DIRECTED STUDY IN STATISTICS - FINAL REPORT

GAN DAILIN
Supervisor: Prof. Jeff J.F. Yao

December, 2019

Abstract

Multivariate statistics analysis flourishes in recent decades due to some problems arising in genetics and social science, which require a large number of data [1]. It would be natural to consider the distribution of large sample data and explore some properties of it. Wishart distribution is of great importance in estimating the covarince matrices in multivariate statistics. With this distribution, hypothesis testing for multivariate samples could be conducted and the close form of the likelihood ratio (LR) criterion could be found [2]. Besides, the joint density of the eigenvalues was found based on Wishart distribution, which indicated a new era for multivariate distribution theory. Another great result would be Marcenko-Pastur Law or The Quarter Circle Law, which gives the limiting distribution of eigenvalues when the ratio between the number of variables and sample size converges to some constant [1].

Apart from the derivation of Wishart distribution and the joint density of the eigenvalues, hypothesis testing for multivariate samples may be another interesting topic. When the ratio between the number of variables and sample size is quite small, the classic LR test may give a good result for the hypothesis testing result given simulated data [2]. However, when the ratio becomes larger and larger but is still smaller than 1, LR test may fail in this situation but some alternative methods, such as Nagao's criterion [3] and Bai & Yao's criterion [4], may perform better than the classic one.

This report will firstly show the relationship between Wishart distribution and the maximum likelihood estimator of covariance matrices under normal setting. Then, classic methods of deriving the probability density function of Wishart distribution and the joint density of the eigenvalues will be reviewed. An alternative method to derive the eigenvalue distribution will also be discussed. Besides that, three hypothesis testing methods, which are LR criterion, Nagao's criterion and Bai & Yao's criterion, will be compared. The results show that when the number of variables is much smaller than the sample size, LR test could give a good result. While the number of variables becomes larger, LR test may not be satisfactory but Nagao's and Bai & Yao's criterion perform better.

1 Introduction

This section will introduce the basic setting and derive the maximum likelihood estimators of the mean μ and the covariance Σ for multivariate normal distribution. Hence, find out the relationship between Wishart distribution and the covariance estimator.

Let X_1, X_2, $\cdots$, X_N be N independent and identically distributed random variables with $X_i \sim \mathcal{N}_p(\mu, \Sigma)$ for each X_i. Note that the dimensions for μ and Σ are $p \times 1$ and $p \times p$ respectively. The probability density function for X_i is

$$f(X_i) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x_i-\mu)^\intercal \Sigma^{-1}(x_i-\mu)}. \tag{1}$$

The likelihood function is

$$L = \prod_{i=1}^{N} f(X_i) = (2\pi)^{-\frac{1}{2}pN} |\Sigma|^{\frac{1}{2}N} e^{-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^\intercal \Sigma^{-1}(x_i-\mu)}. \tag{2}$$

Hence, the log-likelihood function is

$$\ln L = -\frac{1}{2}pN \ln 2\pi - \frac{1}{2}N \ln |\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^\intercal \Sigma^{-1}(x_i-\mu). \tag{3}$$

Define the sample mean vector as $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. The matrix of sum of squares and cross product of deviations about the mean is $A = \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^\intercal$

Lemma 1. Let X_1, X_2, $\cdots$, X_N be N vectors with dimension p. $\bar{X}$ is defined above. For any vector b with dimension p, we have that

$$\sum_{i=1}^{N}(x_i - b)(x_i - b)^\intercal = \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^\intercal + N(\bar{x} - b)(\bar{x} - b)^\intercal. \tag{4}$$

Proof. $\sum_{i=1}^{N}(x_i-b)(x_i-b)^\intercal = \sum_{i=1}^{N}(x_i-\bar{X}+\bar{X}-b)(x_i-\bar{X}+\bar{X}-b)^\intercal = \sum_{i=1}^{N}((x_i-\bar{x})(x_i-\bar{x})^\intercal + (\bar{X}-b)(X_i-\bar{X})^\intercal+(X_i-\bar{X})(\bar{X}-b)^\intercal+(\bar{X}-b)(\bar{X}-b)^\intercal) = \sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^\intercal+N(\bar{x}-b)(\bar{x}-b)^\intercal$. □

According to Lemma 1 and the property of trace, we can rewrite the last term in (3) as

$$\begin{aligned}
&\sum_{i=1}^{N}(x_i-\mu)^\intercal \Sigma^{-1}(x_i-\mu) \\
&= tr(\sum_{i=1}^{N} \Sigma^{-1}(x_i-\mu)(x_i-\mu)^\intercal) \\
&= tr(\Sigma^{-1}(A + N(\bar{x}-\mu)(\bar{x}-\mu)^\intercal)) \\
&= tr(\Sigma^{-1}A) + N(\bar{x}-\mu)^\intercal \Sigma^{-1}(\bar{x}-\mu).
\end{aligned}$$

Therefore, the log-likelihood function can be rewritten as

$$\ln L = -\frac{1}{2}pN \ln 2\pi - \frac{1}{2}N \ln |\Sigma| - \frac{1}{2}(tr(\Sigma^{-1}A) + N(\bar{x}-\mu)^\intercal \Sigma^{-1}(\bar{x}-\mu)). \tag{5}$$

Note that Σ is positive definite, so is Σ^{-1}. Hence, $(\bar{x}-\mu)^\intercal\Sigma^{-1}(\bar{x}-\mu) \geq 0$ and the equality holds when $\bar{x}=\mu$. Then, we want to maximize $-\frac{1}{2}(N\ln|\Sigma| + tr(\Sigma^{-1}A))$.

Lemma 2. If D is positive definite matrix of order p, the maximum of $f(G) = -N\ln|G| - tr(G^{-1}D)$ with respect to positive definite matrix G exists and occurs at $G = \frac{D}{N}$, with value $f(\frac{D}{N}) = -N\ln|D| + Np\ln N - pN$.

Proof. Let $D = EE^\intercal$ and $G = EH^{-1}E^\intercal$. Note that $|G| = |E||H^{-1}||E^\intercal| = |H^{-1}||D| = |D|/|H|$ and $tr(G^{-1}D) = tr(G^{-1}EE^\intercal) = tr(E^\intercal G^{-1}E) = tr(H)$. Hence, we can obtain $f(G) = -N\ln|G| - tr(G^{-1}D) = N\ln|H| - N\ln|D| - tr(H)$. By Cholesky decomposition, let $H = TT^\intercal$, where T is a lower triangular matrix. Then, $f(G) = -N\ln|D| + N\ln TT^\intercal - tr(TT^\intercal) = -N\ln|D| + \sum_{i=1}^{p}(2N\ln t_{ii} - t_{ii}^2) - \sum_{i>j} t_{ij}^2$, which is maximized when $t_{ii}^2 = N, t_{ij} = 0$ for $i \neq j$. Note that t_{ij} is an entry from the lower triangular matrix T. Therefore, when $H = NI$ and $G = EE^\intercal/N = D/N$, $f(G)$ obtain its maximum value. □

With above lemma 2, we can get the results for the likelihood estimators immediately, which is summarized in the following theory.

Theory 3. If $x_1, x_2, \ldots, x_N$ are from a sample following $\mathcal{N}_p(\mu, \Sigma)$ with $p < N$, the maximum likelihood estimators for the mean and covariance matrix are $\hat{\mu} = \sum_{i=1}^{N} x_i/N$ and $\hat{\Sigma} = A/N$, where $A = \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^\intercal$.

Note that the maximum likelihood estimator for the covariance matrix equals A/N, where A is the matrix of which we want to obtain the distribution. In the next section, we want to find the distribution of A which is called the Wishart distribution.

2 The Wishart Distribution

This section will show the classic approach to obtain the probability density function of the Wishart distribution. We will begin the proof from the matrix A which is defined in the previous section.

In order to obtain the general formula of the distribution of $A = \sum_{i=1}^{N+1}(x_i - \bar{x})(x_i - \bar{x})^\intercal$, we may first consider the case when the mean of the sample is 0, in which $A = \sum_{i=1}^{N} z_i z_i^\intercal$ and $z_1, z_2, \ldots, z_N$ are independent with $\mathcal{N}_p(0, \Sigma)$. The density function of the positive definite matrix A is given below

$$f(A|\mu = 0, \Sigma) = \frac{|A|^{\frac{1}{2}(N-p-1)}e^{-\frac{1}{2}tr(\Sigma^{-1}A)}}{2^{\frac{1}{2}Np}\pi^{\frac{p(p-1)}{4}}|\Sigma|^{\frac{N}{2}}\prod_{i=1}^{p}\Gamma(\frac{1}{2}(N+1-i))}. \tag{6}$$

Case 1: $\Sigma = I$. We first consider the case when $\Sigma = I$. Define $\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{bmatrix} := \begin{bmatrix} z_1 & z_2 & \cdots & z_N \end{bmatrix}$. Note that the dimensions for v_i and z_i are $1 \times N$ and $p \times 1$ respectively. In the following discussion, we always treat v_i as a column vector for simplification of illustration. Hence, each entry of the matrix A could be represented as $a_{ij} = v_i^\intercal v_j$, where each v_i is independently distributed with $\mathcal{N}_N(0, I_N)$.

Then, we want to construct a new orthogonal coordinate based on v_i, such that the matrix A could be represented by a lower triangular matrix T by Cholesky decomposition. For each element t_{ij} in T, it can be written in terms of the new orthogonal coordinate. Details of the construction are provided in the following paragraphs.

Firstly, we want to construct the new orthogonal coordinate system. By Gram-Schmidt orthogonalization, define $w_i := v_i - \sum_{j=1}^{i-1} proj_{\mathbf{w_j}}\mathbf{v_i}$, where $w_i^\intercal w_j = 0$ for each $i \neq j$. Define $t_{ii} := ||w_i|| = (w_i^\intercal w_i)^{\frac{1}{2}}$ for $i = 1, 2, \cdots, p$ and $t_{ij} := v_i^\intercal \frac{w_j}{||w_j||}$ for $j = 1, 2, \cdots, i-1$. Note that $v_i = c_1 w_1 + c_2 w_2 + \cdots + c_{i-1} w_{i-1} + w_i$ for some constant c_j. The formula $t_{ij} = v_i^\intercal \frac{w_j}{||w_j||} = c_j \frac{w_j^\intercal w_j}{||w_j||} = c_j ||w_j||$ will give the $j-th$ coefficient of v_i in the w_j coordinate. Therefore, we have $v_i = \sum_{j=1}^{i} t_{ij} \frac{w_j}{||w_j||}$ and hence, $a_{hi} = v_h^\intercal v_i = \sum_{j=1}^{min(h,i)} t_{hj} t_{ij}$. By Cholesky decomposition, define th lower triangular matrix $T := (t_{ij})$ with $t_{ii} > 0$ for $i = 1, 2, \cdots, p$ and $t_{ij} = 0$ for $i < j$. Then, $A = TT^\intercal$.

According to previous definition of t_{ii} and t_{ij}, we may notice that t_{ii} includes the multiplication of two vectors which contains normal random vectors and t_{ij} represents an entry from a normal random vector. Then, the following lemma could be proved without too much difficulty.

Lemma 4. Conditional on $w_1, w_2, \cdots, w_{i-1}$, $t_{i,1}, t_{i,2}, \cdots, t_{i,i-1}$ and t_{ii}^2 are independently distributed, where $t_{ij} \sim \mathcal{N}(0,\, 1)$ for $i > j$ and $t_{ii}^2 \sim \chi^2_{N-(i-1)}$.

The degree of freedom of the Chi-square distribution is $N - (i-1)$ because the first $i-1$ terms are given in the condition and hence, only $N - (i-1)$ terms are left. With the results from lemma 4, we can obtain the following corollary.

Corollary 5. Let $z_1, z_2, \cdots, z_N$ be independently distributed vectors following $\mathcal{N}_p(0,\, I_p)$ for $N \geq p$. Let $A = \sum_{i=1}^{N} z_i z_i^\intercal = TT^\intercal$, where T is a lower triangular matrix and each entry is independent. Then, $t_{ij} \sim \mathcal{N}(0,\, 1)$ for $i > j$ and $t_{ii}^2 \sim \chi^2_{N-(i-1)}$.

Next, we want to find the joint density function of t_{ij} for $j = 1, 2, \cdots, i$ and $i = 1, 2, \cdots, p$. This density function is the probability density function of the matrix A when the covariance $\Sigma = I$. Since the distribution for t_{ij} and t_{ii}^2 are given in the previous corollary, we only need to multiply the Jacobian term to obtain the probability density function of t_{ii}, which is

$$\begin{aligned} f_{t_{ii}}(t_{ii}) &= f_{t_{ii}^2}(t_{ii}) \left| \frac{\partial t_{ii}^2}{\partial t_{ii}} \right| \\ &= \frac{2^{-\frac{1}{2}(N-i-1)} t_{ii}^{N-1} e^{-\frac{t_{ii}^2}{2}}}{\Gamma(\frac{1}{2}(N-i+1))}. \end{aligned}$$

Then, the joint density function is

$$\begin{aligned} f(t_{11}, t_{12}, \cdots, t_{pp}) &= \prod_{i=1}^{p} \left(\frac{2^{-\frac{1}{2}(N-i-1)} t_{ii}^{N-1} e^{-\frac{t_{ii}^2}{2}}}{\Gamma(\frac{1}{2}(N-i+1))} \left(\prod_{j=1}^{i-1} (2\pi)^{-\frac{1}{2}} e^{-\frac{t_{ij}^2}{2}}\right)\right) \\ &= \prod_{i=1}^{p} \left(\frac{t_{ii}^{N-1} e^{-\frac{1}{2}\sum_{j=1}^{i} t_{ij}^2}}{2^{\frac{N}{2}-1} \pi^{\frac{i-1}{2}} \Gamma(\frac{1}{2}(N+1-i))}\right) \\ &= \frac{\prod_{i=1}^{p} t_{ii}^{N-i} e^{-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{i} t_{ij}^2}}{2^{\frac{1}{2}p(N-2)} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma(\frac{1}{2}(N+1-i))}. \end{aligned} \tag{7}$$

The joint density function in (7) is the required density function of A when the covariance matrix $\Sigma = I$. With the help of previous derivation, when the covariance is Σ, the joint density function can be found by using linear transformation and Jacobian terms.

Case 2: $\Sigma \neq I$**.** Since Σ is positive definite, by Cholesky decomposition, $\Sigma = CC^\intercal$ for some lower triangular matrix C. Using linear transformation, define $T_* := CT$, where T is the defined lower triangular matrix in the previous section. Hence, $t_{*ij} = \sum_{k=j}^{i} c_{ik} t_{kj}$ for $i \geq j$. If we consider the vector $\begin{bmatrix} t_{*11} & t_{*21} & t_{*22} & t_{*31} & \cdots & t_{*pp} \end{bmatrix}^\intercal$, it can be viewed as the dot product of a lower triangular matrix whose diagonal entries are $\begin{bmatrix} c_{11} & c_{22} & c_{22} & c_{33} & \cdots & c_{pp} \end{bmatrix}$ and a vector $\begin{bmatrix} t_{11} & t_{21} & t_{22} & t_{31} & \cdots & t_{pp} \end{bmatrix}^\intercal$. Therefore, the Jacobian term of the transformation from T to T_* is $(\prod_{i=1}^{p} c_{ii}^i)^{-1}$, which is the inverse of the multiplication of the diagonal entries.

Note that $t_{ii} = t_{*ii}/c_{ii}$, $\prod_{i=1}^{p} c_{ii}^2 = |C||C| = |\Sigma|$ and

$$\begin{aligned} \sum_{i=1}^{p}\sum_{j=1}^{i} t_{ij}^2 &= tr(TT^\intercal) \\ &= tr(T_* T_*^\intercal (C^\intercal C)^{-1}) \\ &= tr(T_*^\intercal \Sigma^{-1} T_*) \end{aligned}$$

Therefore, we can get the probability density function of $T_* T_*^\intercal$ based on above transformation, which is

$$\begin{aligned} f(t_{*11}, t_{*12}, \cdots, t_{*pp}) &= \frac{\prod_{i=1}^{p} (\frac{t_{*ii}}{c_{ii}})^{N-i} e^{-\frac{1}{2}tr(\Sigma^{-1} T_* T_*^\intercal)}}{2^{\frac{1}{2}p(N-2)} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma(\frac{1}{2}(N+1-i))} \left(\prod_{i=1}^{p} c_{ii}^i\right)^{-1} \\ &= \frac{\prod_{i=1}^{p} (t_{*ii})^{N-i} e^{-\frac{1}{2}tr(\Sigma^{-1} T_* T_*^\intercal)}}{2^{\frac{1}{2}p(N-2)} \pi^{\frac{p(p-1)}{4}} |\Sigma|^{\frac{N}{2}} \prod_{i=1}^{p} \Gamma(\frac{1}{2}(N+1-i))} \end{aligned} \tag{8}$$

It is natural to consider the relationship between $A = \sum_{i=1}^{N+1} (x_i - \bar{x})(x_i - \bar{x})^\intercal$ and $T_* T_*^\intercal$. We may notice that $a_{hi} = \sum_{j=1}^{i} t_{*hj} t_{*ij}$ for $h > i$. The Jacobian term of the transformation from A to T_* can be similarly found according to the previous section's method, which is also the determinant of a lower triangular matrix by multiplying the diagonal entries. Hence, the Jacobian term from T_* to A is $(2^p \prod_{i=1}^{p} t_{*ii}^{p+1-i})^{-1}$.

With above derivation, the probability density function of the Washart distribution is summarized in the following theory.

Theory 6. Let $z_1, z_2, \cdots, z_N$ be independently distributed vectors with $\mathcal{N}_p(0, \Sigma)$. The density of $A = \sum_{i=1}^{N} z_i z_i^\intercal$ is the equation (6).

Proof. By replacing T_* with A and multiplying the Jacobian term, the density function can be written as

$$f(A|\mu = 0, \Sigma) = \frac{(\prod_{i=1}^{p}(t_{*ii})^2)^{\frac{1}{2}(N-p-1)} e^{-\frac{1}{2}tr(\Sigma^{-1}A)}}{2^{\frac{1}{2}pN}\pi^{\frac{p(p-1)}{4}}|\Sigma|^{\frac{N}{2}}\prod_{i=1}^{p}\Gamma(\frac{1}{2}(N+1-i))}$$

Since $|T_*||T_*| = \prod_{i=1}^{p} t_{ii}^2 = |A|$, the density function can be further simplified to the equation (6). □

In the end, the distribution of A is called the Wishart distribution, which can be denoted by $W(\Sigma, N)$ for $N = (N+1) - 1$.

3 Distribution of Eigenvalues and Marcenko-Pastur Law

This section will review the the classic proof of the joint density of eigenvalues under double Wishart setting and an alternative proof of the eigenvalue distribution provided by Dumitriu and Edelman[3]. After that, the result of Marcenko-Pastur Law will also be given.

3.1 Joint density of eigenvalues under double Wishart setting

At the beginning, we give out the basic settings and some definitions. Let A^* and B^* be independent $p \times p$ random matrices, such that $A^* \sim W(\Sigma, m)$ and $B^* \sim W(\Sigma, n)$. If l satisfies that $|A^* - lB^*| = 0$, then l is called a characteristic root of A^* in the metric of B^*. If a vector x satisfies $(A^* - lB^*)x = 0$, then x is called a characteristic vector of A^* in the metric of B^*. The classic method is provided in the following.

We may first want to reduce Σ into the identity matrix I. which would be easier to deal with. Let C be matrix, such that $C\Sigma C^\intercal = I$. Let $A = CA^*C^\intercal$ and $B = CB^*C^\intercal$. Then, we can get that $A \sim W(I, m)$ and $B \sim W(I, n)$. The following results can be derived with no surprise.

Lemma 7. The characteristic roots for A^* and B^* are the same with those for A and B. The characteristic vectors only differ with a matrix C

Proof. For the characteristic roots, it can be noted that $|A - lB| = 0 \Leftrightarrow |C(A^* - lB^*)C^\intercal| = 0 \Leftrightarrow |C||(A^* - lB^*)||C^\intercal| = 0 \Leftrightarrow |A^* - lB^*| = 0$.

For the characteristic vectors, $(A - lB)x = 0 \Leftrightarrow C^{-1}(A - lB)x = 0 \Leftrightarrow C^{-1}C((A^* - lB^*))C^\intercal x = 0 \Leftrightarrow (A^* - lB^*)C^\intercal x = 0$. Then, we can get $x^* = C^\intercal x$. □

The above lemma 7 ensures that we only need to consider A and B in the following proof. Let $l = \frac{f}{1-f}$ for $f \neq 1$. The the roots and characteristic vectors can be rewritten as

$$|A - f(A + B)| = 0 \tag{9}$$

$$(A - f(A+B))y = 0 \tag{10}$$

We may first consider the distribution of the roots and vectors in (9) and (10) and then the target distribution will differ with a Jacobian term.

Let roots f_i be ordered, such that $f_1 > f_2 > \cdots > f_p$ for $f_i \neq f_j$ if $i \neq j$. Suppose the corresponding vectors y_i are normalised by

$$y_i^\intercal (A+B) y_i = 1 \tag{11}$$

for $i = 1, 2, \cdots, p$. These vectors also satisfy that

$$y_i^\intercal (A+B) y_j = 0. \tag{12}$$

Let F be a $p \times p$ diagonal matrix whose diagonal entries are $\begin{bmatrix} f_1 & f_2 & \cdots & f_p \end{bmatrix}^\intercal$. Let $Y := \begin{bmatrix} y_1 & y_2 & \cdots & y_p \end{bmatrix}$, where y_i are column vectors. Hence, the equation (10) can be rewritten as

$$AY = (A+B)YF. \tag{13}$$

According to (11) and (12), we can get

$$Y^\intercal (A+B) Y = I. \tag{14}$$

From (13) and (14), let $Y^{-1} = E$ and we can get the solution of A and B with regard to F and E

$$\begin{cases} A = E^\intercal F E \\ B = E^\intercal (I - F) E \end{cases} \tag{15}$$

It should be noted that the distributions for A and B are given in the very beginning. The matrix E is related to the characteristic vectors while the matrix F contains the characteristic roots. The equation (15) separates the known and unknown parts. The following steps include deriving the joint density function of E and F, finding the marginal distribution of E and getting the distribution of characteristic roots.

The Jacobian term from A and B to E and F is

$$\left| \frac{\partial(A,B)}{\partial(E,F)} \right| = 2^p |E|^{p+2} \prod_{i<j} (f_i - f_j). \tag{16}$$

The detailed proof can be found in [2].

The joint density of A and B is

$$w(A|I,m) w(B|I,n) = C_1 |A|^{\frac{1}{2}(m-p-1)} |B|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2} tr(A+B)}, \tag{17}$$

where C_1 is the constant term.

Note that

$$\begin{aligned} |E^\intercal F E| &= |F||E^\intercal E| = (\prod_{i=1}^{p} f_i)|E^\intercal E| \\ |E^\intercal (I-F) E| &= |(I-F)||E^\intercal E| = (\prod_{i=1}^{p} (1-f_i))|E^\intercal E| \end{aligned} \tag{18}$$

With (15), (16), (17) and (18), we can get the joint distribution function of E and F, which is

$$\begin{aligned} D(E,F) &= C_1|E^\intercal FE|^{\frac{1}{2}(m-p-1)}|E^\intercal(I-F)E|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}tr(E^\intercal E)}|\frac{\partial(A,B)}{\partial(E,F)}| \\ &= 2^p C_1|E^\intercal E|^{\frac{1}{2}(m+n-p)}e^{-\frac{1}{2}tr(E^\intercal E)}\prod_{i=1}^{p} f_i^{\frac{1}{2}(m-p-1)}(1-f_i)^{\frac{1}{2}(n-p-1)}\prod_{i<j}(f_i-f_j). \end{aligned} \tag{19}$$

Integrate out F and change f_i with l_i, the required function for the joint distribution of eigenvalues is

$$C_2\prod_{i=1}^{p}(l_1^{\frac{1}{2}(m-p-1)}(l_i+1)^{-\frac{1}{2}(m+n)})\prod_{i<j}(l_i-l_j), \tag{20}$$

where $C_2 = \frac{\pi^{\frac{1}{2}p^2}\Gamma_p(\frac{1}{2}(m+n))}{\Gamma_p(\frac{1}{2}n)\Gamma_p(\frac{1}{2})\Gamma_p(\frac{1}{2}p)}$.

3.2 Tridiagonalizing method

This section will review an alternative method to derive the joint density function of eigenvalues, which was found by Dumitriu and Edelman [5]. The tridiagonalizing method provides an inductive way to find the distribution under single Wishart setting.

The target eigenvalue density function is

$$f_\beta(l) = C(\beta,\alpha)\prod_{i<j}|l_i-l_j|^\beta\prod_{i=1}^{p} l_i^{\alpha-k}e^{-\sum_{i=1}^{p} l_i/2}, \tag{21}$$

where $\beta = 1$ for real numbers, $\alpha = \frac{\beta}{2}n$, $k = 1+\frac{\beta}{2}(p-1)$ and

$$C(\beta,\alpha) = 2^{-pa}\prod_{j=1}^{p}\frac{\Gamma(1+\beta/2)}{\Gamma(1+\frac{\beta}{2}j)\Gamma(a-\frac{\beta}{2}(m-j))}$$

It can be noticed that $l_i^{\alpha-k}e^{-\sum_{i=1}^{p} l_i/2}$ is the kernel of Chi-square distribution. The following theory summarizes the main result of this method.

Theory 8. Let $W = GG^\intercal$ be a $p\times p$ Wishart real matrix, where G is a $p\times q$ matrix whose entries are $i.i.d$ standard Gaussians. By reducing G into a bidiagonal matrix B, the matrix $T = BB^\intercal$ has the joint eigenvalue density function (21), which is the eigenvalue distribution of W.

The proof for theory 8 is briefly discussed in the following. Let $G = \begin{bmatrix} x^\intercal \\ G_1 \end{bmatrix}$, where x is a $q\times 1$ vector and G_1 is a $(p-1)\times q$ matrix. Let R be a right reflector matrix such that $x^\intercal R = ||x||e_1$. Define $G_1R := [y, G_2]$. Similarly, let L be a left reflector matrix such that $Ly = ||x||e_1$. Then, the first step can be summarized as

$$\begin{bmatrix} 1 & 0 \\ 0 & L \end{bmatrix} GR = \begin{bmatrix} ||x|| & 0 \\ ||y||e_1 & LG_2 \end{bmatrix}. \tag{22}$$

It can be noticed that $||x||^2 \sim \chi^2_{q-1}$ and $||y||^2 \sim \chi^2_{p-1}$, because each entry of the matrix G is $i.i.d$ standard Gaussian. Then, we can repeat the above step inductively to obtain a bidiagonal matrix B. Since only orthogonal multiplications are used in the process, the singular values of G will not change. Therefore, the eigenvalue distributions of W and T are the same.

3.3 Marcenko-Pastur Law

This section will briefly discuss the results of Marcenko-Pastur Law, which was found by Marcenko and Pastur[1]. This law gives the limiting distribution of eigenvalues when both the number of variables p and the sample size n become very large but the ratio p/n converges to some constant γ. The setting and results are in the following.

Let A be a $p \times p$ random matrix following $W(I, n)$. The empirical cumulative density function of eigenvalues is

$$F_p(t) = \frac{1}{p} \# \left[l_i \leq t \right], \tag{23}$$

where $\# \left[l_i \leq t \right]$ the number of eigenvalues which are smaller than t.

When p and n increase together but the ratio p/n converges, the empirical cumulative density function has a limiting density, which is

$$f(t) = \frac{((b_+ - t)(t - b_-))^{\frac{1}{2}}}{2\pi\gamma t}, \tag{24}$$

where $b_+ = (1 + \gamma^{\frac{1}{2}})^2$ and $b_- = (1 - \gamma^{\frac{1}{2}})^2$.

It can be noticed that when p is relatively large to n, the convergent ratio γ will increase. Hence, the limiting density will more spread out across its domain.

The next section is about the hypothesis testing for the covariance matrix under normal assumption. All the results obtained in previous sections, like the maximum likelihood estimators, the probability density function and the Marcenko-Pastur will be used to construct different methods for hypothesis testing.

4 Hypothesis Testing

Let $x_1, x_2, \cdots, x_N$ be N independent samples from $\mathcal{N}_p(\mu, \Sigma)$ with dimension p. The key problem of this section is to test whether the hypothesis that $H : \Sigma = \sigma^2 I$ is true or not with some significance level α.

4.1 Derivation of the Likelihood Ratio Criterion

Before finding the likelihood ratio (LR) criterion, it can be noted that the hypothesis $H : \Sigma = \sigma^2 I$ is a combination of two hypothesises, which are

- H_1: Σ is a diagonal matrix

- H_2: the diagonal elements of Σ are equal given Σ is diagonal.

The following lemma can help us derive the criterion based on the combination of two hypothesises. The detailed proof for this lemma is in [2].

Lemma 9. Let x be an observed vector from a distribution with the parameter vector θ and $\theta \in \Omega$. Suppose three hypothesises are

- H_a: $\theta \in \Omega_a \subset \Omega$
- H_b: $\theta \in \Omega_b \subset \Omega_a$, given $\theta \in \Omega_a$
- H_{ab}: $\theta \in \Omega_b$, given $\theta \in \Omega$.

Denote the LR criterion of H_a, H_b and H_{ab} as λ_a, λ_b and λ_{ab} respectively. Then, the LR criterion of H_{ab} can be uniquely defined as $\lambda_{ab} = \lambda_a \lambda_b$.

Criterion for H_1. Let X be a $p \times 1$ random vector from $\mathcal{N}_p(\mu, \Sigma)$. Partition X into q sub-vectors, which are $X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(q)} \end{bmatrix}$, $\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \\ \vdots \\ \mu^{(q)} \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1q} \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \Sigma_{qq} \end{bmatrix}$. To test H_1 is equivalent to test that $X^{(1)}, X^{(2)}, \cdots, X^{(q)}$ are mutually independent, which is to say $\Sigma_{ij} = 0$ for $i \neq j$. Therefore, $H_1 : \Sigma_0 = \begin{bmatrix} \Sigma_{11} & 0 & \cdots & 0 \\ \vdots & \Sigma_{22} & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma_{qq} \end{bmatrix}$.

Let L be the likelihood function which is obtained in (2). The maximum likelihood estimators for the mean and covariance matrix are $\hat{\mu} = \sum_{i=1}^{N} x_i/N$ and $\hat{\Sigma} = A/N$, where $A = \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^\intercal$. Hence, we can obtain $max_{\mu,\Sigma} L(\mu, \Sigma) = (2\pi)^{-\frac{1}{2}pN} |\hat{\Sigma}|^{-\frac{1}{2}N} e^{-\frac{1}{2}pN}$.

The next step is to find the maximum value of L under null hypothesis. The method to find the maximum value is similar to that mentioned in the Introduction section. We can first deal with the likelihood function of one partition and them consider all the partitions together. The process is summarized below

$$\begin{aligned} max_{\mu,\Sigma_0} L(\mu, \Sigma_0) &= \prod_{i=1}^{q} max_{\mu^i, \Sigma_{ii}} L_i(\mu^i, \Sigma_{ii}) \\ &= \prod_{i=1}^{q} (2\pi)^{-\frac{1}{2}p_i N} |\hat{\Sigma}_{ii}|^{-\frac{1}{2}N} e^{-\frac{1}{2}p_i N} \\ &= (2\pi)^{-\frac{1}{2}pN} (\prod_{i=1}^{q} |\hat{\Sigma}_{ii}|^{-\frac{1}{2}N}) e^{-\frac{1}{2}pN} \end{aligned}$$

Where $\hat{\Sigma}_{ii} = \frac{1}{N} \sum_{a=1}^{N} (x_a^{(i)} - \bar{x}^{(i)})(x_a^{(i)} - \bar{x}^{(i)})^\intercal = \frac{1}{N} A_{ii}$.

Therefore, the criterion for H_1 is $\lambda_1 = \frac{max_{\mu,\Sigma_0} L(\mu,\Sigma_0)}{max_{\mu,\Sigma} L(\mu,\Sigma)} = \frac{|\hat{\Sigma}|^{\frac{1}{2}N}}{\prod_{i=1}^{q} |\hat{\Sigma}_{ii}|^{\frac{1}{2}N}} = \frac{|A|^{\frac{1}{2}N}}{\prod_{i=1}^{q} |A_{ii}|^{\frac{1}{2}N}}$. If $q = p$,

$$\lambda_1 = \frac{|A|^{\frac{1}{2}N}}{\prod_{i=1}^{q} |a_{ii}|^{\frac{1}{2}N}}, \tag{25}$$

which is the required criterion for the hypothesis H_1.

Criterion for H_2. Let x_a^g be observations from the $g-th$ population with $\mathcal{N}_p(\mu^g, \Sigma_g)$ for $a = 1, 2, \cdots, N_g$ and $g = 1, 2, \cdots, q$. We want to test $H_2 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_q$, given H_g are diagonal matrices. Define $N := \sum_{g=1}^{q} N_g$, $A_g := \sum_{a=1}^{N_g} (x_a^{(g)} - \bar{x}^{(g)})(x_a^{(g)} - \bar{x}^{(g)})^\intercal$ and $A := \sum_{g=1}^{q} A_g$.

The likelihood function for all the N samples is

$$L = \prod_{g=1}^{q} \prod_{a=1}^{N_g} (2\pi)^{-\frac{1}{2}p} |\Sigma_g|^{-\frac{1}{2}} e^{-\frac{1}{2}(x_a^{(g)} - \mu^{(g)})^\intercal \Sigma_g^{-1} (x_a^{(g)} - \mu^{(g)})}.$$

Define Ω as the full parameter space for the alternative hypothesis and ω as the reduced parameter space for the null hypothesis. Similar to the process mentioned in the Introduction section, we can get the maximum likelihood estimators for the mean and covariance matrix in the whole space Ω, which are $\hat{\mu}_\Omega^{(g)} = \bar{x}^{(g)}$ and $\hat{\Sigma}_{g,\Omega} = A_g/N_g$, for each $g = 1, 2, \cdots, q$. Under the null hypothesis, we can set $\Sigma_1 = \Sigma_1 = \cdots = \Sigma_q = \Sigma$. The maximum estimator of the mean under the null hypothesis can be easily found, which is $\hat{\mu}_\omega^{(g)} = \bar{x}^{(g)}$. Then, put back the estimated mean and maximize the likelihood function with regard to Σ. The result is $\hat{\Sigma}_\omega = A/N$.

Put back all the maximum likelihood estimators in the likelihood function. Then, we can obtain the likelihood ratio, which is

$$\lambda_2 = \frac{max_\omega L}{max_\Omega L} = \frac{\prod_{g=1}^{q} |\hat{\Sigma}_{g,\Omega}|^{\frac{1}{2}N_g}}{|\hat{\Sigma}_\omega|^{\frac{1}{2}N}} = \frac{\prod_{g=1}^{q} |A_g|^{\frac{1}{2}N_g}}{|A|^{\frac{1}{2}N}} \frac{N^{\frac{1}{2}pN}}{\prod_{g=1}^{q} N_g^{\frac{1}{2}pN_g}}. \tag{26}$$

Here we can replace q in (26) as p, N_g in (26) as N and N in (26) as pN. Hence, $A_i = \sum_{a=1}^{N} (x_{a,i} - \bar{x}_i)^2$ and $A = \sum_{i=1}^{p} \sum_{a=1}^{N} (x_{a,i} - \bar{x}_i)^2$. After simplification, the likelihood ratio for H_2 can be rewritten as

$$\lambda_2 = \frac{\prod_{i=1}^{p} a_{ii}^{\frac{1}{2}N}}{(tr(A)/p)^{\frac{1}{2}pN}}, \tag{27}$$

which is the required criterion for H_2.

According to lemma 9, (25) and (27), the criterion for $H : \Sigma = \sigma^2 I$ is

$$\begin{aligned} \lambda &= \lambda_1 \lambda_2 \\ &= \frac{|A|^{\frac{1}{2}N}}{\prod_{i=1}^{q} |a_{ii}|^{\frac{1}{2}N}} \frac{\prod_{i=1}^{p} a_{ii}^{\frac{1}{2}N}}{(tr(A)/p)^{\frac{1}{2}pN}} \\ &= \frac{|A|^{\frac{1}{2}N}}{(tr(A)/p)^{\frac{1}{2}pN}}. \end{aligned} \tag{28}$$

In order to do hypothesis testing with the LR criterion, the next step is to find the distribution of this criterion. It can be noted that that matrix A follows Wishart distribution, which may be helpful for finding the distribution of LR criterion.

4.2 Distribution of LR Criterion

We first restate the basic settings and some results from previous sections. The LR criterion is $\lambda^* = (\frac{e}{n})^{\frac{1}{2}pn}|A|^{\frac{1}{2}n}e^{-\frac{1}{2}tr(A)}$ for $A = \sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})^\intercal$ and $n = N - 1$. The null hypothesis is $H : \Sigma = \sigma^2 I$. Note that $A \sim W(\Sigma, n)$. The idea to find the distribution of the LR criterion is that we want to first derive the moments of LR criterion and then find the characteristic function of it. At last, compare its characteristic function with some known distributions.

Denote $w(A|\Sigma, n)$ as the probability density function for the random matrix A. The $h-th$ moment of LR criterion is

$$\begin{aligned}\mathbb{E}(\lambda^{*h}) &= \int\cdots\int\left[(\frac{e}{n})^{\frac{1}{2}pn}|A|^{\frac{1}{2}n}e^{-\frac{1}{2}tr(A)}\right]^h w(A|\Sigma, n)dA \\ &= (\frac{e}{n})^{\frac{1}{2}pnh}\int\cdots\int|A|^{\frac{1}{2}nh}e^{-\frac{1}{2}tr(A)h}w(A|\Sigma, n)dA.\end{aligned} \tag{29}$$

Then, we want to rearrange the terms inside the integral so that the close form solution can be easily observed. Expand the terms in the integral and we can get

$$\begin{aligned}|A|^{\frac{1}{2}nh}e^{-\frac{1}{2}tr(A)h}w(A|\Sigma, n) &= \frac{|A|^{\frac{1}{2}(nh+n-p-1)}e^{-\frac{1}{2}tr(A(\Sigma^{-1}+hI))}}{2^{\frac{1}{2}pn}|\Sigma|^{\frac{1}{2}n}\Gamma_p(\frac{1}{2}n)} \\ &= w(A|(\Sigma^{-1}+hI)^{-1}, n+nh)\frac{\Gamma_p(\frac{1}{2}(n+nh))2^{\frac{1}{2}pnh}|\Sigma|^{\frac{1}{2}nh}}{\Gamma_p(\frac{1}{2}n)|I+h\Sigma|^{\frac{1}{2}(n+nh)}},\end{aligned} \tag{30}$$

which is a combination of a Wishart distribution and a constant term. Therefore, with (30) the $h-th$ moment in (29) can be written as

$$\mathbb{E}(\lambda^{*h}) = (\frac{2e}{n})^{\frac{1}{2}pnh}\frac{|\Sigma|^{\frac{1}{2}nh}}{|I+h\Sigma|^{\frac{1}{2}(n+nh)}}\frac{\prod_{j=1}^{p}\Gamma_p(\frac{1}{2}(n+nh+1-j))}{\prod_{j=1}^{p}\Gamma_p(\frac{1}{2}(n+1-j))}. \tag{31}$$

Hence, the characteristic function of $-2\ln(\lambda^*)$ can be easily derived, which is

$$\begin{aligned}\mathbb{E}(e^{-2it\ln\lambda^*}) &= \mathbb{E}(\lambda^{*-2it}) \\ &= (\frac{2e}{n})^{\frac{1}{2}-itpn}\frac{|\Sigma|^{\frac{1}{2}-itn}}{|I+h\Sigma|^{\frac{n}{2}-itn}}\frac{\prod_{j=1}^{p}\Gamma_p(\frac{1}{2}(n+1-j)-itn)}{\prod_{j=1}^{p}\Gamma_p(\frac{1}{2}(n+1-j))}\end{aligned} \tag{32}$$

By Stirling's formula for gamma function, which states that $\Gamma(x) \approx (\frac{2\pi}{x})^{\frac{1}{2}}(\frac{x}{e})^x$, and under the null hypothesis which states that $\Sigma = I$, the characteristic function can be further simplified as

$$\mathbb{E}(\lambda^{*-2it}) = (\frac{2e}{n})^{\frac{1}{2}-itpn}(1-2it)^{p(int-\frac{n}{2})}e^{int}\prod_{j=1}^{p}\frac{(\frac{1}{2}(n+1-j)-int)^{\frac{1}{2}(n-j)-int}}{(\frac{1}{2}(n+1-j)^{\frac{1}{2}(n-j)}} \tag{33}$$

Define

$$f_j(t) := (\frac{2e}{n})^{\frac{1}{2}-itpn}(1-2it)^{p(int-\frac{n}{2})}e^{int}\frac{(\frac{1}{2}(n+1-j)-int)^{\frac{1}{2}(n-j)-int}}{(\frac{1}{2}(n+1-j)^{\frac{1}{2}(n-j)}} \tag{34}$$

As n goes to infinity, $f_j(t)$ will converge to $(1-2it)^{-\frac{1}{2}j}$, which is exactly the characteristic function of χ^2_j. Therefore, we can conclude that

$$-2\ln\lambda^* \sim \chi^2_{\frac{1}{2}p(p+1)}, \tag{35}$$

which is the asymptotic distribution of LR criterion and it also matches the Wilks' theorem for large sample LR test.

4.3 Comparison of Hypothesis Testing Methods for High Dimensional Data

This section will compare the performance of unbiased LR criterion, Nagao's criterion [3] and corrected likelihood ratio test (CLRT) criterion which was found by Bai and Yao's group[4]. In our setting, high dimensional data means that the ratio $p/N \in [0,1]$ is quite large where p stands for the number of variables and N stands for sample size.

The testing problem can be formulated in the following. Let the $p \times 1$ vectors $X_1, X_2, \cdots, X_N$ be a random sample from $\mathcal{N}_p(\mu, \Sigma)$. Our target is to test the null hypothesis $H : \Sigma = I$.

Unbiased LR criterion The unbiased LR criterion, which was shown by Sugiura and Nagao [2], and is similar to the LR criterion derived in previous sections, is

$$\lambda_{LR} = (\frac{e}{n})^{\frac{1}{2}pn}|A|^{\frac{1}{2}n}e^{-\frac{1}{2}tr(A)} \tag{36}$$

for $n = N-1$, in which A is the same as defined in section 4.2. When n goes to infinity, $-2\ln\lambda_{LR}$ converges to χ^2 in distribution with degree of freedom $\frac{1}{2}p(p+1)$. It should be noticed that for likelihood ratio test, the condition $p < n$ is always necessary for us to derive the close form of the criterion. Otherwise, this method cannot be used to test our hypothesis since the target covariance matrix may not be invertible.

Nagao's criterion Nagao proposed a test criterion, which is

$$\lambda_{Na} = \frac{n}{2}tr(A/n - I)^2 \tag{37}$$

This criterion can be treated as a square loss, which characterizes the difference between the null hypothesis and the sample covariance matrix. Besides, this criterion does not require the condition $p < n$ since the covariance may not be invertible. Nagao also gave out the asymptotic distribution function for this criterion, which is

$$\begin{aligned} P(\lambda_{Na} \le x) = P_f + \frac{1}{n}(\frac{p}{12}(4p^2+9p+7)P_{f+6} - \frac{p}{8}(6p^2+13p+9)P_{f+4} \\ + \frac{p}{2}(p+1)^2P_{f+2} - \frac{p}{24}(2p^2+3p-1)P_f), \end{aligned} \tag{38}$$

in which $f = \frac{1}{2}p(p+1)$ and $P_f = P(\chi_f^2 \leq x)$. χ_f^2 denotes the Chi-square distribution with degree of freedom f. The detailed proof can be found in [3].

CLRT criterion Bai and Yao's group [4] proposed another test criterion for high dimension data, which is

$$\lambda_C = v(g)^{-\frac{1}{2}} \left[L^* - pF^{y_n}(g) - m(g)\right] \tag{39}$$

where y_n is the convergent value of the ratio p/n when both p and n go to infinity, F^{y_n} is the Marcenko-Pastur distribution of index y_n. In the formula (39), $S_n = \frac{1}{n}A$, $L^* = tr(S_n) - \ln|S_n| - p$, $m(g) = -\frac{\ln 1-y}{2}$, $v(g) = -2\ln 1 - y - 2y$ and $F^{y_n}(g) = 1 - \frac{y_n-1}{y_n}\ln 1 - y_n$, where y stands for the ratio p/n. Under the null hypothesis and when the sample size n goes to infinity, it was proved in (ref) that the CLRT criterion will converge to $\mathcal{N}(0,\, 1)$.

We perform a simulation study for different values of (p, N) in order to compare the performance of all three methods, where N is set to be 500 while p is chosen to be $5, 10, 50, 100, 300$. In the simulation study, we compute the realized sizes of the classic likelihood ratio test (LRT), Nagao's test and the corrected likelihood ratio test (CLRT). The significance level is set to be $\alpha = 0.05$. For each pair (p, N), we run 1,000 independent replications. The data is simulated from $\mathcal{N}_p(0,\, I)$. Our results are summarized in Table 1.

(p, n)	LRT size	Nagao size	CLRT size
$(5, 500)$	0.0520	0.0510	0.0550
$(10, 500)$	0.0640	0.0680	0.0540
$(50, 500)$	0.2140	0.0610	0.0320
$(100, 500)$	0.9740	0.0650	0.0340
$(300, 500)$	1.0000	0.0870	0.0300

Table 1: Hypothesis testing results

According to the above table, it can be readily noticed that the classic likelihood ratio test always reject the null hypothesis when the number of variables p becomes larger and larger, like $p = 100$ or 300. However, the sample is exactly generated from the distribution $\mathcal{N}_p(0,\, I)$, which means the LRT performs quite poorly for high dimensional data. For the Nagao's criterion, since this criterion will not be affected by the ratio p/n from its criterion formula, it may have the ability to deal with high dimensional data. But it may also be noted that, when the ratio p/n becomes larger, Nagao's test may be inclined to reject the null hypothesis. For the corrected likelihood ratio test, the size of it is always around $alpha = 0.05$ no matter how large the ratio p/n is, which means CLRT can handle high dimensional data as well as low dimensional data.

All in all, we may conclude that the classic LRT can handle low dimensional data well, Nagao's test can handle low dimensional data and may have the potential to handle high dimensional data, and CLRT can handle both kinds.

5 Further Discussions

Although we have showed that the corrected likelihood ratio test may be the best among the three tests when dealing high dimensional data, we still have a lot things to answer. In the derivations of the distributions of the criterion for all three tests, the close form distribution can only be obtained asymptotically, which means the sample size n is always required or assumed to be large enough. However, when the sample size is not that large, the real distribution may not match the asymptotic distribution, which may cause problems in our hypothesis testing. Another point is that we always require the condition $p/n \leq 1$. However, when it comes to some genetic data which may not satisfy this condition, hypothesis testing would be quite hard since there may not be suitable criterion or close form distribution to analyze our samples. Both problems mentioned before would surely be interesting topics to explore.

6 Acknowledgements

This report is written as a summary of the work done in the course STAT3799 Directed Studies in Statistics under the supervision of Prof. Jeff J. F. Yao. I sincerely thank Prof. Yao for his valuable suggestions and kind support.

References

[1] Johnstone, I. M. (2006). High dimensional statistical inference and random matrices. arXiv preprint math/0611589.

[2] Anderson, T. W. (1962). An introduction to multivariate statistical analysis (No. 519.9 A53). New York: Wiley.

[3] Nagao, H. (1973). On some test criteria for covariance matrix. The Annals of Statistics, 700-709.

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

[5] Dumitriu, I., Edelman, A. (2002). Matrix models for beta ensembles. Journal of Mathematical Physics, 43(11), 5830-5847.