A Likelihood Ratio Test of Independence of Components for High-dimensional Normal Vectors

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

Lin Zhang

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

Yongcheng Qi

May, 2013
Acknowledgements

I would like to thank my advisor, Yongcheng Qi for his support and guidance these years. Thank Dr. Kang Ling James, she gave me cherished advise and encouragement to my project and me. Thank Dr Barry James and Dr. Richard Green for serving on my project committee, taking the time to read this paper, and hearing the defense. Thank Xingguo Li, he helps me a lot on software use.
Dedication

To my dear mother.
Abstract

Consider a $p$-variate normal population. We are interested in testing the independence of its components based on a random sample of size $n$ from this population. In classic multivariate analysis, the dimension $p$ is fixed or relatively small compared with the sample size $n$, and the likelihood ratio test (LRT) is an effective way to test the hypothesis of independence, and the limiting distribution of the LRT is a chi-squared distribution. When $p$ goes to infinity, the chi-square approximation of the LRT may be invalid. In multivariate analysis, testing the independence of grouped components is one topic of interest. When the grouping is well balanced and the number of groups is fixed, the LRT, properly normalized, has a normal limit as proved in the literature. In practice, grouping can be unbalanced, and the number of groups can be arbitrarily large. In this project, we prove that the LRT statistic converges to a normal distribution under quite general conditions. Simulation results including histograms and comparisons of sizes and powers of tests with those in the classical chi-square approximations are presented as well.

Keywords: Likelihood ratio test, High-dimensional normal vector, Blocked Diagonal, covariance matrix.
# Contents

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td></td>
</tr>
<tr>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Description of the Model</td>
<td>2</td>
</tr>
<tr>
<td>2 Likelihood Ratio test</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Likelihood Ratio Test</td>
<td>3</td>
</tr>
<tr>
<td>2.2 Wilks' Statistic</td>
<td>5</td>
</tr>
<tr>
<td>3 Main Result</td>
<td>7</td>
</tr>
<tr>
<td>4 Some Lemmas</td>
<td>9</td>
</tr>
<tr>
<td>4.1 Multivariate Gamma Function and (\xi(x)) function</td>
<td>9</td>
</tr>
<tr>
<td>4.2 Moment of Wilks' Statistic</td>
<td>11</td>
</tr>
<tr>
<td>4.3 Infinitesimal and estimates</td>
<td>11</td>
</tr>
<tr>
<td>5 Proof of Theorem</td>
<td>17</td>
</tr>
</tbody>
</table>
6 Simulation 21

References 23
List of Tables

6.1 Size and Power of LRT for Specified Normal Distribution . . . . . . . . 21
List of Figures

6.1 Histograms of CLT and chi-square Approximation . . . . . . . . . . . . 22
Chapter 1

Introduction

1.1 Background

In classic statistical inference, the likelihood ratio test (LRT) is one widely-used method for hypothesis testing. An advantage of using the LRT is that one doesn’t have to estimate the variance of the test statistics. It is well known that the asymptotic distribution of the LRT is chi-square under certain regularity conditions when the dimension $p$ is a small constant or is negligible compared with the sample size $n$. However, the chi-square approximation doesn’t fit the distribution of the LRT very well for the high-dimension case, especially when $p$ grows with the sample size $n$. For many modern datasets, their dimensions can be proportionally large compared with the sample size. For example, financial data, consumer data, modern manufacturing data and multimedia data all have this feature. To deal with this feature, Schott(2001, 2005, 2007), Ledoit and Wolf(2002), Bai et al. (2009), Chen et al(2010), Jiang et al (2012), and Jiang and Yang(2012) derived different methods to study the classical likelihood ratio test when the dimension $p$ is large.

In this project, we are interested in testing the independence of grouped components from a high dimensional normal vector. The same problem has been considered by Jiang and Yang(2013), and Jiang and Qi (2013) when the number of the partition is fixed. The aim of the project is to extend the test to an arbitrary partitions and allow the number of the partition to change with the sample size $n$. The results are very applicable in practice. For example, in the analysis of microarray data on genes, it is meaningful to
check whether there is correlation among pieces of genes.

1.2 Description of the Model

This problem could be abstracted as the statistical model below. For a multivariate distribution $N_p(\mu, \Sigma)$, we partition a set of $p$ variates with a joint normal distribution into $k$ subsets and ask whether the $k$ subsets are mutually independent, or equivalently, we want to test whether variables among different subsets are dependent. More specifically, let $x_1, x_2, \cdots, x_n$ be i.i.d $\mathbb{R}^p$-valued random vectors with normal distribution $N_p(\mu, \Sigma)$. We write the blocked diagonal covariance matrix $\Sigma_0$ as

$$
\Sigma_0 = \begin{pmatrix}
\Sigma_{11} & 0 & \cdots & 0 \\
0 & \Sigma_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \Sigma_{kk}
\end{pmatrix}.
$$

Then we want to test the hypotheses

$$
H_0 : \Sigma = \Sigma_0 \ vs \ H_a : \Sigma \neq \Sigma_0. \quad (1.1)
$$

Let $W_n$ be Wilks’ likelihood ratio test (LRT) statistic (to be given in (2.11)). The traditional theory of multivariate analysis shows that $-\rho n \log W_n$ (Theorem 11.2.5 of Muirhead(1982)) goes to a chi-square distribution when $n$ tends to infinity and $p$ is fixed, where $\rho$ is given in (2.12). Jiang and Yang(2013) showed that the chi-square approximation is no longer true when $p \to \infty$. In fact, their results show that the central limit theorem (CLT) holds, i.e. $(\log W_n - \mu_n)/\sigma_n$ actually converges to the standard normal for a fixed number of partition $k$, where $\mu_n$ and $\sigma_n$ can be expressed explicitly as a function of sample size and partition.

In this project we will prove the CLT for the LRT, allowing that $k$ changes with $n$ and the partition can be unbalanced in the sense that numbers of components within subsets are not necessarily proportional.

The rest of the paper is organized as follows. In section 2, we introduce the LRT and give a brief literature review. In section 3, we state our main result. In section 4, we present a few lemmas, and in section 5, we provide the proof of the theorem given in section 3.
Chapter 2

Likelihood Ratio test

2.1 Likelihood Ratio Test

Let the $p$-component vector $x$ be distributed according to $N_p(\mu, \Sigma)$. We partition $x$ into $k$ subvectors.

$$x = (x^{(1)}, \cdots, x^{(k)})'$$  \hspace{1cm} (2.1)

where each $x^{(i)}$ has dimension $q_i$ respectively, with $p = \sum_{i=1}^{k} q_i$. The vector of means $\mu$ and the covariance matrix $\Sigma$ are partitioned similarly:

$$\mu = (\mu^{(1)}, \cdots, \mu^{(k)})'$$  \hspace{1cm} (2.2)

and

$$\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1k} \\
\Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
\Sigma_{k1} & \Sigma_{k2} & \cdots & \Sigma_{kk}
\end{pmatrix}.
$$

The null hypothesis is that the subvectors $x^{(1)}, \cdots, x^{(k)}$ are mutually independently distributed, i.e, the density of $x$ factors into the product of the density functions of $x^{(1)}, \cdots, x^{(k)}$:

$$H_0: f(x|\mu, \Sigma) = \prod_{i=1}^{k} f(x^{(i)}|\mu^{(i)}, \Sigma_{ii}).$$  \hspace{1cm} (2.3)
If $x^{(1)}, \cdots, x^{(k)}$ are independent subvectors, then the covariance matrix is block diagonal and denoted by $\Sigma_0$:

$$
\Sigma_0 = \begin{pmatrix}
\Sigma_{11} & 0 & \cdots & 0 \\
0 & \Sigma_{22} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \Sigma_{kk}
\end{pmatrix}
$$

with $\Sigma_{ii}$ unspecified for $1 \leq i \leq k$. Given a sample of size $n$, $x_1, \cdots, x_n$ are $n$ observations on $x$, the likelihood ratio is

$$
\Lambda_n = \frac{\max_{\{\mu, \Sigma_0\}} L(\mu, \Sigma_0)}{\max_{\{\mu, \Sigma\}} L(\mu, \Sigma)}, \quad (2.4)
$$

where

$$
L(\mu, \Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}p} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (x_i - \mu)' \Sigma^{-1} (x_i - \mu) \right\}. \quad (2.5)
$$

$L(\mu, \Sigma)$ is $L(\mu, \Sigma)$ with $\Sigma_{ij} = 0, \ i \neq j$, for all $0 \leq i, j \leq k$; and the maximum is taken with respect to all vectors $\mu$ and positive definite $\Sigma$ and $\Sigma_0$. According to Theorem 11.2.2 of Muirhead(1982), we have

$$
\max_{\mu, \Sigma} L(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}pn} |\hat{\Sigma}_\Omega|^{\frac{1}{2}n}} \exp\left\{ -\frac{1}{2} pn \right\}, \quad (2.6)
$$

where

$$
\hat{\Sigma} = \frac{1}{n-1} A = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x}). \quad (2.7)
$$

Under the null hypothesis,

$$
\max_{\mu, \Sigma_0} L(\mu, \Sigma_0) = \prod_{i=1}^{k} \max_{\mu^{(i)}, \Sigma_{ii}} L_i(\mu^{(i)}, \Sigma_{ii})
$$

$$
= \prod_{i=1}^{k} \frac{1}{(2\pi)^{\frac{1}{2}q_i n} |\hat{\Sigma}_{ii}|^{\frac{1}{2}n}} \exp\left\{ -\frac{1}{2} q_in \right\}
$$

$$
= \frac{1}{k} \exp\left\{ -\frac{1}{2} pn \right\}, \quad (2.8)
$$

$$
= \frac{1}{(2\pi)^{\frac{1}{2}pn} \prod_{i=1}^{k} |\hat{\Sigma}_{ii}|^{\frac{1}{2}n}}
$$
where
\[
\hat{\Sigma}_{ii} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j^{(i)} - \bar{x}^{(i)})(x_j^{(i)} - \bar{x}^{(i)})'.
\] (2.8)

If we partition \(A\) and \(\hat{\Sigma}\) in the same way for \(\Sigma\), we see that
\[
\hat{\Sigma}_{ii} = \frac{1}{n-1} A_{ii}.
\]

Then the likelihood ratio becomes
\[
\Lambda_n = \frac{\max_{\{\mu, \Sigma_0\}} L(\mu, \Sigma_0)}{\max_{\{\mu, \Sigma\}} L(\mu, \Sigma)} = \frac{\mid\hat{\Sigma}\mid^{\frac{1}{2}}n}{\prod_{i=1}^{k} \mid\hat{\Sigma}_{ii}\mid^{\frac{1}{2}}n} = \frac{\mid A\mid^{\frac{1}{2}}n}{\prod_{i=1}^{k} \mid A_{ii}\mid^{\frac{1}{2}}n}.
\] (2.9)

The critical region of the likelihood ratio test is
\[
\Lambda_n \leq \Lambda_n(\alpha),
\] (2.10)

where \(\Lambda_n(\alpha)\) is a number such that the probability of (2.10) is \(\alpha\) when \(\Sigma = \Sigma_0\).

### 2.2 Wilks’ Statistic

Now we employ Wilks’ statistic to do the test. Let
\[
W_n = \frac{\mid A\mid}{\prod_{i=1}^{k} \mid A_{ii}\mid};
\] (2.11)

\(W_n\) can be expressed entirely in terms of sample correlation coefficients. Note to see that \(\Lambda_n = W_n^{\frac{1}{2}}n\) is a monotonically increasing function of \(W_n\). The critical region can be equivalently written as \(W_n \leq W_n(\alpha)\). Notice that \(W_n = 0\) if \(p > n\), since the matrix \(A\) is not of full rank in this case. We see that the LRT test of level \(\alpha\) for testing \(H_0\) in (2.3) is \(\{\Lambda_n \leq C_\alpha\} = \{W_n \leq C_\alpha^{\frac{n}{2}}\}\). Set
\[
f = \frac{1}{2}(p^2 - \sum_{i=1}^{k} q_i^2) \quad \text{and} \quad \rho = 1 - \frac{2(p^3 - \sum_{i=1}^{k} q_i^2) + 9(p^2 - \sum_{i=1}^{k} q_i^2)}{6(n+1)(p^2 - \sum_{i=1}^{k} q_i^2)}. \] (2.12)
When \( n \to \infty \) while all \( q_i \)'s remain fixed, the traditional \( \chi^2 \) approximation to the distribution of \( \Lambda_n \) is found in Theorem 11.2.5 in Muirhead (1982):

\[
-2\rho \log(\Lambda_n) \xrightarrow{d} \chi_f^2.
\]

When \( p \) is large enough or is proportional to \( n \), this chi-square approximation may fail (Jiang and Yang (2013)).

Considering the insufficiency of the LRT when \( p \) is large, Bai et al. (2009) introduced a corrected likelihood ratio test for covariance matrices of Gaussian populations when the dimension is large compared to the sample size. They also developed a LRT to fit high-dimensional normal distribution \( N_p(\mu, \Sigma) \) with \( H_0 : \Sigma = I_p \). In their derivation, the dimension \( p \) is no longer a fixed constant, but rather is a variable that goes to infinity along with the sample size \( n \), and the ratio between \( p = p_n \) and converges to a constant \( y \), i.e.,

\[
\lim_{n \to \infty} \frac{p_n}{n} = y \in (0, 1)
\]

Jiang and Yang (2013) further extended by Bai et al. (2009) to cover the case of \( y = 1 \), and obtained the CLT of the LRT used for testing dependence of \( k \) groups of components for high-dimensional datasets, where \( k \) is a fixed number.
Chapter 3

Main Result

In this chapter, we will give the main result of the project.

Let \( k \geq 2, q_1, \cdots, q_k \) be an arbitrary partition of dimension \( p \). Denote \( p = \sum_{i=1}^{k} q_i \), and let

\[
\Sigma = (\Sigma_{ij})_{p \times p}
\]  

be the covariance matrix (positive definite), where \( \Sigma_{ij} \) is a \( q_i \times q_j \) sub-matrix for all \( 1 \leq i, j \leq k \). Consider the following hypotheses,

\[
H_0 : \Sigma = \Sigma_0 \quad \text{vs} \quad H_a : \Sigma \neq \Sigma_0
\]  

which is equivalent to hypothesis (2.3). Let \( S \) be the sample covariance matrix. Then partition \( A := (n - 1)S \) in the following way:

\[
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1k} \\
A_{21} & A_{22} & \cdots & A_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
A_{k1} & A_{k2} & \cdots & A_{kk}
\end{pmatrix}
\]

where \( A_{ij} \) is a \( q_i \times q_j \) matrix. Wilks(1935) presented that the likelihood ratio statistic for test (2.3)

\[
\Lambda_n = \frac{|A|^{n/2}}{\prod_{i=1}^{k} |A_{ii}|^{n/2}} = (W_n)^{n/2}.
\]  

When \( p > n + 1 \), the matrix \( A \) is not full rank, therefore, \( \Lambda_n \) is degenerate.

We have the following theorem.
Theorem. Let \( p \) satisfy \( p < n - 1 \) and \( p \to \infty \) as \( n \to \infty \). \( q_1, \ldots, q_k \) are \( k \) integers such that \( p = \sum_{i=1}^{k} q_i \) and \( \frac{\max \sum q_i}{p} \leq 1 - \delta \), for a fixed \( \delta \in (0, \frac{1}{2}) \) and all large \( n \). \( W_n \) is the Wilks likelihood ratio statistic described as (3.3). Then

\[
\frac{\log W_n - \mu_n}{\sigma_n} \overset{d}{\to} N(0,1) \tag{3.4}
\]
as \( n \to \infty \), where

\[
\mu_n = -c \log(1 - \frac{p}{n-1}) + \sum_{i=1}^{k} c_i \log(1 - \frac{q_i}{n-1}),
\]

\[
\sigma_n^2 = -\log \left(1 - \frac{p}{(n-1)}\right) + \sum_{i=1}^{k} \log \left(1 - \frac{q_i}{(n-1)}\right),
\]

with \( c = p - n + \frac{3}{2}, c_i = q_i - n + \frac{3}{2} \).

Remarks

- In the theorem above, integers \( k, q_1, \ldots, q_k \) and \( p \) all can depend on sample size \( n \). Compared with Jiang and Yang(2013) and Jiang and Qi(2013), constraint on the partition \( q_1, \ldots, q_k \) are relaxed.

- The assumption that \( \max \sum q_i \) \( p \leq 1 - \delta \) for some \( \delta \in (0,1) \) rules out the situation where \( \max \sum q_i \) \( p \to 1 \) along the entire sequence or any subsequence.

- This theorem works for all situations below.

  1. For \( q_1 = \cdots = q_p = 1 \), this satisfies the condition and it’s complete independence. (Theorem 6 of [6]).

  2. For relative balanced partition, i.e, when the number of each group is mutually proportional. (Theorem 2 of [6]).

  3. For unbalanced partitions, i.e, it is allowed for some of the groups dominating the partition. This will be verified by section 6 in the simulation part.
Chapter 4

Some Lemmas

4.1 Multivariate Gamma Function and $\xi(x)$ function

For two sequences of numbers $\{a_n\}$ and $\{b_n\}$ the notation $a_n = O(b_n)$ as $n \to \infty$ means $\lim \sup \frac{a_n}{b_n} < \infty$. $a_n = o(b_n)$ as $n \to \infty$ means $\lim \frac{a_n}{b_n} = 0$. $a_n \sim b_n$ as $n \to \infty$ stands for $\lim \frac{a_n}{b_n} = 1$.

Throughout the paper $\Gamma(x)$ stands for the Gamma function defined on $\mathbb{R}$, which is defined via an improper integral that converges. Define the multivariate Gamma function as:

$$\Gamma_p(x) := \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(x - \frac{1}{2}(i - 1)\right)$$

(4.1)

with $x > (n - p)/2$. See p. 62 in Muirhead (1982).

Define

$$\xi(x) = -2(\log(1 - x) + x), x \in [0, 1).$$

(4.2)

$\xi(x)$ is nonnegative in its domain. By the definition of $\xi(x)$, $\xi(0) = 0$, and $\xi'(x) = \frac{2x}{1-x}$.

We have

$$\xi(x) = \xi(x) - \xi(0) = \int_{0}^{x} \xi'(t)dt = \int_{0}^{x} \frac{2t}{1-t}dt$$

Substitute $t = ux$, then $dt = xdu$. Therefore

$$\xi(x) = \int_{0}^{1} \frac{2ux^2}{1-ux}du = 2 \int_{0}^{1} \frac{ux^2}{1-ux}du.$$  

(4.3)
Lemma 4.1.1. Let \( p \) satisfy \( p < n - 1 \) and \( p \to \infty \) as \( n \to \infty \). And let \( \{q_i\}_1^k \) be a partition of \( p \), i.e., \( p = \sum_{i=1}^k q_i \). Then

\[
\sigma_n^2 = \xi\left(\frac{p}{n-1}\right) - \sum_{i=1}^k \xi\left(\frac{q_i}{n-1}\right),
\]

and

\[
[1 - \sum_{i=1}^k \left(\frac{q_i}{p}\right)^2] \xi\left(\frac{p}{n-1}\right) \leq \sigma_n^2 \leq \xi\left(\frac{p}{n-1}\right).
\]

Furthermore, if for some \( \delta \in (0, \frac{1}{2}) \), \( \frac{\max_i q_i}{p} \leq 1 - \delta \) for all large \( n \), we have

\[
(1 - \delta) \xi\left(\frac{p}{n-1}\right) \leq \sigma_n^2 \leq \xi\left(\frac{p}{n-1}\right).
\]

Proof. Since \( \xi(x) \) is nonnegative in the domain, it follows that \( \sigma_n^2 \leq \xi\left(\frac{p}{n-1}\right) \).

By the integral expression of \( \xi(x) \) (4.3)

\[
\begin{align*}
\sigma_n^2 &= \xi\left(\frac{p}{n-1}\right) - \sum_{i=1}^k \xi\left(\frac{q_i}{n-1}\right) \\
&= 2 \int_0^1 \frac{u\left(\frac{p}{n-1}\right)^2}{1-u\left(\frac{p}{n-1}\right)} \, du - \sum_{i=1}^k 2 \int_0^1 \frac{u\left(\frac{q_i}{n-1}\right)^2}{1-u\left(\frac{q_i}{n-1}\right)} \, du \\
&\geq 2 \int_0^1 \frac{u\left(\frac{p}{n-1}\right)^2}{1-u\left(\frac{p}{n-1}\right)} \left[1 - \sum_{i=1}^k \left(\frac{q_i}{p}\right)^2\right] \, du \\
&= 2 \int_0^1 \frac{u\left(\frac{p}{n-1}\right)^2}{1-u\left(\frac{p}{n-1}\right)} \, du \\
&= [1 - \sum_{i=1}^k \left(\frac{q_i}{p}\right)^2] \xi\left(\frac{p}{n-1}\right).
\end{align*}
\]

This proves (4.5).

Since \( \max_i q_i/p \leq 1 - \delta \) and \( \sum_i q_i = p \), we have

\[
\sum_{i=1}^k \left(\frac{q_i}{p}\right)^2 \leq \frac{\max q_i}{p} \sum_{i=1}^k \frac{q_i}{p} = \frac{\max q_i}{p} \leq \delta,
\]
combined with (4.5), implying
\[ \sigma_n^2 \geq (1-\delta)\xi\left(\frac{p}{n-1}\right). \]
This completes the proof of Lemma 4.1.1. \(
\square
\)

4.2 Moment of Wilk’s Statistic

Lemma 4.2.1. \((\text{Theorem 11.2.3 from Muirhead}(1982))\) Let \(p = \sum_{i=1}^{k} q_i\) and \(W_n\) be Wilks’ likelihood ratio statistic defined as (3.3). Then
\[
EW_n^t = \frac{\Gamma_p(\frac{n-1}{2} + t)}{\Gamma_p(\frac{n-1}{2})} \prod_{i=1}^{k} \frac{\Gamma_{q_i}(\frac{n-1}{2})}{\Gamma_{q_i}(\frac{n-1}{2} + t)}
\]
for any \(t > (p-n)/2\), where \(\Gamma_p(x)\) is defined as (4.1).

4.3 Infinitesimal and estimates

Lemma 4.3.1. Let \(r_n \to \infty\) and \(r_n/n \to 0\) as \(n \to \infty\). Let \(q\) be a variable changing according to \(n\) and \(r_n \leq q < n - 1\).
\[
\lim_{n \to \infty} \sup_{r_n \leq q < n-1} \frac{q}{(n-1)(n-1-q)} = 0.
\]
And furthermore,
\[
\lim_{n \to \infty} \sup_{r_n \leq q < n-1} \frac{(q)(n-1)(n-1-q)^2}{\xi\left(\frac{q}{n-1}\right)} = 0.
\]
Proof. Set \(a = \frac{q}{n-1}\), then it is needed to prove
\[
\sup_{r_n \leq q < n-1} \frac{q}{(n-1)(n-1-q)} = \sup_{r_n \leq q < n-1} \frac{a}{n-1-q} = 0.
\]
Recall
\[
\xi(x) = 2\int_{0}^{1} \frac{ux^2}{1-ux}du = 2x^2\eta(x), \text{ where } \eta(x) = \int_{0}^{1} \frac{u}{1-ux}du.
\]
\(\eta(x)\) is a monotonically increasing function and \(\frac{1}{2} \leq \eta(x) \leq \infty\) for \(x \in [0,1]\).
Let $h$ be any positive constant in $h \in (0, \frac{1}{2})$.

When $r_n \leq q < h(n - 1)$, $\eta(a) \geq \frac{1}{2}$ and
\[
\sup_{r_n \leq q < h(n-1)} \frac{a}{n-1-q} \leq 2 \sup_{r_n \leq q < h(n-1)} \frac{1}{n-1-q} \frac{n-1}{q} = 2 \sup_{r_n \leq q < h(n-1)} \frac{1}{q(1-a)} \leq \frac{2}{r_n(1-h)}.
\]

When $h(n-1) \leq q \leq (1-h)(n-1)$, $h \leq a \leq 1-h$, we have
\[
\sup_{h(n-1) \leq q \leq (1-h)(n-1)} \frac{a}{n-1-q} \leq h(n-1) \sup_{h(n-1) \leq a \leq (1-h)(n-1)} \frac{1}{n-1-a} \frac{1}{q} \eta(a) = \frac{1}{n-1} \frac{1}{h^2} \log(1-h).
\]

When $(1-h)(n-1) < q < n-1$,
\[
\sup_{(1-h)(n-1) < q < n-1} \frac{a}{n-1-q} \leq \sup_{(1-h)(n-1) < q < n-1} \frac{1}{1-h} \frac{1}{\eta(1-h)}.
\]

Combining the 3 cases above and selecting $h_n = \frac{1}{\sqrt{n}}$, we get (4.7).

Since
\[
\frac{q}{(n-1)(n-q-1)} \leq 1 \quad \text{for} \quad r_n \leq q < n-1
\]
(4.8) follows from (4.7).

This completes the proof of Lemma 4.3. \qed

Lemma 4.3.2. Let $r_n \to \infty$ and $r_n/n \to 0$ as $n \to \infty$. Then
\[
\sum_{i=1}^{q} \left( \frac{1}{n-i} - \frac{1}{n-1} \right) = -\log(1 - \frac{q}{n-1}) - \frac{q}{n-1} + O\left( \frac{q}{(n-1)(n-1-q)} \right) \quad (4.10)
\]
and
\[
\sum_{i=1}^{q} (\log(n-1) - \log(n-i)) = (n-q-\frac{1}{2}) \log(1 - \frac{q}{n-1}) + q + O\left( \frac{q}{(n-1)(n-q-1)} \right) \quad (4.11)
\]
uniformly on $r_n \leq q < n-1$. 

Proof. By the partial sum of harmonic series,
\[ \sum_{i=1}^{k} \frac{1}{i} = \log k + \gamma + \frac{1}{2k} + \tau(k), \]
where \( \gamma \) is the Euler-Mascheroni constant and \( \tau(k) = o\left(\frac{1}{k^2}\right) \) as \( k \to \infty \). See "Sloane’s A082912 : Sum of \( a_n \) terms of harmonic series is > 10\(^n\)”, The On-Line Encyclopedia of Integer Sequences. OEIS Foundation.

First, note that
\[
\sum_{i=1}^{q} \left( \frac{1}{n-i} - \frac{1}{n-1} \right) = \sum_{i=1}^{n-1} \frac{1}{i} - \sum_{i=1}^{n-1-q} \frac{1}{i} - \frac{q}{n-1} = -\log(1 - \frac{q}{n-1}) - \frac{q}{n-1} - \frac{q}{2(n-1)(n-1-q)} + \tau(n-1-q).
\]
To show (4.10), it suffices to show \( \tau(n-1-q) = O\left(\frac{q}{(n-1)(n-1-q)}\right) \). Note that \( \tau(k) \leq \frac{c}{k^2} \), \( k \geq 1 \) for some \( c > 0 \). We can also verify that
\[
\frac{n-1}{q(n-1-q)} \leq 2.
\]
As a result,
\[
\tau(n-1-q) = \frac{n-1}{q(n-1-q)} \frac{q}{(n-1)(n-1-q)} \leq \frac{2cq}{(n-1)(n-1-q)} \quad (4.12)
\]
This finishes the proof of (4.10).

To show (4.11), we apply the Stirling formula. Recall the Stirling formula from Ahlfors(1979) is equivalent to
\[
\log \Gamma(x) = (x - \frac{1}{2}) \log(x) - x + \frac{1}{2} \log(2\pi) + \frac{1}{12x} + O\left(\frac{1}{x^2}\right) \quad (4.13)
\]
as \( x \to \infty \).

Take \( x = n-1 \) and \( x = n-q-1 \) respectively and take the difference, then we have
as $n \to \infty$, 

$$
\begin{align*}
\log \Gamma(n - 1) - \log \Gamma(n - q - 1) \\
= (n - \frac{3}{2}) \log(n - 1) - (n - q - \frac{3}{2}) \log(n - q - 1) - q \\
+ \frac{1}{12} \left( \frac{1}{n - 1} - \frac{1}{n - q - 1} \right) + O\left( \frac{1}{(n - 1)^2} \right) + O\left( \frac{1}{(n - q - 1)^2} \right) \\
= (n - \frac{3}{2}) \log(n - 1) - (n - q - \frac{3}{2}) \log(n - q - 1) - q \\
- \frac{q}{12(n - 1)(n - q - 1)} + O\left( \frac{1}{(n - 1)^2} \right) + O\left( \frac{1}{(n - q - 1)^2} \right) \\
= (n - \frac{3}{2}) \log(n - 1) - (n - q - \frac{3}{2}) \log(n - q - 1) - q \\
+ O\left( \frac{q}{(n - 1)(n - q - 1)} \right).
\end{align*}
$$

In the last step, we have used the following estimate

$$
\frac{1}{(n - 1)^2} + \frac{1}{(n - q - 1)^2} = O\left( \frac{q}{(n - 1)(n - q - 1)} \right). \tag{4.14}
$$

For any $r_n \leq q < n - 1$, uniformly,

$$
\frac{1}{(n - 1)^2} \leq \frac{q}{(n - 1)(n - q - 1)}
$$

which, combined with (4.11), gives (4.13).

Note that for any integer $m \geq 1$, $\Gamma(m) = (m - 1)!$. Then we have

$$
\begin{align*}
\sum_{i=1}^{q} (\log(n - 1) - \log(n - i)) \\
=q \log(n - 1) - (\log \Gamma(n - 1) - \log \Gamma(n - q - 1)) + \log(1 - \frac{q}{n - 1}) \\
=q \log(n - 1) - (n - \frac{3}{2}) \log(n - 1) + (n - q - \frac{3}{2}) \log(n - q - 1) + q \\
+ O\left( \frac{q}{(n - 1)(n - q - 1)} \right) + \log(1 - \frac{q}{n - 1}) \\
= (n - q - \frac{1}{2}) \log(1 - \frac{q}{n - 1}) + q + O\left( \frac{q}{(n - 1)(n - q - 1)} \right).
\end{align*}
$$

This finishes the proof of (4.11). \qed
Lemma 4.3.3. (Lemma 2.1 from Jiang and Qi(2013)) Let \( b := b(x) \) be a real-valued function defined on \((0, \infty)\).

\[
\log \left( \frac{\Gamma(x + b)}{\Gamma(x)} \right) = (x + b) \log(x + b) - x \log x - b - \frac{b^2 + |b| + 1}{2x} + O\left(\frac{b^2 + |b| + 1}{x^2}\right) \tag{4.15}
\]

holds uniformly on \( b \in [-lx, lx] \) for any given constant \( l \in (0, 1) \). Furthermore, as \( x \to \infty \),

\[
\log \left[ \frac{\Gamma(x + b)}{\Gamma(x)} \cdot \frac{\Gamma(z)}{\Gamma(z + b)} \right] = b(\log x - \log z) + \frac{b^2 - b}{2} \left( \frac{1}{x} - \frac{1}{z} \right) + O\left(\frac{|b|^3(x - z)}{x^3}\right) + O\left(\frac{b^2 + |b| + 1}{x^2}\right) \tag{4.16}
\]

uniformly over \( b \in [-lx/2, lx/2] \) and \( z \in [x/2, x] \) for any \( l \in (0, 1) \).

Lemma 4.3.4. Let \( t = t_n \) be a bounded variable with respect to \( n \). Let \( r_n \to \infty \) and \( r_n/n \to 0 \) as \( n \to \infty \). Then

\[
\log \left[ \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2} + t\right)} \right]^q \log \left[ \frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2} - t\right)} \right] = t[(q-n+\frac{3}{2})\log(1 - \frac{q}{n-1}) - \frac{n-2}{n-1}q] + t^2 \frac{\xi(q, n)}{2} + O\left(\frac{|t|^3 q^2}{n^3}\right) + O\left(\frac{|t|^2 + |t| + 1}{n^2}\right). \tag{4.17}
\]

holds uniformly for \( r_n \leq q < n - 1 \).

Proof. From the definition of \( \Gamma_p(x) \) in (4.1) we have

\[
\log \left[ \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2} + t\right)} \right]^q \log \left[ \frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2} - t\right)} \right] = - \sum_{i=1}^{q} \log \left[ \frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2} - i\right)} \right]. \tag{4.17}
\]

By applying Lemma 4.3.3 to the summands, we have

\[
\log \left[ \frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2} - i\right)} \right] = t(\log(n - 1) - \log(n - i)) + O\left(\frac{|t|^3 (i - 1)}{n^3}\right) + O\left(\frac{|t|^2 + |t| + 1}{n^2}\right) = t(\log(n - 1) - \log(n - i) - \frac{1}{n - 1} + \frac{1}{n - i}) + O\left(\frac{|t|^3 (i - 1)}{n^3}\right) + O\left(\frac{|t|^2 + |t| + 1}{n^2}\right)
\]
uniformly for $1 \leq i \leq q$. Then by Lemma 4.3.2,

\[
\sum_{i=1}^{q} \log \left[ \frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{\Gamma\left(\frac{n-i}{2}\right)}{\Gamma\left(\frac{n-i}{2} + t\right)} \right] = t\left[ \sum_{i=1}^{q} \left( \frac{1}{n-i} - \frac{1}{n-1} \right) \right] - t^2 \sum_{i=1}^{q} \left( \frac{1}{n-i} - \frac{1}{n-1} \right) + O\left( \sum_{i=1}^{q} \left( \frac{|t|^3(i-1)}{n^3} \right) \right) + O\left( \frac{q(|t|^2 + |t| + 1)}{n^2} \right)
\]

\[
= -t\left[ (n-q-\frac{1}{2}) \log(1 - \frac{q}{n-1}) + q + O\left( \frac{q}{(n-1)(n-q-1)} \right) - \log(1 - \frac{q}{n-1}) \right] - \frac{q}{n-1} - t^2 \left[ \frac{1}{2} \xi\left( \frac{q}{n-1} \right) - O\left( \frac{q}{(n-1)(n-1-q)} \right) \right] + O\left( \frac{|t|^3q(q-1)/2}{n^3} \right)
\]

\[
+ O\left( \frac{q(|t|^2 + |t| + 1)}{n^2} \right)
\]

\[
= -t\left[ (q-n+\frac{3}{2}) \log(1 - \frac{q}{n-1}) - \frac{n-2}{n-1}q \right] - t\left[ O\left( \frac{q}{(n-1)(n-q-1)} \right) \right] - t^2 \left[ \frac{1}{2} \xi\left( \frac{q}{n-1} \right) \right] - t^2 \left[ O\left( \frac{q}{(n-1)(n-1-q)} \right) \right] + O\left( \frac{|t|^3q^2}{n^3} \right) + O\left( \frac{q(|t|^2 + |t| + 1)}{n^2} \right)
\]

which together with (4.17) completes the proof of the lemma. \qed
Chapter 5

Proof of Theorem

To prove (3.4), it suffices to show that

\[ E \exp\left(\frac{\log W_n - \mu_n}{\sigma_n}\right)s = \exp\left(-\frac{\mu_n s}{\sigma_n}\right)E[W_n^{\sigma_n}] \rightarrow e^{s^2/2} \]

as \( n \rightarrow \infty \) for all \( s \) such that \(|s| \leq 1\) and for corresponding \( \mu_n \) and \( \sigma_n \), or equivalently

\[ \log EW_n^t = \mu_n t + \frac{\sigma_n^2 t^2}{2} + o(1) \]  

(5.1)
as \( n \rightarrow \infty \) with \(|\sigma_n t| \leq 1\). By Lemma 4.2.1,

\[ EW_n^t = \frac{\Gamma_p\left(\frac{n-1}{2} + t\right)}{\Gamma_p\left(\frac{n-1}{2}\right)} \prod_{i=1}^k \frac{\Gamma_{q_i}\left(\frac{n-1}{2}\right)}{\Gamma_{q_i}\left(\frac{n-1}{2} + t\right)} \]

(5.2)

By taking logarithms on both sides, we have

\[ \log EW_n^t = -\sum_{i=1}^p \log\left[\frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{\Gamma\left(\frac{n-i}{2}\right)}{\Gamma\left(\frac{n-i}{2} + t\right)}\right] + \sum_{i=1}^k \sum_{j=1}^{q_i} \log\left[\frac{\Gamma\left(\frac{n-1}{2} + t\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{\Gamma\left(\frac{n-j}{2}\right)}{\Gamma\left(\frac{n-j}{2} + t\right)}\right]. \]
For the first sum of the right hand side, set $c = p - n + \frac{3}{2}$ and $g(t) = t^2 + |t| + 1$. By Lemma 4.3.4 with $q = p$, we have

\[
\begin{align*}
&- \sum_{i=1}^{p} \log \left[ \frac{\Gamma \left( \frac{n-1}{2} + t \right)}{\Gamma \left( \frac{n-1}{2} \right)} \frac{\Gamma \left( \frac{n-1}{2} + t \right)}{\Gamma \left( \frac{n-1}{2} + t \right)} \right] \\
&= t \left[ c \log(1 - \frac{p}{n-1}) - \frac{n-2}{n-1} p + \frac{t^2}{2} \xi \left( \frac{p}{n-1} \right) \right] \\
&+ (t + t^2)O \left( \frac{p}{(n-1)(n-p-1)} \right) + O\left( \frac{|t|^3p^2}{n^3} \right) + O\left( \frac{pg(t)}{n^2} \right)
\end{align*}
\]

For the second sum on the right hand side of (5.2), we set $c_i = q_i - n + \frac{3}{2}$ and note that $p = \sum q_i$ and $\frac{\max_i q_i}{p} \leq 1 - \delta$ for some $\delta \in (0, \frac{1}{2})$. Again using Lemma 4.3.4 with $q = q_i$, we have

\[
\begin{align*}
&\sum_{i=1}^{k} \sum_{j=1}^{q_i} \log \left[ \frac{\Gamma \left( \frac{n-1}{2} + t \right)}{\Gamma \left( \frac{n-1}{2} \right)} \frac{\Gamma \left( \frac{n-1}{2} + t \right)}{\Gamma \left( \frac{n-1}{2} + t \right)} \right] \\
&= - \sum_{i=1}^{k} \left( t \left[ c_i \log(1 - \frac{q_i}{n-1}) - \frac{n-2}{n-1} q_i + \frac{t^2}{2} \xi \left( \frac{q_i}{n-1} \right) \right] \right) \\
&- (t + t^2) \sum_{i=1}^{k} O \left( \frac{q_i}{(n-1)(n-1-q_i)} \right) + O\left( \frac{|t|^3 \sum q_i^2}{n^3} \right) + O\left( \frac{pg(t)}{n^2} \right)
\end{align*}
\]

\[
\begin{align*}
&= t \left[ - \sum_{i=1}^{k} c_i \log(1 - \frac{q_i}{n-1}) \right] + \frac{n-2}{n-1} p - \frac{t^2}{2} \sum_{i=1}^{k} \xi \left( \frac{q_i}{n-1} \right) \\
&- (t + t^2) \sum_{i=1}^{k} O \left( \frac{q_i}{(n-1)(n-1-q_i)} \right) + O\left( \frac{pg(t)}{n^2} \right).
\end{align*}
\]
Therefore,

\[
\log EW_n^t = t[c \log(1 - \frac{p}{n-1}) - \sum_{i=1}^{k} c_i \log(1 - \frac{q_i}{n-1})] + \frac{t^2}{2} [\xi(\frac{p}{n-1}) - \sum_{i=1}^{k} \xi(\frac{q_i}{n-1})] \\
+ (t + t^2) [O(\frac{p}{(n-1)(n-p-1)}) + \sum_{i=1}^{k} O(\frac{q_i}{(n-1)(n-1-q_i)})] + O(\frac{pg(t)}{n^2}) \\
= t\mu_n + \frac{t^2}{2} \sigma_n^2 \\
+ (t + t^2) [O(\frac{p}{(n-1)(n-p-1)}) + \sum_{i=1}^{k} O(\frac{q_i}{(n-1)(n-1-q_i)})] + O(\frac{pg(t)}{n^2}),
\]

where

\[
\mu_n = -c \log(1 - \frac{p}{n-1}) + c_i \sum_{i=1}^{k} \log(1 - \frac{q_i}{n-1}), \\
\sigma_n^2 = \xi(\frac{p}{n-1}) - \sum_{i=1}^{k} \xi(\frac{q_i}{n-1}).
\]

Now it is sufficient to prove as \( n \to \infty , \)

\[
(t + t^2) [O(\frac{p}{(n-1)(n-p-1)}) + \sum_{i=1}^{k} O(\frac{q_i}{(n-1)(n-1-q_i)})] + O(\frac{pg(t)}{n^2}) = o(1) \quad (5.3)
\]

for all \( |\sigma_n t| \leq 1. \) Recall the constraint that \( p \leq n - 2, \) \( p \to \infty \) and \( \frac{\max_i q_i}{p} \leq 1 - \delta \) for some \( \delta \in (0, \frac{1}{2}) . \)

We can verify that

\[
\sum_{i=1}^{k} O(\frac{q_i}{(n-1)(n-1-q_i)}) \leq \sum_{i=1}^{k} O(\frac{q_i}{(n-1)(n-1-p)}) \\
\leq O(\frac{\sum q_i}{(n-1)(n-1-p)}) = O(\frac{p}{(n-1)(n-1-p)}).
\]

Equation (5.3) becomes

\[
(t + t^2) O(\frac{p}{(n-1)(n-p-1)}) + O(\frac{pg(t)}{n^2}) = o(1). \quad (5.4)
\]

From (4.5), we have

\[
|t| \leq \frac{1}{\sigma_n} = O(\frac{1}{\sqrt{\xi(\frac{p}{n-1})}}).
\]
By (4.8) and (4.9) of Lemma 4.3.1 with \( q = p \),

\[
t[tO\left(\frac{p}{(n-1)(n-1-p)}\right)] = O\left(\frac{p}{\sqrt{\xi(p/n-1)}}\right) = o(1),
\]

(5.5)

\[
t^2[O\left(\frac{p}{(n-1)(n-1-p)}\right)] = O\left(\frac{p}{\xi(p/n-1)}\right) = o(1)
\]

(5.6)

Recall \( g(t) = t^2 + |t| + 1 \) and for any \( p \to \infty \) and \( p < n - 1 \), \( \xi(p/n-1) = O((p/n)^2) \), with \( p/n-1 \leq \theta \), when \( \theta \) is small in \( \theta \in (0, 1) \). Then

\[
O\left(\frac{pg(t)}{n^2}\right) = O\left(\frac{pt^2}{n^2}\right) + O\left(\frac{p|t|}{n^2}\right) + O\left(\frac{p}{n^2}\right)
\]

\[
= O\left(\frac{p (n-1)^2}{n^2 p^2} + O\left(\frac{p n - 1}{n^2 p}\right) + o(1)
\]

\[
= O\left(\frac{1}{p}\right) + O\left(\frac{1}{n}\right) + o(1) = o(1).
\]

Also when \( \theta \leq \frac{p}{n-1} \leq 1 \), by the increasing property of \( \xi(x) \) in (4.2),

\[
\frac{pt^2}{n^2} \leq \frac{p}{n^2}/\xi\left(\frac{p}{n-1}\right) \leq \frac{p}{n^2}/\xi(\theta) \to 0.
\]

Similarly

\[
\frac{pt}{n^2} \leq \frac{p}{n^2}/\sqrt{\xi\left(\frac{p}{n-1}\right)} \leq \frac{p}{n^2}/\sqrt{\xi(\theta)} \to 0,
\]

and

\[
O\left(\frac{pg(t)}{n^2}\right) = O\left(\frac{pt^2}{n^2}\right) + O\left(\frac{p|t|}{n^2}\right) + O\left(\frac{p}{n^2}\right) = o(1).
\]

This completes the proof of (3.4).
Chapter 6

Simulation

In this section, we compare the performance of the chi-square approximation and the normal approximation through a finite sample simulation study. We plot the histograms for the chi-square statistics which are used for the chi-square approximation and compare with their corresponding limiting chi-square curves. Similarly, we plot the histograms of the statistic which is used for the normal approximation and compare with the standard normal curve.

We also report estimated sizes and powers of tests for the LRT tests based on the chi-square approximation and the normal approximation. All simulations have been done by using R, and the histograms, estimates of the sizes and powers are based on 10,000 replicates.

Table 6.1: Size and Power of LRT for Specified Normal Distribution

<table>
<thead>
<tr>
<th></th>
<th>Size under $H_0$</th>
<th>Power under $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CLT</td>
<td>$\chi^2$ approx.</td>
</tr>
<tr>
<td>$p = 10, q_1 = 3, q_2 = 2, q_3 = 2, q_4 = 3$</td>
<td>0.0359</td>
<td>0.049</td>
</tr>
<tr>
<td>$p = 80, q_1 = 30, q_2 = 20, q_3 = 20, q_4 = 10$</td>
<td>0.0482</td>
<td>0.0814</td>
</tr>
<tr>
<td>$p = 80, q_1 = 3, q_2 = 2, q_3 = 2, q_4 = 73$</td>
<td>0.0487</td>
<td>0.2462</td>
</tr>
<tr>
<td>$p = 80, q_1 = 3, q_2 = 77$</td>
<td>0.0423</td>
<td>0.4877</td>
</tr>
</tbody>
</table>

Sizes (alpha errors) are estimated based on 10,000 simulations from $N_p(0, I_p)$. The powers are estimated under the alternative hypothesis that $\Sigma = 0.15J_p + 0.85I_p$, where $J_p$ is matrix with all entries as 1.
Figure 6.1: Histograms of CLT and chi-square Approximation
References


