}\right)(n_i-1)\right.\\ &\left.-(n-k)\left(p-n+k+\frac{1}{2}\right)\log\left(1-\frac{p}{n-k+1}\right)+(k-1)p\right],\end{aligned} \tag{1.19}$$

$$\sigma_n^2 = \frac{1}{2}\left[p(k-1) + (n-k)^2 \log\left(1 - \frac{p}{n-k+1}\right) - \sum_{i=1}^{k}(n_i - 1)^2 \log\left(1 - \frac{p}{n_i}\right)\right]. \quad (1.20)$$

Based on the above Theorem, the rejection region of the test (1.1) is $(\log \Lambda_n^* - \mu_n)/\sigma_n < -z_\alpha$, where z_α is $100\alpha\%$ upper quantile of the standard normal distribution. When k is a fixed constant, there is a similar theorem (Theorem 4) in [10], which is under the condition that $\min_{1\le i\le k} n_i > p - c$, where $1 \le c \le 4$.

In this project, we proceed as follows. The preliminary results used for the proof of Theorem 1.1 are presented in section 2. In section 2.1, we introduce the multivariate gamma function and Stirling formula. Section 2.2 is devoted to a key proposition, which is proved through several lemmas. In section 3, we present the proof of the main result. In section 4, we provide numerical simulations for the size and power of the tests and compare graphically the classical chi-square approximation with the normal approximation.

2 Preliminary results

2.1 Basic definitions

In this subsection, we introduce the multivariate gamma function and Stirling's Formula, as well as some standard notations.

Recall that the gamma function $\Gamma(z)$ is defined for complex numbers z with $\mathrm{Re}(z) > 0$ through a convergent improper integral:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt. \quad (2.1)$$

In multivariate analysis, the multivariate gamma function $\Gamma_p(z)$, defined on the complex plane $\mathbb{C}$, is a generalization of the univariate gamma function. The definition of $\Gamma_p(z)$ is as follow.

Definition 2.1. *The multivariate gamma function, denoted by $\Gamma_p(z)$, is defined to be*

$$\Gamma_p(z) = \int \exp(-\mathrm{trace}(S))|S|^{z-(p+1)/2} dS, \quad (2.2)$$

where $Re(z) \ge \frac{1}{2}(p-1)$ and $S_{p\times p}$ is positive-definite matrix (the integral is over the space of positive definite $p \times p$ matrices).

$\Gamma_p(z)$ can also be written as the product of univariate gamma functions, as the following theorem shows (see, e.g., Muirhead [13, p. 62]).

Theorem 2.1.

$$\Gamma_p(z) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left[z - \frac{1}{2}(i-1)\right], \tag{2.3}$$

for complex numbers z with $Re(z) > \frac{1}{2}(p-1)$.

By (2.3), it follows immediately that

$$\Gamma_p(z) = \pi^{(p-1)/2} \Gamma_{p-1}(z) \Gamma\left[z + \frac{(1-p)}{2}\right]. \tag{2.4}$$

Note that when $p = 1$, the multivariate gamma function reduces to the univariate gamma function. In addition, the multivariate gamma function appears in the probability density function of the Wishart distribution and inverse Wishard distribution. However, the formula (2.3) is much more useful in practice and is used throughout this project.

The logarithm of the gamma function defined on the complex plane $\mathbb{C}$ has the following expansion, which is called Stirling's formula (see, e.g., Gamelin [8, p. 368] or Equation (37) of Ahlfors [1, p. 204]):

$$\log \Gamma(z) = \left(z - \frac{1}{2}\right) \log z - z + \log \sqrt{2\pi} + \frac{1}{12z} + O\left(\frac{1}{x^3}\right), \tag{2.5}$$

as $x = \mathrm{Re}(z) \to \infty$. Formula (2.5) is a powerful tool that can be used to obtain the asymptotic approximation to $\Gamma(z)$ leading to an accurate approximation for factorials, see more details in Gamelin [8, p. 368].

The standard notation to describe the limiting behavior of a function is denoted by big O and small o. Precisely, for two sequences of numbers $\{a_n; n \geq 1\}$ and $\{b_n; n \geq 1\}$, $a_n = O(b_n)$ as $n \to \infty$ means that $\limsup_{\infty} |a_n/b_n| < \infty$ and $a_n = o(b_n)$ as $n \to \infty$ means that $\lim_{n\to\infty} a_n/b_n = 0$. The symbol $a_n \sim b_n$ stands for $\lim_{n\to\infty} a_n/b_n = 1$. For two functions f and g, we also use $f(x) = O(g(x))$, $f(x) = o(g(x))$, and $f(x) \sim g(x)$ with the similar meaning as $x \to x_0 \in [-\infty, \infty]$.

2.2 Preliminary proposition and lemmas

The key point to prove Theorem 1.1 is the following expansion for $\Gamma_p(z)$.

Proposition 2.1. *Let $p = p_m$ and $t = t_m$ satisfy that, when $m \to \infty$, (1) $p_m \to \infty$; (2) $t = t_m = O(m/(p\sqrt{k}))$; (3) $m/p_m \geq c > 1$. Then,*

$$\log \frac{\Gamma_p(\frac{m-1}{2} + t)}{\Gamma_p(\frac{m-1}{2})} = \alpha_m t + \beta_m t^2 + \gamma_m + I_m, \tag{2.6}$$

where

$$\begin{aligned}
\alpha_m &= \left(p - m + \frac{3}{2}\right)\log\left(1 - \frac{p}{m}\right) - 2p;\\
\beta_m &= -\frac{p}{m} - \log\left(1 - \frac{p}{m}\right);\\
\gamma_m &= p\left[-\frac{m}{2}\log\left(\frac{m}{2}\right) + \left(\frac{m}{2} + t\right)\log\left(\frac{m}{2} + t\right)\right];\\
I_m &= O\left(\frac{1}{mk^{1/2}}\right) + O\left(\frac{1}{pk}\right).
\end{aligned} \tag{2.7}$$

The proof of Proposition 2.1 is split into three lemmas. The following Lemma 2.1 is actually Lemma 5.1 of Jiang and Qi [10].

Lemma 2.1. *For any given $\delta \in (0,1)$,*

$$\log\frac{\Gamma(x+b)}{\Gamma(x)} = (x+b)\log(x+b) - x\log x - b - \frac{b}{2x} + O\left(\frac{b^2+1}{x^2}\right) \text{ as } x \to +\infty, \tag{2.8}$$

uniformly on $b \in [-\delta x, \delta x]$.

Lemma 2.2. *Let $p = p_m$ be such that $1 \le p < m$, $m/p \ge c > 1$ as $p \to \infty$. Then*

$$\sum_{i=1}^{p}\left(\frac{1}{m-i} - \frac{1}{m}\right) = -\frac{p}{m} - \log\left(1 - \frac{p}{m}\right) + O\left(\frac{p}{m^2}\right), \tag{2.9}$$

$$\sum_{i=1}^{p}[\log(m-i) - \log m] = \left(p - m + \frac{1}{2}\right)\log\left(1 - \frac{p}{m}\right) - p + O\left(\frac{p}{m^2}\right). \tag{2.10}$$

Proof. Note that

$$\sum_{i=1}^{p}\left(\frac{1}{m-i} - \frac{1}{m}\right) = \sum_{n=1}^{p}\frac{1}{m-i} - \frac{p}{m}. \tag{2.11}$$

By the asymptotic expansion of the harmonic series [16]:

$$\begin{aligned}
\sum_{i=1}^{p}\frac{1}{i} &= \log p + \gamma + \frac{1}{2p} - \frac{1}{12p^2} + \frac{1}{120p^4} - \cdots\\
&= \log p + \gamma + \frac{1}{2p} + O(p^{-2}),
\end{aligned} \tag{2.12}$$

as $p \to \infty$, where γ is the Euler's constant, we obtain

$$\begin{aligned}
&\sum_{i=1}^{p}\frac{1}{m-i} = \sum_{i=1}^{m}\frac{1}{i} - \sum_{i=1}^{m-p}\frac{1}{i} - \frac{1}{m} + \frac{1}{m-p}\\
&= \log m - \log(m-p) + \frac{p}{m(m-p)} - \frac{p}{2m(m-p)} + O\left(\frac{1}{m^2}\right) + O\left(\frac{1}{(m-p)^2}\right)\\
&= -\log\left(1 - \frac{p}{m}\right) + O\left(\frac{p}{m^2}\right) \text{ as } p \to \infty.
\end{aligned} \tag{2.13}$$

In the last step of (2.13), we used the fact

$$O\left(\frac{p}{m^2-mp}\right)+O\left(\frac{1}{m^2}\right)+O\left(\frac{1}{(m-p)^2}\right)=O\left(\frac{p}{m^2}\right). \tag{2.14}$$

Equation (2.9) then follows from (2.13).

Next we show (2.10). Since for any integer $n \geq 1$, $\Gamma(n)=(n-1)!$, we have

$$\sum_{i=1}^{p}\log(m-i)=\sum_{i=1}^{m-1}\log i-\sum_{i=1}^{m-p-1}\log i=\log\Gamma(m)-\log\Gamma(m-p). \tag{2.15}$$

Applying Stirling's formula (2.5) to $x=m$ and $x=m-p$ to obtain

$$\begin{aligned}
&\log\Gamma(m)-\log\Gamma(m-p)\\
&=\left(m-\frac{1}{2}\right)\log m-\left(m-p-\frac{1}{2}\right)\log(m-p)-p-\frac{1}{12}\left(\frac{1}{m}-\frac{1}{m-p}\right)+O\left(\frac{1}{m^3}\right)\\
&=\left(m-\frac{1}{2}\right)\log m-\left(m-p-\frac{1}{2}\right)\log(m-p)-p+O\left(\frac{p}{m^2}\right).
\end{aligned} \tag{2.16}$$

Therefore,

$$\begin{aligned}
&\sum_{i=1}^{p}[\log(m-i)-\log m]\\
&=-p\log m+\sum_{i=1}^{p}\log(m-i)\\
&=-p\log m+\left(m-\frac{1}{2}\right)\log m-\left(m-p-\frac{1}{2}\right)\log(m-p)-p+O\left(\frac{p}{m^2}\right)\\
&=\left(m-p-\frac{1}{2}\right)\log m-\left(m-p-\frac{1}{2}\right)\log(m-p)-p+O\left(\frac{p}{m^2}\right)\\
&=\left(p-m+\frac{1}{2}\right)\log\left(1-\frac{p}{m}\right)-p+O\left(\frac{p}{m^2}\right)
\end{aligned} \tag{2.17}$$

as $m\to\infty$. ■

Lemma 2.3. *Let $p=p_m$ be such that $1\leq p<m$, $m/p\geq c>1$ as $p\to\infty$. Define*

$$g_i(x)=\left(\frac{m-i}{2}+x\right)\log\left(\frac{m-i}{2}+x\right)-\left(\frac{m}{2}+x\right)\log\left(\frac{m}{2}+x\right), \tag{2.18}$$

for $1\leq i\leq p$ and $x>-(m-p)/2$. Set $t=O(m/(p\sqrt{k}))$, we have that

$$\begin{aligned}
&\sum_{i=1}^{p}(g_i(t)-g_i(0))\\
&=\left[\left(p-m+\frac{1}{2}\right)\log\left(1-\frac{p}{m}\right)-p\right]t+\left[-\frac{p}{m}-\log\left(1-\frac{p}{m}\right)\right]t^2+B,
\end{aligned} \tag{2.19}$$

as $m \to \infty$, where

$$B = O\left(\frac{1}{mk^{1/2}}\right) + O\left(\frac{1}{pk}\right). \tag{2.20}$$

Proof. A straightforward computation shows that

$$\begin{aligned} g_i'(x) &= \log\left(\frac{m-i}{2}+x\right) - \log\left(\frac{m}{2}+x\right), \\ g_i''(x) &= \left(\frac{m-i}{2}+x\right)^{-1} - \left(\frac{m}{2}+x\right)^{-1}, \\ g_i^{(3)}(x) &= -\left(\frac{m-i}{2}+x\right)^{-2} + \left(\frac{m}{2}+x\right)^{-2}. \end{aligned} \tag{2.21}$$

Hence, there exists a constant C that is independent of m, such that

$$\sup_{|x|\le|t|,\ 1\le i\le p} |g_i^{(3)(x)}| \le C\frac{p}{m(m-p)^2}. \tag{2.22}$$

We expand the function $g_i(t)$ at $t=0$ by Taylor's expansion, obtaining

$$\begin{aligned} g_i(t) &= g_i(0) + g_i'(0)t + \frac{t^2}{2}g_i''(0) + R_2(t) \\ &= g_i(0) + (\log(m-i) - \log(m-1))t + \left(\frac{1}{m-i} - \frac{1}{m-1}\right)t^2 + R_2(t), \end{aligned} \tag{2.23}$$

where R_2 is the remainder such that

$$|R_2(t)| \le C\frac{p}{m(m-p)^2}t^3. \tag{2.24}$$

Therefore,

$$\begin{aligned} &\sum_{i=1}^{p}(g_i(t) - g_i(0)) \\ =&t\sum_{i=1}^{p}(\log(m-i) - \log m) + t^2\sum_{i=1}^{p}\left(\frac{1}{m-i} - \frac{1}{m}\right) + O\left(\frac{p^2t^3}{m(m-p)^2}\right) \\ =&\left[\left(p - m + \frac{1}{2}\right)\log\left(1-\frac{p}{m}\right) - p\right]t + \left[-\frac{p}{m} - \log\left(1-\frac{p}{m}\right)\right]t^2 + A, \end{aligned} \tag{2.25}$$

where when $t = O(m/(p\sqrt{k}))$,

$$\begin{aligned} B &= O\left(\frac{p^2t^3}{m(m-p)^2}\right) + O\left(\frac{pt}{m^2-mp}\right) + O\left(\frac{pt^2}{m^2-mp}\right) \\ &= O\left(\frac{m^2}{pk^{3/2}(m-p)^2}\right) + O\left(\frac{1}{(m-p)k^{1/2}}\right) + O\left(\frac{m}{pk(m-p)}\right) \\ &= O\left(\frac{1}{mk^{1/2}}\right) + O\left(\frac{1}{pk}\right). \end{aligned} \tag{2.26}$$

■

Proof of Proposition 2.1. By the definition of Γ_p function in (2.3), we have

$$\log\frac{\Gamma_p(\frac{m-1}{2}+t)}{\Gamma_p(\frac{m-1}{2})}=\sum_{i=1}^{p}\log\frac{\Gamma(\frac{m-i}{2}+t)}{\Gamma(\frac{m-i}{2})}. \tag{2.27}$$

Consider the terms in "Σ" first. By Lemma 2.1, we write

$$\begin{aligned}\log\frac{\Gamma(\frac{m-i}{2}+t)}{\Gamma(\frac{m-i}{2})}=&\left(\frac{m-i}{2}+t\right)\log\left(\frac{m-i}{2}+t\right)-\frac{m-i}{2}\log\frac{m-i}{2}\\&-t-\frac{t}{m-i}+O\left(\frac{t^2+1}{m^2}\right).\end{aligned} \tag{2.28}$$

Performing sum with respect to i yields

$$\begin{aligned}\log\frac{\Gamma_p(\frac{m-1}{2}+t)}{\Gamma_p(\frac{m-1}{2})}=&\sum_{i=1}^{p}\left[\left(\frac{m-i}{2}+t\right)\log\left(\frac{m-i}{2}+t\right)-\frac{m-i}{2}\log\frac{m-i}{2}\right]\\&-tp-t\sum_{i=1}^{p}\frac{1}{m-i}+O\left(\frac{pt^2+p}{m^2}\right).\end{aligned} \tag{2.29}$$

Set

$$g_i(x)=\left(\frac{m-i}{2}+x\right)\log\left(\frac{m-i}{2}+x\right)-\left(\frac{m}{2}+x\right)\log\left(\frac{m}{2}+x\right). \tag{2.30}$$

By Lemma 2.3, we obtain

$$\begin{aligned}&\sum_{i=1}^{p}\left[\left(\frac{m-i}{2}+t\right)\log\left(\frac{m-i}{2}+t\right)-\frac{m-i}{2}\log\frac{m-i}{2}\right]\\=&p\left[-\frac{m}{2}\log\left(\frac{m}{2}\right)+\left(\frac{m}{2}+t\right)\log\left(\frac{m}{2}+t\right)\right]+\sum_{i=1}^{p}(g_i(t)-g_i(0))\\=&p\left[-\frac{m}{2}\log\left(\frac{m}{2}\right)+\left(\frac{m}{2}+t\right)\log\left(\frac{m}{2}+t\right)\right]\\&+\left[\left(p-m+\frac{1}{2}\right)\log\left(1-\frac{p}{m}\right)-p\right]t+\left[-\frac{p}{m}-\log\left(1-\frac{p}{m}\right)\right]t^2+A.\end{aligned} \tag{2.31}$$

This, together with (2.29), shows that

$$\begin{aligned}&\log\frac{\Gamma_p(\frac{m-1}{2}+t)}{\Gamma_p(\frac{m-1}{2})}\\=&\left[\left(p-m+\frac{3}{2}\right)\log\left(1-\frac{p}{m}\right)-2p\right]t+\left[-\frac{p}{m}-\log\left(1-\frac{p}{m}\right)\right]t^2\\&+p\left[-\frac{m}{2}\log\left(\frac{m}{2}\right)+\left(\frac{m}{2}+t\right)\log\left(\frac{m}{2}+t\right)\right]+I_m,\end{aligned} \tag{2.32}$$

where

$$I_m = B + O\left(\frac{pt^2+p}{m^2}\right) = B + O\left(\frac{1}{pk}\right) = O\left(\frac{1}{mk^{1/2}}\right) + O\left(\frac{1}{pk}\right). \tag{2.33}$$

■

To prove our Theorem 1.1, we also need study the σ_n^2 as defined in (1.20). We have the following lemma for σ_n^2.

Lemma 2.4. *Let $\sigma_n^2 = \sigma_n^2(k,p,n)$ be defined in (1.20). Assume $\min_{1\le i\le k} n_i > p$, $n_i > k \ge 2$, $p \ge 3$, then we have*

$$\sigma_n^2(k,p,n) \ge \frac{p^2}{36}(k-1). \tag{2.34}$$

The proof of lemma (2.4) is also split into three lemmas. The idea of the proof is to find the minimal value with respect to each individual variable k, n, p in σ_n^2. We start from $G(p) = 2\sigma_n^2$. In the following three lemmas, we treat p, n, k, n_i as continuous variables.

Lemma 2.5. *Define*

$$G(p) = p(k-1) + (n-k)^2 \log\left(1 - \frac{p}{n-k+1}\right) - \sum_{i=1}^{k}(n_i-1)^2 \log\left(1-\frac{p}{n_i}\right), \tag{2.35}$$

where $\min_{1\le i\le k} n_i > p$, $n_i > k \ge 2$, $p \ge 2$, then we have

$$\inf_{p\ge 1} G(p) = G(1). \tag{2.36}$$

Proof. It suffices to show that $G(p)$ is an increasing function with respect to p under the given condition. Differentiating (2.35) with respect to p, we obtain

$$G'(p) = k - 1 - \frac{(n-k)^2}{n-k+1-p} + \sum_{i=1}^{k}\frac{(n_i-1)^2}{n_i-p}. \tag{2.37}$$

Observe that

$$\begin{aligned} & n-k+1-p \ge n-kp \\ \Leftrightarrow\ & kp-k-p+1 \ge 0 \\ \Leftrightarrow\ & (k-1)(p-1) \ge 0 \\ \Leftrightarrow\ & k \ge 1 \text{ and } p \ge 1. \end{aligned} \tag{2.38}$$

This gives

$$\begin{aligned} (n-k+1-p)\sum_{i=1}^{k}\frac{(n_i-1)^2}{n_i-p} &\ge (n-kp)\sum_{i=1}^{k}\frac{(n_i-1)^2}{n_i-p} \\ &= \sum_{i=1}^{k}(n_i-p)\sum_{i=1}^{k}\frac{(n_i-1)^2}{n_i-p}. \end{aligned} \tag{2.39}$$

By Cauchy-Schwarz inequality we have

$$\sum_{i=1}^{k}(n_i-p)\sum_{i=1}^{k}\frac{(n_i-1)^2}{n_i-p}\geq\left(\sum_{i=1}^{k}(n_i-1)\right)^2=(n-k)^2. \tag{2.40}$$

Hence

$$\sum_{i=1}^{k}\frac{(n_i-1)^2}{n_i-p}-\frac{(n-k)^2}{n-k+1-p}\geq 0, \tag{2.41}$$

which implies

$$G'(p)\geq k-1\geq 1. \tag{2.42}$$

Therefore, $G(p)$ is an increasing function of p, which leads to (2.36). ■

Then, we fix $p=1$ in $G(p)$ and consider the following function $H(n_1,\ldots,n_k)=G(1)$.

Lemma 2.6. *Define*

$$H(n_1,\ldots,n_k)=(k-1)+(n-k)^2\log\left(1-\frac{1}{n-k+1}\right)-\sum_{i=1}^{k}(n_i-1)^2\log\left(1-\frac{1}{n_i}\right), \tag{2.43}$$

where $\min_{1\leq i\leq k}n_i>k\geq 2$. *Under the constraint* $\sum_{i=1}^{k}n_i=n$, *we have*

$$\inf_{n_i\geq 2}H(n_1,\cdots,n_k)=H\left(\frac{n}{k},\cdots,\frac{n}{k}\right). \tag{2.44}$$

Proof. To find the maximized or minimized value of function $H(n_1,\cdots,n_k)$ subject to the constraint $\sum_{i=1}^{k}n_i=n$ (this n is treated as a fixed constant), the method of Lagrange multipliers says to look for solutions of the following $k+1$ equations:

$$\begin{cases}\dfrac{\partial}{\partial n_i}\left(H(n_1,\cdots,n_k)-\lambda\left(\sum\limits_{i=1}^{k}n_i-n\right)\right)=0,\ 1\leq i\leq k,\\ \sum\limits_{i=1}^{k}n_i-n=0,\end{cases} \tag{2.45}$$

where λ is the Lagrange multiplier. Note that

$$\frac{\partial}{\partial n_i}\left(H(n_1,\cdots,n_k)-\lambda\left(\sum_{i=1}^{k}n_i-n\right)\right)=-2(n_i-1)\log\left(1-\frac{1}{n_i}\right)+\frac{1}{n_i}-\lambda-1 \tag{2.46}$$

and

$$\left[-2(n_i-1)\log\left(1-\frac{1}{n_i}\right)+\frac{1}{n_i}\right]'=-\frac{2n_i^2\log\left(1-\frac{1}{n_i}\right)+2n_i+1}{n_i^2}>0. \tag{2.47}$$

(2.47) implies that each equation is a monotone function and then, by applying the symmetric property of n_i, the solution of (2.45) is unique and must be $n_1 = n_2 = \cdots = n_k = n/k$. However, when $n_1 = n_2 = \cdots = n_k$, we don't know the function $H(n_1, \cdots, n_k)$ has minimal value or maximum value. To prove (2.44), we need check if the function attains minimal value at $(n/k, \cdots, n/k)$, for which it suffices to show

$$M = \left(\frac{\partial^2 H}{\partial n_i \partial n_j}\right)_{k\times k} \text{ is a positive definite matrix.} \tag{2.48}$$

Let $f(x) := (x-1)^2 \log(1-1/x), \forall\, x \geq 2$, then

$$f'(x) = 2(x-1)\log\left(1-\frac{1}{x}\right) + 1 - \frac{1}{x}, \tag{2.49}$$

and

$$f''(x) = 2\log\left(1-\frac{1}{x}\right) + \frac{2}{x} + \frac{1}{x^2}. \tag{2.50}$$

Using a taylor expansion of $\log(1-x)$ when $0 < |x| < 1$, we have

$$f''(x) = -2\sum_{i=1}^{\infty}\frac{1}{jx^j} + \frac{2}{x} + \frac{1}{x^2} = -2\sum_{i=3}^{\infty}\frac{1}{jx^j} < 0. \tag{2.51}$$

Note that

$$M_{ij} = \frac{\partial^2 H}{\partial n_i \partial n_j} = \begin{cases} -f''(n_i) > 0 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, \tag{2.52}$$

which leads to (2.48), then (2.44). ■

Then we let $n_i = n/k$ and consider $L(n) = H(n/k, \ldots, n/k)$.

Lemma 2.7. *Define*

$$L(n) = k - 1 + (n-k)^2 \log\left(1 - \frac{1}{n-k+1}\right) - \frac{(n-k)^2}{k}\log\left(1-\frac{k}{n}\right), \tag{2.53}$$

where $n > k \geq 2$. We have that

$$L(n) \geq \frac{1}{2}(k-1). \tag{2.54}$$

Proof. We show that $L(n)$ is a non-increasing function, in which case we have $L(n) \geq \lim_{n\to\infty} L(n)$. Set

$$l(n) = -2\log\left(1 - \frac{1}{n-k+1}\right) + \frac{2}{k}\log\left(1-\frac{k}{n}\right) - \frac{k-1}{n(n-k+1)}, n > k \geq 2. \tag{2.55}$$

The derivate function of $l(n)$ is given by

$$l'(n) = \frac{(k-1)\left(k^2 - k(n+1) - n\right)}{n^2(n-k)(n-k+1)^2}, \tag{2.56}$$

which is non-positive when $k < n$, implying $l(n)$ is a non-increasing function, thus

$$l(n) \geq \inf_{n>k\geq 2} l(n) = \lim_{n\to\infty} l(n) = 0. \tag{2.57}$$

A straightforward computation shows

$$\begin{aligned} L'(n) &= -(n-k)\left(-2\log\left(1 - \frac{1}{n-k+1}\right) + \frac{2}{k}\log\left(1-\frac{k}{n}\right) - \frac{k-1}{n(n-k+1)}\right) \\ &= -(n-k)l(n) \leq 0. \end{aligned} \tag{2.58}$$

Letting n go to infinity and using Taylor expansion of $\log(1-x)$ for $|x|< 1$, we get

$$\begin{aligned} \lim_{n\to\infty} L(n) &= \lim_{n\to\infty}\left[k - 1 - (n-k)^2\left(\frac{1}{n-k+1} + \frac{1}{2(n-k+1)^2}\right) + (n-k)^2\left(\frac{1}{n} + \frac{k}{2n^2}\right)\right] \\ &= \lim_{n\to\infty}\left[\frac{3}{2}(k-1) + \frac{(n-k)^2}{n} - \frac{(n-k)^2}{n-k+1}\right] \\ &= \lim_{n\to\infty}\left[\frac{3}{2}(k-1) + (1-k)\frac{(n-k)^2}{n(n-k+1)}\right] \\ &= \frac{1}{2}(k-1) > 0. \end{aligned} \tag{2.59}$$

Therefore, by (2.58), we get

$$L(n) \geq \inf_{n>k\geq 2} L(n) = \lim_{n\to\infty} L(n) = \frac{1}{2}(k-1). \tag{2.60}$$

■

Proof of Lemma 2.4. Let $G(p)$, $H(n_1,\ldots,n_k)$ and $L(n)$ be defined in (2.35), (2.43), (2.53) respectively. Note that $G(1) = H(n_1,\ldots,n_k)$ and $H(n/k,\ldots,n/k) = L(n)$, thus, by Lemmas 2.5, 2.6 and 2.7, we have

$$\inf_{p\geq 2}\frac{G(p)}{2} = \frac{G(1)}{2} \geq \inf_{n_i\geq 2}\frac{H(n_1,\ldots,n_k)}{2} \geq \inf_{n\geq 2}\frac{L(n)}{2} = \frac{k-1}{4}. \tag{2.61}$$

We then consider the following function

$$\begin{aligned} G_2(p) &= \frac{p(k-1) + (n-k)^2\log\left(1-\frac{p}{n-k+1}\right) - \sum_{i=1}^{k}(n_i-1)^2\log\left(1-\frac{p}{n_i}\right)}{p^2} \\ &= \frac{G(p)}{p^2}, \end{aligned} \tag{2.62}$$

where $G(p)$ is defined in (2.35). Note that $\sigma_n^2 = 2G(p)$. To show (2.34), it is suffice to show $G_2(p)$ is an increasing function with repsect to p, since when $G_2(p)$ is increasing, we can easily have, by (2.61),

$$\frac{\sigma_n^2}{p^2} = \frac{G_2(p)}{2} \geq \frac{G_2(3)}{2} = \frac{G(3)}{18} \geq \frac{G(1)}{18} = \frac{k-1}{36}. \tag{2.63}$$

Differentiate (2.62) with respect to p to obtain

$$G_2'(p) = \frac{G'(p)p - 2G(p)}{p^3}. \tag{2.64}$$

Since p is a positive integer, the positive or negative nature of $G_2'(p)$ is only dependent on $G'(p)p - 2G(p)$, i.e.

$$G_2'(p) \geq 0 \Leftrightarrow G'(p)p - 2G(p) \geq 0. \tag{2.65}$$

A direct calculation gives

$$\begin{aligned} G'(p)p - 2G(p) =& -\frac{p(n-k)^2}{n-k-p+1} - 2(n-k)^2 \log\left(1 - \frac{p}{n-k+1}\right) + p \\ &+ \sum_{i=1}^{k} \frac{p(n_i-1)^2}{n_i - p} + 2\sum_{i=1}^{k} (n_i-1)^2 \log\left(1 - \frac{p}{n_i}\right) - pk \\ =& q\left(\sum_{i=1}^{k} t_i\right) - \sum_{i=1}^{k} q(t_i), \end{aligned} \tag{2.66}$$

where $t_i = n_i - 1$ and for $x > p - 1 \geq 1$,

$$q(x) := -\frac{px^2}{x-p+1} - 2x^2 \log\left(1 - \frac{p}{x+1}\right) + p - px. \tag{2.67}$$

We write $q(x)$ as another form so that we can see that $q(x)$ is an increasing function with respect to x. Denote $\phi(x) = -\log(1-x) - x$, for $0 < x < 1$. Easily, we have that $\phi(x)$ is a positive increasing function, thus $x^2\phi(x)$ is increasing too. Then

$$\begin{aligned} q(x) &= p - px - \frac{px^2}{x-p+1} + \frac{2px^2}{x+1} + 2x^2\phi(x) \\ &= -p^2 + p\left[\frac{2}{x+1} - \frac{(p-1)^2}{x-p+1}\right] + 2x^2\phi(x). \end{aligned} \tag{2.68}$$

By evaluating the derivative function of the second term, it's easy to see that the second term is non-decreasing with respect to x when $p \geq 3$, which implies $q(x)$ (2.68) is also non-decreasing.

Letting x goes to infinity, we have that

$$\begin{aligned}\lim_{x\to\infty} q(x) &= -\frac{px^2}{x-p+1} - 2x^2\left(-\frac{p}{x+1} - \frac{p^2}{2(x+1)^2}\right) + p - px \\ &= \frac{p^3x^2 + p^2x + p^2 - px^2 - 2px - p}{(x+1)^2(p-x-1)} \\ &= 0,\end{aligned} \tag{2.69}$$

which implies $q(x) < 0$ for $x > p-1 \geq 2$. Then we find the one lower bound for $G'(p)p-2G(p)$ in (2.66) :

$$G'(p)p - 2G(p) \geq q(t_1) - \sum_{i=1}^{k} q(t_i) = -\sum_{i\neq 1}^{k} q(t_i) \geq 0. \tag{2.70}$$

This leads to (2.65), then (2.63) follows from (2.65). ■

3 Proof of the main result

Before beginning the proof of the main result, we state the following lemma to express the moments of W_n as a product of Γ_p functions, which is available in Muirhead [13, p. 302].

Lemma 3.1. *Let Λ_n^* be as in (1.10). Set*

$$W_n = \frac{\prod_{i=1}^{k}|A|^{(n_i-1)/2}}{|A|^{(n-k)/2}} = \Lambda_n^* \cdot \frac{\prod_{i=1}^{k}(n_i-1)^{(n_i-1)p/2}}{(n-k)^{(n-k)p/2}}. \tag{3.1}$$

Assume $n_i > p$ for $1 \leq i \leq k$. Then under H_0 in (1.1), we have

$$E(W_n^t) = \frac{\Gamma_p\left(\frac{1}{2}(n-k)\right)}{\Gamma_p\left(\frac{1}{2}(n-k)(1+t)\right)} \cdot \prod_{i=1}^{k} \frac{\Gamma_p\left(\frac{1}{2}(n_i-1)(1+t)\right)}{\Gamma_p\left(\frac{1}{2}(n_i-1)\right)}, \tag{3.2}$$

for all $t > \max_{1\leq i\leq k} \frac{p-1}{n_i-1} - 1$, where Γ_p is defined as in (2.3).

Proof of Theorem 1.1. By (3.1), we write

$$\log \Lambda_n^* = \log W_n + \frac{(n-k)p}{2}\log(n-k) - \sum_{i=1}^{k}\frac{(n_i-1)p}{2}\log(n_i-1). \tag{3.3}$$

Let

$$\mu_n' = \mu_n - \frac{(n-k)t}{2}\log\frac{n-k}{2} + \sum_{i=1}^{k}\frac{(n_i-1)t}{2}\log\frac{n_i-1}{2}. \tag{3.4}$$

To prove Theorem 1.1, it is equivalent to show that

$$\frac{\log W_n - \mu_n'}{\sigma_n} \xrightarrow{d} N(0,1), \tag{3.5}$$

as $n \to \infty$. Note that

$$\begin{aligned}
&\frac{\log W_n - \mu_n'}{\sigma_n} \xrightarrow{d} N(0,1) \\
\Leftrightarrow &E\exp\left\{\frac{\log W_n - \mu_n'}{\sigma_n}s\right\} = \log Ee^{t\log W_n} - \mu_n' t \to e^{s^2/2} \\
\Leftrightarrow &\log E(W_n^t) = \mu_n' t + \frac{1}{2}s^2 + o(1) = \mu_n' t + \frac{1}{2}\sigma_n^2 t^2 + o(1)
\end{aligned} \tag{3.6}$$

for all $|s| \leq 1$, where $t = s/\sigma_n$. By Lemma 3.1, we write the $\log E(W_n^t)$ as sum of $\log(\Gamma_p)$ function:

$$\begin{aligned}
&\log E(W_n^t) \\
=&-\log\frac{\Gamma_p\left(\frac{1}{2}(n-k)(1+t)\right)}{\Gamma_p\left(\frac{1}{2}(n-k)\right)} + \sum_{i=1}^{k}\log\frac{\Gamma_p\left(\frac{1}{2}(n_i-1)(1+t)\right)}{\Gamma_p\left(\frac{1}{2}(n_i-1)\right)} \\
=&-\log\frac{\Gamma_p\left(\frac{1}{2}(n-k)+\frac{1}{2}(n-k)t\right)}{\Gamma_p\left(\frac{1}{2}(n-k)\right)} + \sum_{i=1}^{k}\log\frac{\Gamma_p\left(\frac{1}{2}(n_i-1)+\frac{1}{2}(n_i-1)t\right)}{\Gamma_p\left(\frac{1}{2}(n_i-1)\right)}.
\end{aligned} \tag{3.7}$$

By Lemma 2.4, we have

$$\frac{\sigma_n^2}{kp^2} \geq \frac{1}{36}\left(1-\frac{1}{k}\right) \Rightarrow \lim_{n\to\infty}\frac{\sigma_n^2}{kp^2} = c > 0 \text{ or } \infty, \tag{3.8}$$

as $p \to \infty$, $k \to \infty$, which follows $t = O(1/(p\sqrt{k}))$ and when the limit is positive constant c or $t = o(1/(p\sqrt{k}))$ when the limit is infinity. Since $k/n \to 0$, $(n-k)t = O(n/(p\sqrt{k}))$ and $(n_i-1)t = O(n_i/(p\sqrt{k}))$. By Proposition 2.1, the first term on the last line of the left-hand side of (3.7) is

$$\begin{aligned}
&\log\frac{\Gamma_p\left(\frac{1}{2}(n-k)+\frac{1}{2}(n-k)t\right)}{\Gamma_p\left(\frac{1}{2}(n-k)\right)} \\
=&\gamma_{n-k+1} + \alpha_{n-k+1}\frac{(n-k)t}{2} + \beta_{n-k+1}\frac{(n-k)^2t^2}{4} + I_n \\
=&p\Bigg[-\frac{n-k+1}{2}\log\left(\frac{n-k+1}{2}\right) \\
&+\left(\frac{n-k+1}{2}+\frac{(n-k)t}{2}\right)\log\left(\frac{n-k+1}{2}+\frac{(n-k)t}{2}\right)\Bigg] \\
&+\left[\left(p-n+k+\frac{1}{2}\right)\log\left(1-\frac{p}{n-k+1}\right)-2p\right]\frac{(n-k)t}{2} \\
&+\left[-\frac{p}{n-k+1}-\log\left(1-\frac{p}{n-k+1}\right)\right]\frac{(n-k)^2t^2}{4} + I_n,
\end{aligned} \tag{3.9}$$

as $p \to \infty$, $k \to \infty$, where α_{n-k+1}, β_{n-k+1} and γ_{n-k+1} are defined in (2.7). Replacing $n-k+1$ with n_i to obtain a similar form for the second term on the last line of the right-hand side of (3.7)

$$\begin{aligned}
&\log \frac{\Gamma_p\left(\frac{1}{2}(n_i-1)+\frac{1}{2}(n_i-1)t\right)}{\Gamma_p\left(\frac{1}{2}(n_i-1)\right)} \\
=&\gamma_{n_i}+\alpha_{n_i}\frac{(n_i-1)t}{2}+\beta_{n_i}\frac{(n_i-1)^2t^2}{4}+I_n \\
=&p\left[-\frac{n_i}{2}\log\left(\frac{n_i}{2}\right)+\left(\frac{n_i}{2}+\frac{(n_i-1)t}{2}\right)\log\left(\frac{n_i}{2}+\frac{(n_i-1)t}{2}\right)\right] \\
&+\left[\left(p-n_i+\frac{3}{2}\right)\log\left(1-\frac{p}{n_i}\right)-2p\right]\frac{(n_i-1)t}{2} \\
&+\left[-\frac{p}{n_i}-\log\left(1-\frac{p}{n_i}\right)\right]\frac{(n_i-1)^2t^2}{4}+I_{n_i},
\end{aligned} \tag{3.10}$$

as $p \to \infty$, $k \to \infty$. Observe that

$$I_n = O\left(\frac{1}{nk^{1/2}}\right)+O\left(\frac{1}{pk}\right)=o(1) \tag{3.11}$$

and

$$\sum_{i=1}^{k} I_{n_i} = \sum_{i=1}^{k}\left[O\left(\frac{1}{n_ik^{1/2}}\right)+O\left(\frac{1}{pk}\right)\right] = o(1)+\sum_{i=1}^{k} O\left(\frac{1}{n_ik^{1/2}}\right)=o(1). \tag{3.12}$$

Then taking the sum with respect to i in (3.10) and adding all terms in (3.9) gives

$$\begin{aligned}
\log E(W_n^t) =& \sum_{i=1}^{k}\left[\gamma_{n_i}+\alpha_{n_i}\frac{(n_i-1)t}{2}+\beta_{n_i}\frac{(n_i-1)^2t^2}{4}\right]+kI_n \\
&-\gamma_{n-k+1}-\alpha_{n-k+1}\frac{(n-k)t}{2}-\beta_{n-k+1}\frac{(n-k)^2t^2}{4}+I_n \\
=&\left[\sum_{i=1}^{k}\gamma_{n_i}-\gamma_{n-k+1}\right]+\left[\sum_{i=1}^{k}\frac{\alpha_{n_i}(n_i-1)}{2}-\frac{\alpha_{n-k+1}(n-k)}{2}\right]t \\
&+\left[\sum_{i=1}^{k}\frac{\beta_{n_i}(n_i-1)^2}{4}-\frac{\beta_{n-k+1}(n-k)}{4}\right]t^2+o(1) \\
=&F_1+F_2t+F_3t^2+o(1),
\end{aligned} \tag{3.13}$$

as $n \to \infty$. Then we compute F_1, F_2 and F_3 in the last line of (3.13) respectively. Note that

$$\begin{aligned}\frac{\gamma_{n-k+1}}{p} =& -\frac{n-k+1}{2}\log\left(\frac{n-k+1}{2}\right)\\ &+\left(\frac{n-k+1}{2}+\frac{(n-k)t}{2}\right)\log\left(\frac{n-k+1}{2}+\frac{(n-k)t}{2}\right)\\ =&\frac{n-k+1}{2}\log(1+t)+\frac{(n-k)t}{2}\log\frac{n-k+1}{2}+\frac{(n-k)t}{2}\log(1+t)\\ &+\left(\frac{n-k+1}{2}+\frac{(n-k)t}{2}\right)\log\left(1-\frac{t}{(n-k+1)(1+t)}\right),\end{aligned} \tag{3.14}$$

where we used the fact

$$\begin{aligned}&\log\left(\frac{n-k+1}{2}+\frac{(n-k)t}{2}\right)\\ =&\log[(n-k+1)(1+t)-t]-\log 2\\ =&\log\left(\frac{n-k+1}{2}\right)+\log(1+t)+\log\left(1-\frac{t}{(n-k+1)(1+t)}\right).\end{aligned} \tag{3.15}$$

To get γ_{n_i}/p, we replace $n-k+1$ with n_i in (3.14), then

$$\begin{aligned}\frac{\gamma_{n_i}}{p} =& -\frac{n_i}{2}\log\left(\frac{n_i}{2}\right)+\left(\frac{n_i}{2}+\frac{(n_i-1)t}{2}\right)\log\left(\frac{n_i}{2}+\frac{(n_i-1)t}{2}\right)\\ =&\frac{n_i}{2}\log(1+t)+\frac{(n_i-1)t}{2}\log\frac{n_i}{2}+\frac{(n_i-1)t}{2}\log(1+t)\\ &+\left(\frac{n_i}{2}+\frac{(n_i-1)t}{2}\right)\log\left(1-\frac{t}{n_i(1+t)}\right).\end{aligned} \tag{3.16}$$

Subtracting (3.16) from (3.14) gives

$$\begin{aligned}\frac{F_1}{p} =&\frac{\sum_{i=1}^k \gamma_{n_i}-\gamma_{n-k+1}}{p}\\ =&\frac{k-1}{2}\log(1+t)-\frac{(n-k)t}{2}\log\frac{n-k+1}{2}+\sum_{i=1}^k\frac{(n_i-1)t}{2}\log\frac{n_i}{2}\\ &-\left(\frac{n-k+1}{2}+\frac{(n-k)t}{2}\right)\log\left(1-\frac{t}{(n-k+1)(1+t)}\right)\\ &+\sum_{i=1}^k\left(\frac{n_i}{2}+\frac{(n_i-1)t}{2}\right)\log\left(1-\frac{t}{n_i(1+t)}\right)\\ =&\frac{k-1}{2}\log(1+t)-\frac{(n-k)t}{2}\log\frac{n-k}{2}+\sum_{i=1}^k\frac{(n_i-1)t}{2}\log\frac{n_i-1}{2}\\ &+\frac{k-1}{2}t-\frac{k-1}{2}t+o(1)\\ =&\frac{(k-1)}{2}t-\frac{(n-k)t}{2}\log\frac{n-k}{2}+\sum_{i=1}^k\frac{(n_i-1)t}{2}\log\frac{n_i-1}{2}+o(t^2),\end{aligned} \tag{3.17}$$

as $n \to \infty$. Observe that $o(t^2) \cdot p = o(1)$, therefore

$$F_1 = \frac{(k-1)}{2}pt - \frac{(n-k)pt}{2}\log\frac{n-k}{2} + \sum_{i=1}^{k}\frac{(n_i-1)pt}{2}\log\frac{n_i-1}{2} + o(1). \tag{3.18}$$

By the fact $\sum_{i=1}^{k} n_i = n \Rightarrow \sum_{i=1}^{k}(n_i - 1) = n - k$, the coefficient of t, F_2, in (3.13) is given by

$$\begin{aligned} F_2 =& \sum_{i=1}^{k}\frac{\alpha_{n_i}(n_i-1)}{2} - \frac{\alpha_{n-k+1}(n-k)}{2} \\ =& \frac{1}{2}\sum_{i=1}^{k}\left[\left(p - n_i + \frac{3}{2}\right)\log\left(1-\frac{p}{n_i}\right)(n_i-1)\right] \\ & - \frac{1}{2}(n-k)\left(p - n + k + \frac{1}{2}\right)\log\left(1 - \frac{p}{n-k+1}\right) \\ =& \mu_n - \frac{(k-1)}{2}p, \end{aligned} \tag{3.19}$$

where μ_n is defined in (1.19). Similarly, we obtain

$$\begin{aligned} 4F_3 =& \sum_{i=1}^{k}\beta_{n_i}(n_i-1)^2 - \beta_{n-k+1}(n-k)^2 \\ =& -np + 2kp - \sum_{i=1}^{k}\frac{p}{n_i} - \sum_{i=1}^{k}(n_i-1)^2\log\left(1-\frac{p}{n_i}\right) \\ & + p(n-k+1) - 2p + \frac{p}{n-k+1} + (n-k)^2\log\left(1-\frac{p}{n-k+1}\right) \\ =& p(k-1) + (n-k)^2\log\left(1-\frac{p}{n-k+1}\right) + \frac{p}{n-k+1} \\ & - \sum_{i=1}^{k}\frac{p}{n_i} - \sum_{i=1}^{k}(n_i-1)^2\log\left(1-\frac{p}{n_i}\right) \\ =& 2\sigma_n^2 + \frac{p}{n-k+1} - \sum_{i=1}^{k}\frac{p}{n_i}, \end{aligned} \tag{3.20}$$

where σ_n^2 is defined in (1.20). Thus, we have

$$F_3 = \frac{\sigma_n^2}{2} + \frac{p}{4(n-k+1)} - \sum_{i=1}^{k}\frac{p}{4n_i}. \tag{3.21}$$

Plugging F_1 (3.18), F_2 (3.19), and F_3 (3.21) back in (3.13) yields

$$\begin{aligned}\log E(W_n^t) &= F_1 + F_2 t + F_3 t^3 + o(1) \\ &= \left(\mu_n - \frac{(k-1)}{2}p\right) t + \left(\frac{\sigma_n^2}{2} + \frac{p}{4(n-k+1)} - \sum_{i=1}^{k} \frac{p}{4n_i}\right) t^2 + \frac{(k-1)}{2}pt \\ &\quad - \frac{(n-k)pt}{2}\log\frac{n-k}{2} + \sum_{i=1}^{k}\frac{(n_i-1)pt}{2}\log\frac{n_i-1}{2} + o(1).\end{aligned} \tag{3.22}$$

By the given condition $k = o(n_i)$, we have $\sum_{i=1}^{k} 1/(kn_i) = o(1)$, thus

$$\sum_{i=1}^{k}\frac{p}{n_i}t^2 = O\left(\frac{1}{p}\right)\cdot O\left(\sum_{i=1}^{k}\frac{1}{kn_i}\right) = o(1) \tag{3.23}$$

and

$$\frac{p}{4(n-k+1)}t^2 = O\left(\frac{1}{p}\right)\cdot O\left(\frac{1}{n}\right) = o(1), \tag{3.24}$$

as $p \to \infty$, $k \to \infty$. Therefore (3.22) becomes

$$\begin{aligned}&\log E(W_n^t) \\ =&\left[\mu_n - \frac{(n-k)t}{2}\log\frac{n-k}{2} + \sum_{i=1}^{k}\frac{(n_i-1)t}{2}\log\frac{n_i-1}{2}\right] t + \frac{1}{2}\sigma_n^2 t^2 + o(1) \\ =&\mu_n' t + \frac{1}{2}s^2 + o(1),\end{aligned} \tag{3.25}$$

as $p \to \infty$, $k \to \infty$, where μ_n' is defined in (3.4). This completes the proof. ■

4 Numerical simulations

In this section, we present numerical simulation with 10,000 iterations to estimate the size and power of the likelihood ratio test using the central limit theorem approximation and the classical χ^2 approximation (1.11). First, we show the histogram of two tests comparing them with normal curves and chi-square curves, which gives a intuitive distinction between these two approximations. The R code for this section is given in Appendix.

In Figure 1, we display the normal curves and the histogram of $(\log\Lambda_n^* - \mu_n)/\sigma_n$. Let $n_i = 100$, $1 \le i \le 10$ with $p = 5, 20, 40, 60$. We see that the histogram of $(\log\Lambda_n^* - \mu_n)/\sigma_n$ fits better with the normal curve along with growth of dimension p. In Figure 2 we present the chi-square curves and the histogram of $-2\rho\log\Lambda_n^*$ under the same conditions as in Figure 1. Observe that the histogram of $-2\rho\log\Lambda_n^*$ parts from the chi-square curves quickly. The

comparison between the CLT and the classical chi-square approximation shows that the CLT is better when p and k are large.

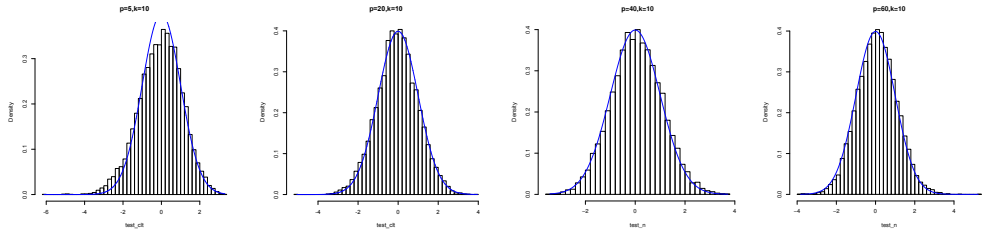

Figure 1: Histogram of $(\log \Lambda_n^* - \mu_n)/\sigma_n$ and normal curves, where $n_i = 100$, $1 \leq i \leq 10$ and $p = 5, 20, 40, 60$.

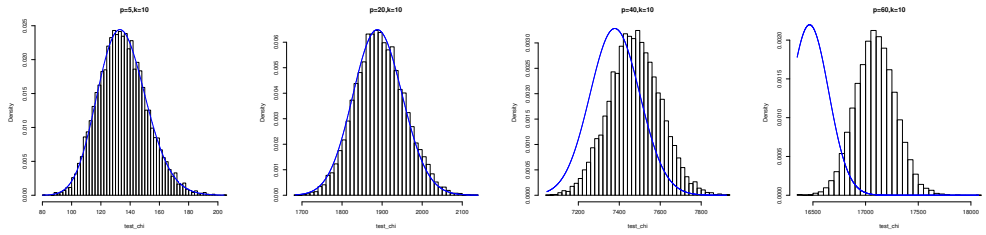

Figure 2: Histogram of $-2\rho \log \Lambda_n^*$ and χ^2 curves, where $n_i = 100$, $1 \leq i \leq 10$ and $p = 5, 20, 40, 60$.

4.1 Estimation of size

Size is the probability of rejecting H_0 when H_0 is true. In our simulation, we choose the significance level α to be 0.05. The estimation of size is the ratio of the number of tests rejected and the number of iterations used under the condition that H_0 is true. If the test is reasonable, the estimate of size should be close to the significant level $\alpha = 0.05$.

In Table 1, we list the sizes of the traditional chi-square approximation and our normal approximation. When $p = 5, 20$, the sizes of chi-square approximation matches the significant level $\alpha = 0.05$ very satisfactorily. However, when p is getting close to n, the size of the chi-square approximation increases to one rapidly, which means the traditional chi-square test almost rejects every true H_0 when $p = 60, 80, 98$ respectively. On the other hand, the normal approximation test is always very close to 0.05.

Table 1: The size of LRT for equality of k covariance matrices

Conditions	CLT	χ^2 approx.
$n_i = 100,\ k = 10,\ p = 5$	0.0773	0.0536
$n_i = 100,\ k = 10,\ p = 20$	0.063	0.0588
$n_i = 100,\ k = 10,\ p = 40$	0.0511	0.1939
$n_i = 100,\ k = 10,\ p = 60$	0.0521	0.9477
$n_i = 100,\ k = 10,\ p = 80$	0.057	1
$n_i = 100,\ k = 10,\ p = 98$	0.0483	1

The sizes are estimated based on 10,000 simulations from $N_p(0, I_p)$.

4.2 Estimation of power

Power represents the probability of rejecting H_0 when H_0 is false, which implies that the test is reasonable when the power is large. Since the estimation of power may vary subject to the choice of the alternative hypothesis, the power of a test is relatively small when the covariance matrices are closed to each other and is larger when they are significantly different. In our numerical simulation, we studied the likelihood-ratio test under two different alternative hypotheses, the first one $H_a : \Sigma_1 = I_p, \Sigma_2 = 1.2I_p, \Sigma_3 = 0.8I_p$ and the second one $H_a : \Sigma_1 = \Sigma_2 = 1.1I_p, \Sigma_3 = 0.8I_p$. The results are listed in Table 2 and Table 3 respectively.

Table 2: The power of LRT for equality of k covariance matrices

Conditions	CLT	χ^2 approx.
$n_i = 100,\ k = 3,\ p = 5$	0.7431	0.6685
$n_i = 100,\ k = 3,\ p = 20$	0.7495	0.7251
$n_i = 100,\ k = 3,\ p = 40$	0.6415	0.7246
$n_i = 100,\ k = 3,\ p = 60$	0.4972	0.9129
$n_i = 100,\ k = 3,\ p = 80$	0.339	1

The powers are estimated based on 10,000 simulations from the alternative hypothesis that $\Sigma_1 = I_p$, $\Sigma_2 = 1.2I_p$, $\Sigma_3 = 0.8I_p$.

Table 3: The power of LRT for equality of k covariance matrices

Conditions	CLT	χ^2 approx.
$n_i = 100,\ k = 3,\ p = 5$	0.633	0.5386
$n_i = 100,\ k = 3,\ p = 20$	0.6079	0.5749
$n_i = 100,\ k = 3,\ p = 40$	0.4971	0.5902
$n_i = 100,\ k = 3,\ p = 60$	0.3927	0.8611
$n_i = 100,\ k = 3,\ p = 80$	0.2739	1

The powers are estimated based on 10,000 simulations from the alternative hypothesis that $\Sigma_1 = \Sigma_2 = 1.1I_p$, $\Sigma_3 = 0.8I_p$.

References

[1] L. V. Ahlfors. *Complex analysis: An introduction of the theory of analytic functions of one complex variable*. Second edition. McGraw-Hill, New York, 1966.

[2] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Publications in Statistics. John Wiley & Sons, New York, 1958.

[3] Z. Bai, D. Jiang, J.-F. Yao, and S. Zheng. Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, 37(6B):3822–3840, 2009.

[4] M. S. Bartlett. Properties of sufficiency and statistical tests. *Proc. R. Soc. A*, 160(901):268–282, 1937.

[5] G. E. P. Box. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36:317–346, 1949.

[6] G. W. Brown. On the power of the l_1 test for equality of several variances. *Ann. Math. Statistics*, 10(2):119–128, 1939.

[7] M. L. Eaton. *Multivariate statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, New York, 1983. A vector space approach.

[8] T. W. Gamelin. *Complex analysis*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2001.

[9] D. Jiang, T. Jiang, and F. Yang. Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *J. Statist. Plann. Inference*, 142(8):2241–2256, 2012.

[10] T. Jiang and Y. Qi. Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions II. Preprint, 2013.

[11] T. Jiang and F. Yang. Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Statist.*, 41(4):2029–2074, 2013.

[12] O. Ledoit and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.*, 30(4):1081–1102, 2002.

[13] R. J. Muirhead. *Aspects of multivariate statistical theory.* John Wiley & Sons, New York, 1982. Wiley Series in Probability and Mathematical Statistics.

[14] M. D. Perlman. Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *Ann. Statist.*, 8(2):247–263, 1980.

[15] N. Sugiura and H. Nagao. Unbiasedness of some test criteria for the equality of one or two covariance matrices. *Ann. Math. Statist*, 39:1686–1692, 1968.

[16] M. B. Villarino. Ramanujan's harmonic number expansion into negative powers of a triangular number. *JIPAM. J. Inequal. Pure Appl. Math.*, 9(3):Article 89, 12, 2008.

[17] S. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4):471–494, 1932.

Appendix

We give the R commands for section 4.1 and 4.2 in this section.

First, we set up the basic conditions. For example, let $p = 40, k = 3, n_i = 100$ and iterate 10000 times.

```
> library(MASS);
> nsim=10000;
> p=40;
> k=3;
> n=rep(100,k);
> N=sum(n);
```

Generate some empty vectors in R to be used in the future:

```
> X=vector("list",k);
> a=vector("list",k);
> logdet=mat.or.vec(1,k);
> logbar=mat.or.vec(1,nsim)
```

The definition of μ_n (1.19), σ_n (1.20), ρ (1.12) and the degree of freedom f (1.12):

```
> df=p*(p+1)*(k-1)/2;
> rho=1-(sum((N-k)/(n-1))-1)*(2*p^2+3*p-1)/(6*(p+1)*(k-1)*(N-k));
> mu_n=((sum((n-1)*(2*p-2*n+3)*log(1-p/n))-(N-k)*(2*p-2*N+2*k+1)
*log(1-p/(N-k+1))))/4+(k-1)*p/2;
> sigma_sq=0.5*((N-k)^2*log(1-p/(N-k+1))-sum((n-1)^2*log(1-p/n)))+(k-1)*p/2;
> sigma=sqrt(sigma_sq);
```

To reproduce the estimation of sizes of normal approximation and chi-square approximation:

```
> for(j in 1:nsim) {
+   for(i in 1:k) X[[i]]=mvrnorm(n[i], rep(0,p), diag(p));
+   for(i in 1:k) a[[i]]=(n[i]-1)*cov(X[[i]]);
+   for(i in 1:k) logdet[i]=log(det(a[[i]]));
+   A=Reduce('+',a);
+   logbar[j]=sum((n-1)*logdet/2)+(p/2)*(N-k)*log(N-k)-(N-k)*log(det(A))/2
-sum((n-1)*p*log(n-1)/2);
+ }
>
> test_clt=(logbar-mu_n)/sigma;
> test_chi=-2*rho*logbar;
>
> size_clt=length(which(test_clt<qnorm(0.05,0,1)))/nsim;
> size_chi=length(which(test_chi>qchisq(0.95,df)))/nsim;
```

To reproduce the estimation of powers of normal approximation and chi-square approximation:

```
> for(j in 1:nsim) {
+   X[[1]]=mvrnorm(n[1], rep(0,p), diag(p));
+   X[[2]]=mvrnorm(n[2], rep(0,p), 1.2*diag(p));
+   X[[3]]=mvrnorm(n[3], rep(0,p), 0.8*diag(p));
+   for(i in 1:k) a[[i]]=(n[i]-1)*cov(X[[i]]);
+   for(i in 1:k) logdet[i]=log(det(a[[i]]));
+   A=Reduce('+',a);
+   logbar[j]=sum((n-1)*logdet/2)+(p/2)*(N-k)*log(N-k)-(N-k)*log(det(A))/2
-sum((n-1)*p*log(n-1)/2);
```

```
+ }
>
> test_clt=(logbar-mu_n)/sigma;
> test_chi=-2*rho*logbar;
>
> power_clt=length(which(test_clt<qnorm(0.05,0,1)))/nsim;
> power_chi=length(which(test_chi>qchisq(0.95,df)))/nsim;
```

Figure 1 and Figure 2 can be reproduced in R as follows:

```
> h_n=hist(test_clt,50,main='p=40,k=10',freq=FALSE);
> seq_n=seq(min(test_clt),max(test_clt),by=0.001);
> density_n=dnorm(seq_n,0,1);
> lines(seq_n,density_n,col="blue",lwd=2);
>
> h_chi=hist(test_chi,50,main='p=5,k=10',freq=FALSE);
> seq_chi=seq(min(test_chi),max(test_chi),by=0.001);
> density_chi=dchisq(seq_chi,df);
> lines(seq_chi,density_chi,col="blue",lwd=2);
```