High-Dimensional Behavior of Some Multivariate Two-Sample Tests

Shan Shi

A Thesis
for The Department of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts (Mathematics) at
Concordia University
Montreal, Quebec, Canada

November 2015

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Shan Shi**

Entitled: **High-Dimensional Behavior of Some Multivariate Two-Sample Tests**

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________ Examiner
Dr. Wei Sun

______________________ Examiner
Dr. Yogendra Chaubey

______________________ Thesis Supervisor
Dr. Arusharka Sen

Approved by ______________________
Chair of Department or Graduate Program Director

____
Dean of Faculty

Date ______________________

ABSTRACT

High-dimensional behavior of some multivariate two-sample tests

Shan Shi

It is a difficult problem to test the equality of distribution of two independent p-dimensional ($p > 1$) samples (of sizes m and n, say) in a nonparametric framework. It is not only because we need deal with issues such as tractability of the null distribution of test-statistics but also the fact that the latter are rarely distribution-free. Several notable nonparametric tests for comparing multivariate distributions are the multivariate runs test of Friedman and Rafsky (1979), the nearest-neighbor test of Henze (1988) and the inter-point distance-based test of Baringhaus and Franz (BF) (2004). Biswas and Ghosh (BG) (2014) recently have shown that in a high dimension, low sample-size (HDLSS) scenario, i.e. where p goes to infinity but m, n are small or fixed, all the tests mentioned do not perform well. However, the BG-test is shown to be consistent in the case of HDLSS. In this work, we study the asymptotic behaviors of BF and BG tests when m, n and p go to infinity and $\min(m, n) = o(p)$. Our results reveal when these tests are expected to work well and when they are not. Results are illustrated by simulated data.

Acknowledgments

I'd like to express my sincere gratitude to my supervisor Dr. Arusharka Sen for all his supports and help throughout my Master's degree. And, I want to thank you to all of the wonderful professors at Concordia University who have helped me to broaden my knowledge in math and encouraged me to pursue my passion for academic research. I would also like to thank Dr. Wei Sun and Dr. Yogendra Chaubey for reviewing my thesis.

Contents

List of Figures **vii**

Introduction **1**

0.1 High-Dimensional Data Problems 1

0.2 Hypothesis Testing For Distributions 2

0.3 $U-$ Statistics 5

0.4 Thesis Organization 6

1 Behavior of the BF-test as $m, n, p \to \infty$ **7**

1.1 Assumptions 7

1.2 Taylor Approximation of T_{mn}^{BF*} 8

1.3 Asymptotic normality of T_{mn}^{BF*} Under H_0 10

1.4 Ratio-Consistent Estimators 15

2 Behavior of the BG-test as $m, n, p \to \infty$ **17**

2.1 Taylor Approximation of $\frac{N}{p}T_{mn}^{BG}$ 17

2.2 Asymptotic distribution of $\frac{N}{p}T_{mn}^{BG}$ 18

2.3 A ratio consistent estimator for $\mathbb{V}(\frac{N}{p}T_{mn}^{BG})$ 20

2.4 Power properties of the BG-test 25

3 Simulation Results **27**

3.1 Simulation of Power curves 27

3.2 Conclusions of Simulation 29

Bibliography **32**

List of Figures

3.1 Power curves of BF-test 28

3.2 Power curves of BG-test in the case of HDLSS 30

3.3 Power curves of BG-test 31

Introduction

Classical statistical data analyses can only be applied in the case where the dimension of observations is fixed and the sample size grows. In many recent practical applications, such as data mining and microarray studies, we face a new challenge that is the number of variables exceeds the number of observations dramatically. Due to the growing dimension of data, many classical statistical data analysis tool are not available any more.

0.1 High-Dimensional Data Problems

In order to show some challenges happening in high-dimensional settings, let's consider two random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$, where $X_i, Y_j \in \mathbb{R}^p$, for all $1 \leq i, j \leq m, n$. Now, let $\mu_1 = \mathbb{E}(X_i)$, $\mu_2 = \mathbb{E}(Y_j)$ where $\mu_1 = (\mu_{11}, \mu_{12}, \ldots, \mu_{1p})'$, $\mu_2 = (\mu_{21}, \mu_{22}, \ldots, \mu_{2p})'$ and the covariance matrices $\Sigma_1 = cov(X_i)$, $\Sigma_2 = cov(Y_j)$. Let's assume we are interested in testing

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2.$$

Traditionally, the Hotteling T^2 test is a widely used mean test. The test statistic is defined as follow

$$T^2 = \frac{mn}{m+n} \left(\bar{X} - \bar{Y}\right) S_N^{-1} \left(\bar{X} - \bar{Y}\right)$$

where $N = m + n$, $\bar{X}, \bar{Y}$ are sample mean vectors and S_N is the pooled sample covariance matrix defined as

$$S_N = \frac{1}{m+n-2} \left[\sum_{i=1}^{m} \left(X_i - \bar{X}\right) \left(X_i - \bar{X}\right)' + \sum_{j=1}^{n} \left(Y_j - \bar{Y}\right) \left(Y_j - \bar{Y}\right)' \right].$$

Under the null hypothesis, $\frac{N-p+1}{Np}T^2$ has a central F-distribution with p and $N-p+1$ degrees of freedom. However, Bai and Saranadasa (1996) show that the asymptotic power of Hotelling's test decrease as the ratio of the dimension to the sample size, p/N, increases to 1. When $p > N$, that is, the dimension is larger than the sample size, the Hotelling's test is not well defined because the sample covariance matrix becomes singular. Thus, Bai and Saranadasa (1996) proposed replacing $(\bar{X}-\bar{Y})'S_N^{-1}(\bar{X}-\bar{Y})$ in the Hottelling's test with $\|\bar{X}-\bar{Y}\|$, where $\|\cdot\|$ denotes the Euclidean norm. The test statistic they established, under some mild conditions, shows attractive power as $p/n \to c < \infty$.

0.2 Hypothesis Testing For Distributions

The high-Dimensional challenges also arise in problems of two-sample hypothesis testing for distributions. Before we proceed, let's have a short review on two-sample hypothesis testing for distributions. In such tests, we are interested in either $H_0 : F = G$ or $H_1 : F \neq G$. In other words, we would like to know if two sets of independent observations $X_i \sim F, 1 \leq i \leq m$ and $Y_j \sim G, 1 \leq j \leq n$ share the same distribution function. The two-sample test problem has been studied for a long time in fixed dimension settings. In the univariate case, some distribution-free and consistent tests such as the Kolmogorov-Smirnov, Wald-Wolfowitz runs and Wilcoxon rank sum tests are commonly applied. However, the multivariate case seems not as straightforward as the univariate case. The easily noticed reason is the fact that the multivariate tests often are not distribution-free under H_0. Notable among many proposals for test-statistics are the multivariate runs test of Friedman and Rafsky (1979), the nearest-neighbor test of Henze (1988) and the inter-point distance-based test of Baringhaus and Franz (2004).

Baringhaus and Franz (2004) proposed the test for arbitrary dimensions setting, which is based on the average sample Euclidean inter-point distances, where

$$T_{m,n}^{BF} = \frac{mn}{m+n}\left[\frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}\|X_j - Y_k\| - \frac{1}{2m^2}\sum_{j=1}^{m}\sum_{k=1}^{m}\|X_j - X_k\| - \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\|Y_j - Y_k\|\right],$$

m, n are sample sizes and $\|\cdot\|$ is the Euclidean norm. The BF-test is simply motivated by the fact proven by Baringhaus and Franz (2004)

$$2\mathbb{E}(\|X-Y\|)-\mathbb{E}(\|X-X^*\|)-\mathbb{E}(\|Y-Y^*\|)\geq 0$$

where $X, X^* \overset{i.i.d}{\sim} F, Y, Y^* \overset{i.i.d}{\sim} G$ and the equality holds if and only if $F = G$. The null hypothesis will be rejected when $T^{BF}_{m,n}$ is large. And, $T^{BF}_{m,n}$ converges in distribution to an integrated, squared Brownian bridge depending on an unknown distribution, where the theorem is the following.

Theorem 0.2.1. *Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be independent p-dimensional random vectors. They have the same distribution function H. Then, as $\min(m,n)\to\infty$, the random variables $T_{m,n}$ converge in distribution to*

$$T=\gamma_p\int B^2_H(a,t)d\mu\otimes\lambda(a,t),$$

where $(B_H(a,t);(a,t)\in S^{p-1}\times\mathbf{R})$ is a H-Brownian bridge having the covariance function

$$Cov(B_H(a,t),B_H(b,s))=\mathbb{P}(a'X_1\leq t, b'X_1\leq s)-\mathbb{P}(a'X_1\leq t)\mathbb{P}(b'X_1\leq s)$$

with $(a,t),(b,s)\in S^{p-1}\times\mathbf{R}$.

Since BF-test depends on an unknown distribution the authors suggested to simulate critical values by using the bootstrap method.

Recently, Biswas and Ghosh (2014) have demonstrated that in a high dimension, low sample-size (HDLSS) scenario, i.e., where $p\to\infty$ but m,n are small, all the tests mentioned above exhibit poor power. So, they proposed another test related to BF-test. The BG-test statistic is $T^{BG}_{m,n}=\|\hat{\mu}_{DF}-\hat{\mu}_{DG}\|^2$, where

$$\hat{\mu}_{DF}=\left[\hat{\mu}_{FF}=\binom{m}{2}^{-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\|X_i-X_j\|,\ \hat{\mu}_{FG}=(mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\|X_i-Y_j\|\right],$$

$$\hat{\mu}_{DG}=\left[\hat{\mu}_{FG}=(mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\|X_i-Y_j\|,\ \hat{\mu}_{GG}=\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\|Y_i-Y_j\|\right].$$

Like BF-test, the null hypothesis will be rejected when $T_{m,n}^{BG}$ is large. The BG-test works because, if we let $\mu_{FF} = \mathbb{E}(\|X - X^*\|),\ \mu_{GG} = \mathbb{E}(\|Y - Y^*\|),\ \mu_{FG} = \mathbb{E}(\|X - Y\|), \mu_{DF} = (\mu_{FF},\ \mu_{FG})',\ \mu_{DG} = (\mu_{FG},\ \mu_{GG})'$, then

$$\|\mu_{DF} - \mu_{DG}\|^2 = 0 \Leftrightarrow \mu_{DF} = \mu_{DG} \Leftrightarrow \mu_{FF} = \mu_{GG} = \mu_{FG}$$

$$\Leftrightarrow 2\mu_{FG} - \mu_{FF} - \mu_{GG} = 0 \Leftrightarrow F = G.$$

The motivation is they observed that in a high dimension, low sample-size (HDLSS) scenario, the BF-test has bad performances only when $v^2 \leq |\sigma_1^2 - \sigma_2^2|$, where $\sigma_1^2 = \lim\limits_{p\to\infty} \frac{trace(\Sigma_1)}{p}$, $\sigma_2^2 = \lim\limits_{p\to\infty} \frac{trace(\Sigma_2)}{p}$, $v^2 = \lim\limits_{p\to\infty} \frac{\|\mu_1 - \mu_2\|^2}{p}$, and $\Sigma_1 = Cov(X_i)$, $\Sigma_2 = Cov(Y_j)$, $\mu_1 = \mathbb{E}(X_i), \mu_2 = \mathbb{E}(Y_j)$, $1 \leq i, j \leq m, n$. The authors realized the fact $v^2 \leq |\sigma_1^2 - \sigma_2^2|$ implies, assuming $\sigma_1 \leq \sigma_2$

$$\sqrt{2}\sigma_1 \leq \sqrt{\sigma_1^2 + \sigma_2^2 + v^2} \leq \sqrt{2}\sigma_2.$$

And, according to their assumptions, $\frac{(\hat{\mu}_{FG} - \hat{\mu}_{FF})}{\sqrt{p}} \to_p (\sqrt{\sigma_1^2 + \sigma_2^2 + v^2} - \sqrt{2}\sigma_1)$ and $\frac{(\hat{\mu}_{FF} - \hat{\mu}_{GG})}{\sqrt{p}} \to_p (\sqrt{\sigma_1^2 + \sigma_2^2 + v^2} - \sqrt{2}\sigma_2)$. Even when $(\hat{\mu}_{FG} - \hat{\mu}_{FF})$ and $(\hat{\mu}_{FF} - \hat{\mu}_{GG})$ may be significantly different from zero, both most likely have different signs. Thus, $T_{m,n}^{BF} = (\hat{\mu}_{FG} - \hat{\mu}_{FF}) + (\hat{\mu}_{FF} - \hat{\mu}_{GG})$ might be close to zero, so H_0 may not be rejected. $T_{m,n}^{BG}$ does not have the same weakness since the cancellation is impossible to happen when $T_{m,n}^{BG} = (\hat{\mu}_{FG} - \hat{\mu}_{FF})^2 + (\hat{\mu}_{FF} - \hat{\mu}_{GG})^2$. When sample size is large and the dimension of data remains fixed, $(m+n)T_{m,n}^{BG}$ converge in distribution to $\frac{2\sigma_0^2}{\lambda(1-\lambda)}\chi_1^2$, where $\lambda = \frac{m}{n}$, $\sigma_0^2 = \mathbb{V}\left(\mathbb{E}\left(\|X_1 - X_2\| | X_1\right)\right)$. While the sample size is fixed and the dimension of data increases, the power of the BG-test of level α converges to 1 if $\lim_{p\to\infty} \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{p} \neq 0$ or $\lim_{p\to\infty} \frac{trace(\boldsymbol{\Sigma_1})}{p} \neq \lim_{p\to\infty} \frac{trace(\boldsymbol{\Sigma_2})}{p}$ is assumed.

It is easy to note that both BF-test and BG-test are linear combinations of $U-$Statistics, which is a very powerful tool and has been widely employed since 1948 when Hoeffding first introduced it to the world. Recently, people have attempted to use the asymptotic theory of $U-$ statistics especially in the degenerating case to tackle high-dimensional problems. However, Ahmad et al. (2014), claimed that no mentionable bibliography indicated that the asymptotic theory of degenerate $U-$ statistics was successfully applied to high-dimensional problems. But, Ahmad et al. (2014) proposed a way to apply degenerate U-statistics theory on high-dimensional problems without explicit proof. In this work, following Jung, Sen and Marron (2012), we are going to simplify Ahmad et al. (2014)'s method by using principal components. We therefore review some basic definitions and theorems of U-statistics.

0.3 $U-$ Statistics

Let $X_1, X_2, X_3, \dots$ be i.i.d random variables with common distribution function $F(x)$. let $m \geq 1$ and $h : \mathbb{R}^m \to \mathbb{R}$ be a measurable function symmetric in its arguments. The U-statistic with kernel h is defined by

$$U_n(h) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} h\left(X_{i_1}, \dots, X_{i_m}\right), n \geq m.$$

The kernel h is called degenerate with respect to $F(x)$ if for all $1 \leq j \leq m$,

$$\int_{\mathbb{R}} h(x_1, x_2, \dots, x_m) dF(x_j) = 0, \ where \ -\infty < x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_m < \infty.$$

Let

$$\theta = \mathbb{E}h(X_1, \dots, X_m)$$

and for $i = 0 \dots, m$ let

$$h_i(x_1, \dots, x_i) = \mathbb{E}h(x_1, \dots, x_i, X_{i+1}, \dots, X_m)$$

$$\sigma_i^2 = \mathbb{V}(h_i(X_1, \dots, X_i))$$

so that

$$\sigma_0^2 = 0$$

$$\sigma_m^2 = \mathbb{V}(h(X_1, \dots, X_m))$$

We say that a U-statistic is degenerate if $\sigma_1^2 = 0$.

Theorem 0.3.1. *Let U_n be a U-statistic based on a kernel function h of degree m, then*

$$\mathbb{V}(U_n) = \binom{n}{m}^{-1} \sum_{i=1}^{m} \binom{m}{i} \binom{n-m}{m-i} \sigma_i^2$$

0.4 Thesis Organization

In this work, we will show how BF-test and BG-test behave in the high dimensional setting where $n, p \to \infty$ and $n = o(p)$. In Chapter 1 and 2 the asymptotic distributions of the test statistics of BF-test and BG-test under the null hypothesis and theirs power properties respectively are given. In Chapter 3, a comparison between BF and BG tests would be made.

Chapter 1

Behavior of the BF-test as $m, n, p \to \infty$

The goal of this chapter is to show how BF-test behaves in the case of the high-dimensional setting. Instead of analyzing the original BF-test, we would like to work on the modified BF-test. Its test statistic is

$$T_{m,n}^{BF*} = (m+n)\left[\frac{2\sum_{j=1}^{m}\sum_{k=1}^{n}\|X_j - Y_k\|}{mn} - \frac{\sum_{j=1}^{m}\sum_{k=1}^{m}\|X_j - X_k\|}{m(m-1)} - \frac{\sum_{j=1}^{n}\sum_{k=1}^{n}\|Y_j - Y_k\|}{n(n-1)}\right].$$

To do the investigation, we need first find out the main contributors of $T_{m,n}^{BF*}$ by analyzing its Taylor approximation. Then, we would derive its limiting distribution under H_0.

1.1 Assumptions

In order to carry out the investigation on these two tests, we need first to make the following assumptions, following Hall and Neeman (2005) and Biswas and Ghosh (2014).

(A1) The fourth moments of the components of X and Y are uniformly bounded, where $X, Y \in \mathbb{R}^p$.

(A2) (i) $\frac{trace(\Sigma_1)}{p} \to \sigma_1^2$, (ii) $\frac{trace(\Sigma_2)}{p} \to \sigma_2^2$, (iii) $\frac{\|\mu_1 - \mu_2\|^2}{p} \to v^2$, (iv) $\frac{trace(\Sigma_i \Sigma_j)}{p^2} \to c_{ij}$, as $p \to \infty$ where $i, j = 1, 2$.

(A3) $X = (X^{(1)}, X^{(2)}, \ldots, X^{(p)}), Y = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(p)})$ are ρ mixing for functions dominated by quadratics if whenever functions f and g of two variables satisfy $|f(u, v) +$

$g(u,v)| \leq Cu^2v^2$, for some $C > 0$, and all u, v, we have

$$\sup_{1\leq l,k<\infty,|k-l|\geq r} |corr\left[f(U^{(k)},V^{(k)}),\ g(U^{(l)},V^{(l)})\right]| \leq \rho(r),$$

for $(U,V) = (X,X),(Y,Y),(X,Y)$

As a consequence of the assumptions, we will have the following

$$(i)\ \|\frac{X_i - X_j}{p}\| \rightarrow_p \sqrt{2}\sigma_1, (ii)\ \|\frac{Y_i - Y_j}{p}\| \rightarrow_p \sqrt{2}\sigma_2, (iii)\ \|\frac{X_i - Y_j}{p}\| \rightarrow_p \sqrt{\sigma_1^2 + \sigma_2^2 + v^2}$$

since the variances of them go to zero as p goes to infinity. For more details, please refer to Hall and Neeman (2005).

1.2 Taylor Approximation of T_{mn}^{BF*}

The way to show how BF-test behaves is to find the main contributors of the test statistic. So, let's begin with the Taylor expansion of $T_{m,n}^{BF*}$.

By simple calculation, we know

$$\begin{aligned}
\frac{\mu_x}{p} &= \frac{\mathbb{E}\left(\|X_i - X_j\|^2\right)}{p} = \frac{2trace\left(\Sigma_1\right)}{p} \approx 2\sigma_1^2 \\
\frac{\mu_y}{p} &= \frac{\mathbb{E}\left(\|Y_i - Y_j\|^2\right)}{p} = \frac{2trace\left(\Sigma_2\right)}{p} \approx 2\sigma_2^2 \\
\frac{\mu_{xy}}{p} &= \frac{\mathbb{E}\left(\|X_i - Y_j\|^2\right)}{p} = \frac{trace(\Sigma_1 + \Sigma_2)}{p} + \frac{\|\mu_1 - \mu_2\|^2}{p} \approx \sigma_3^2 = \sigma_1^2 + \sigma_2^2 + v^2
\end{aligned}$$

where $\mu_1 = \mathbb{E}(X_i)$ and $\mu_2 = \mathbb{E}(Y_j)$, for all $i, j > 0$. And, it is easy to check that by Taylor expansions of the function $x \to \sqrt{x}$ centering at $\sigma_1^2, \sigma_2^2, \sigma_3^3$

$$\begin{aligned}
\frac{\hat{\mu}_{FF}}{\sqrt{p}} &= \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \| \frac{X_i - X_j}{p} \| \\
&= \sqrt{2}\sigma_1 + \frac{1}{m\sqrt{2}\sigma_1} \sum_{i=1}^{m} \left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \sigma_1^2 \right) \\
&- \frac{2}{m(m-1)\sqrt{2}\sigma_1} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{(X_i-\mu_1)^{'}(X_j-\mu_1)}{p} + R_{\mu_{FF}}
\end{aligned}$$

$$\begin{aligned}
\frac{\hat{\mu}_{GG}}{\sqrt{p}} &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \| \frac{Y_i - Y_j}{p} \| \\
&= \sqrt{2}\sigma_2 + \frac{1}{n\sqrt{2}\sigma_2} \sum_{i=1}^{n} \left(\frac{(Y_i-\mu_2)^{'}(Y_i-\mu_2)}{p} - \sigma_2^2 \right) \\
&- \frac{2}{n(n-1)\sqrt{2}\sigma_2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{(Y_i-\mu_2)^{'}(Y_j-\mu_2)}{p} + R_{\mu_{GG}}
\end{aligned}$$

$$\begin{aligned}
\frac{\hat{\mu}_{FG}}{\sqrt{p}} &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \| \frac{X_i - Y_j}{p} \| \\
&= \sigma_3 + \frac{1}{2m\sigma_3} \sum_{i=1}^{m} \frac{(X_i-\mu_1)^{'}(X_i-\mu_1) - p\sigma_1^2}{p} \\
&+ \frac{1}{2n\sigma_3} \sum_{i=1}^{n} \frac{(Y_i-\mu_2)^{'}(Y_i-\mu_2) - p\sigma_2^2}{p} - \frac{1}{mn\sigma_3} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(X_i-\mu_1)^{'}(Y_j-\mu_2)}{p} \\
&- \frac{1}{mn\sigma_3} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{((X_i-\mu_1)-(Y_j-\mu_2))^{'}(\mu_1-\mu_2)}{p} \\
&+ \frac{1}{2\sigma_3} \left(\frac{\|\mu_1-\mu_2\|^2}{p} - v^2 \right) + R_{\mu_{FG}}
\end{aligned}$$

Thus, the modified BF-test can be expressed as the following

$$\frac{T^{BF*}_{m,n}}{\sqrt{p}} = (m+n)\left(2\sigma_3 - \sigma_1 - \sigma_2\right) + (m+n)\frac{1}{2\sigma_3}\left(\frac{\|\mu_1 - \mu_2\|^2}{p} - v^2\right) \tag{1.1}$$

$$+ (m+n)(\frac{2}{m(m-1)\sqrt{2}\sigma_1} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{\left(X_i - \mu_1\right)^{'} \left(X_j - \mu_1\right)}{p} \tag{1.2}$$

$$+ \frac{2}{n(n-1)\sqrt{2}\sigma_2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\left(Y_i - \mu_2\right)^{'} \left(Y_j - \mu_2\right)}{p} \tag{1.3}$$

$$- \frac{2}{mn\sigma_3} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(X_i - \mu_1\right)^{'} \left(Y_j - \mu_2\right)}{p} \tag{1.4}$$

$$+ (\frac{1}{m\sigma_3} - \frac{1}{m\sqrt{2}\sigma_1}) \sum_{i=1}^{m} \frac{\left(X_i - \mu_1\right)^{'} \left(X_i - \mu_1\right) - p\sigma_1^2}{p} \tag{1.5}$$

$$+ (\frac{1}{m\sigma_3} - \frac{1}{n\sqrt{2}\sigma_2}) \sum_{i=1}^{n} \frac{\left(Y_i - \mu_2\right)^{'} \left(Y_i - \mu_2\right) - p\sigma_2^2}{p} \tag{1.6}$$

$$- \frac{1}{mn\sigma_3} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left((X_i - \mu_1) - (Y_j - \mu_2)\right)^{'} \left(\mu_1 - \mu_2\right)}{p} \tag{1.7}$$

$$+ 2R_{\mu_{FG}} - R_{\mu_{FF}} - R_{\mu_{GG}}) \tag{1.8}$$

The term $(1.1), (1.5), (1.6), (1.7)$ from above equal zero under the null hypothesis. And, term (1.8) would be proven negligible in Theorem 1.3.4. Thus, the main contributors of the BF-test statistic, under null hypothesis are $(1.2), (1.3)$ and (1.4).

1.3 Asymptotic normality of T^{BF*}_{mn} Under H_0

We are going to derive the limiting distribution of T^{BF*}_{mn} in the way of Ahmad et al. (2014). First, we need to present the following important lemma. A similar proof has been given by Chen and Qin (2010) by using Martingale Central Limit Theorem. Here, we are going to present a different proof based on properties of degenerate U-statistics and the principal components inspired by Jung, Sen and Marron (2012).

Lemma 1.3.1. *Let $X_i \in \mathbb{R}^p$ and $\mathbb{E}(X_i) = 0$, for all $0 < i < m$, which satisfy (A1)-(A3). Then $T_m = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{X_i^{'}X_j}{p} \rightarrow_d Y = \lim_{p\to\infty}\sum_{i=1}^{p}\frac{\lambda_i}{p}\left(W_i^2 - 1\right)$, where $W_1^2, W_2^2, \ldots$ being independent χ_1^2 random variables, as $p, m \rightarrow \infty$.*

Proof. we shall prove this result by the method of characteristic function, that is, by showing

$$\mathbb{E}\left(e^{ixT_m}\right) \to \mathbb{E}\left(e^{ixY}\right), m, p \to \infty$$

Let $\omega_i = \Lambda^{-1/2}P^{-1}X_i$, where $P\Lambda P^{-1} = \Sigma_1$ and $diag\,(\Lambda) = (\lambda_1, \lambda_2, \ldots, \lambda_p)$, so

$$X_i^{'}X_j = (\Lambda^{-1/2}P^{-1}X_i)^{'}\Lambda(\Lambda^{-1/2}P^{-1}X_j) = \omega_i^{'}\Lambda\omega_j = \sum_{k=1}^{p}\lambda_k\omega_{ki}\omega_{kj}$$

Thus,

$$T_m = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{\omega_i^{'}\Lambda\omega_j}{p}$$

It is noted that each value of $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ depends on $p(m)$, but for convenience we drop $p(m)$. And,

$$|\mathbb{E}\left(e^{ixT_m}\right) - \mathbb{E}\left(e^{ixY}\right)| \leq |\mathbb{E}\left(e^{ixT_m}\right) - \mathbb{E}\left(e^{ixT_{mk}}\right)| + |\mathbb{E}\left(e^{ixT_{mk}}\right) - \mathbb{E}\left(e^{ixY_k}\right)| + |\mathbb{E}\left(e^{ixY_k}\right) - \mathbb{E}\left(e^{ixY}\right)|$$

where

$$T_{mk} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{s=1}^{k}\frac{\lambda_s}{p}\omega_{si}\omega_{sj}$$

and

$$Y_k = \sum_{s=1}^{k}\frac{\lambda_i}{p}\left(W_s^2 - 1\right).$$

Using the inquelity $|e^{iz} - 1| \leq |z|$ we have

$$\begin{aligned}
|\mathbb{E}\left(e^{ixT_m}\right) - \mathbb{E}\left(e^{ixT_{mk}}\right)| &\leq \mathbb{E}|e^{ixT_m} - e^{ixT_{mk}}| \\
&\leq |\mathbb{E}e^{ixT_{mk}}||\mathbb{E}e^{ix(T_m - T_{mk})} - 1| \\
&\leq |\mathbb{E}\left(x(T_m - T_{mk})\right)| \\
&\leq |x||\mathbb{E}(T_m - T_{mk})^2|^{\frac{1}{2}}
\end{aligned}$$

$$\begin{aligned}\mathbb{E}(T_m - T_{mk})^2 &= \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{s=k+1}^{p}\frac{\lambda_s}{p}\omega_{si}\omega_{sj}\right)^2 \\ &= \mathbb{E}\left(\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(\sum_{s=k+1}^{p}\frac{\lambda_s}{p}\omega_{si}\omega_{sj}\right)^2\right) \\ &\leq \sum_{s=k+1}^{p}\frac{\lambda_s^2}{p^2} \leq (\sum_{s=k+1}^{p}\frac{\lambda_s}{p})^2\end{aligned}$$

Since $\frac{trace(\Sigma_1)}{p} \to \sigma_1, p \to \infty$, and $\sum_{s=1}^{p}\frac{\lambda_s}{p} = \frac{trace(\Sigma_1)}{p}$, $\frac{trace(\Sigma_1)}{p}$ is Cauchy. So, there is a P such that $\sum_{s=k+1}^{p}\frac{\lambda_s}{p} \leq \epsilon$ for all $k \geq P$,So, $|\mathbb{E}\left(e^{ixT_m}\right) - \mathbb{E}\left(e^{ixT_{mk}}\right)| \leq \epsilon$ when $k \geq P$.

Next, let's show that $|\mathbb{E}\left(e^{ixT_{mk}}\right) - \mathbb{E}\left(e^{ixY_k}\right)| \leq \epsilon$. We may rewrite T_{mk} as

$$T_{mk} = \frac{1}{m}\sum_{s=1}^{k}\frac{\lambda_s}{p}\left(W_{mk}^2 - Z_{nk}\right),$$

where

$$W_{mk} = m^{-\frac{1}{2}}\sum_{i=1}^{m} w_{ki}$$

and

$$Z_{mk} = m^{-1}\sum_{s=1}^{m} w_{ki}^2.$$

Since $\mathbb{E}W_{ki} = 0, Var(W_{mk}) = 1, Cov\left(W_{mk}, W_{ml}\right) = 0$ for all $k \neq l$. And, w_{ki} depends on m for all k, but for convenience we drop the m. Thus, by Lindeberge-Feller CLT, we have

$$\left(W_{m1}, W_{m2}, \ldots, W_{mk}\right) \to_d N\left(0, \boldsymbol{I}_{k\times k}\right), m \to \infty.$$

And,

$$\left(Z_{m1}, Z_{m2}, \ldots . Z_{mk}\right) \to_p \left(1, 1, \ldots, 1\right), m \to \infty.$$

Consequently, we have $|\mathbb{E}\left(e^{ixT_{mk}}\right) - \mathbb{E}\left(e^{ixY_k}\right)| \leq \epsilon$ for some M such that when all $m \geq M$. Last, we need to show that $|\mathbb{E}\left(e^{ixY_k}\right) - \mathbb{E}\left(e^{ixY}\right)| \leq \epsilon$. If we assume $Y_k \to_d Y$ then we can find a K such that $|\mathbb{E}\left(e^{ixY_k}\right) - \mathbb{E}\left(e^{ixY}\right)| \leq \epsilon$ for all $k \geq K$.

Thus, we can find a $L \geq Max(m(P), M, m(K))$, so $|\mathbb{E}\left(e^{ixT_m}\right) - \mathbb{E}\left(e^{ixY}\right)| \leq 3\epsilon$ for all $m \geq L$.

□

Remark. *The way of deriving the limiting distribution suggested by Ahmad et al. (2014) is same as shown in Serfling (1980). That is, find a L^2 convergent sequence of the kernel. Principal component method clearly gives what we want in our case.*

According to the previous analysis, we present the following theorem about the limiting distribution of the BF-test statistic.

Theorem 1.3.2. *Under $H_0 : F = G$, and the assumptions A(1)-A(3),*

$$\frac{T^{BF*}_{m,n}}{2\sqrt{\frac{tr(\Sigma^2)}{tr(\Sigma)}}} \to_d N(0, \zeta_1^2 + \zeta_2^2 + \frac{\zeta_3^2}{2})$$

as $\min(m, n), p \to \infty$, *$n = o(p)$, where* $\zeta_1 = \lim \frac{m+n}{m-1}, \zeta_2 = \lim \frac{m+n}{n-1}, \zeta_3 = \lim \frac{m+n}{\sqrt{mn}}$.

Proof. Under H_0, without the loss of generality, we can assume that $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = \sigma$ and $\Sigma_1 = \Sigma_2 = \Sigma$. So,

$$\begin{aligned}
\frac{1}{\sqrt{p}} T^{BF*}_{mn} &= \frac{N}{\sqrt{p}} \left(2\hat{\mu}_{FG} - \hat{\mu}_{FF} - \hat{\mu}_{GG}\right) \\
&= \frac{2N}{m(m-1)\sqrt{2}\sigma} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{X_i^{'} X_j}{p} + \frac{2N}{n(n-1)\sqrt{2}\sigma} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{Y_i^{'} Y_j}{p} \\
&- \frac{2N}{mn\sqrt{2}\sigma} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{X_i^{'} Y_j}{p} + 2NR_{\mu_{FG}} - NR_{\mu_{FF}} - NR_{\mu_{GG}}
\end{aligned}$$

where, $N = m + n$. Now, let $\Phi_1 = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{X_i^{'} X_j}{p}$, $\Phi_2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{Y_i^{'} Y_j}{p}$, and $\Phi_3 = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{X_i^{'} Y_j}{p}$. It is easy to see that Φ_1, and Φ_2 are one-sample U- Statistics, and Φ_3 is a two-sample U- Statistic. And, Φ_1, Φ_2 and Φ_3 are degenerate U- statistics, since $\mathbb{E}\left(X_i | X_j\right) = \mathbb{E}\left(Y_i | Y_j\right) = 0$. From Lemma 1, we know that

$$\frac{1}{m} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{X_i^{'} X_j}{p} \to_d \sum_{i=1}^{\infty} \frac{\lambda_i}{p} \left(Z_{1i}^2 - 1\right)$$

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{Y_i^{'} Y_j}{p} \to_d \sum_{i=1}^{\infty} \frac{\lambda_i}{p} \left(Z_{2i}^2 - 1\right)$$

$$\frac{1}{\sqrt{mn}}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{X_i'Y_j}{p} \to_d \sum_{i=1}^{\infty}\frac{\lambda_i}{p}\left(Z_{1i}Z_{2i}\right)$$

as $p, m \to \infty$, where λ_i is a eigenvalue of Σ, Z_{1i} and Z_{2i} are two independent sequences of independent standard normal variables. So,

$$\frac{\sqrt{\frac{2trace(\mathbf{\Sigma})}{p}}}{\sqrt{2}\sigma}\frac{\sqrt{2}\sigma}{2\sqrt{p}}T_{mn}^{BF*} \to_d \frac{N}{m-1}\sum_{i=1}^{\infty}\frac{\lambda_i}{p}\left(Z_{1i}^2-1\right)+\frac{N}{n-1}\sum_{i=1}^{\infty}\frac{\lambda_i}{p}\left(Z_{2i}^2-1\right)-\frac{N}{\sqrt{mn}}\sum_{i=1}^{\infty}\frac{\lambda_i}{p}\left(Z_{1i}Z_{2i}\right).$$

According to

Lemma 1.3.3. *Let X_i be i.i.d. random variables with mean 0 and variance 1. let $b_{ni}, 1 \leq i \leq n$ be a sequence of constants such that $\max_i \ b_{ni}^2 \to 0$ as $n \to \infty$ then*

$$\sum_{i=1}^{n} b_{ni}X_i \to_d N(0,1)$$

as $n \to \infty$.

And, ϕ_1, ϕ_2 and ϕ_3 are uncorrelated. Thus,

$$\frac{1}{\sqrt{\frac{2trace(\Sigma^2)}{p^2}}}\sqrt{\frac{2trace\left(\mathbf{\Sigma}\right)}{p}}\frac{1}{2\sqrt{p}}T_{mn}^{BF*} \to_d N(0, \zeta_1^2+\zeta_2^2+\frac{\zeta_3^2}{2}).$$

□

In order to finish the proof of theorem 1.3.2, we need to show that the remainders converge to 0 in probability. We will show $NR_{\mu_{FF}} \to_p 0$, the others can be shown in the same way.

Theorem 1.3.4. *Follow the same assumptions as Theorem 1.3.2, as $p, min(m,n) \to \infty$ and $min(m,n) = o(p)$, $NR_{\mu_{FF}} \to_p 0$, where*

$$NR_{\mu_{FF}} = \frac{N}{m(m-1)}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{1}{8\xi_{ij}^{\frac{3}{2}}}\left(\frac{\left(X_i-X_j\right)'\left(X_i-X_j\right)}{p}-2\sigma_1^2\right)^2$$

where ξ_{ij} falls between $\frac{(X_i-X_j)'(X_i-X_j)}{p}$ and $2\sigma_1^2$.

Proof. According to Hall et al.(2005), we know that $\|\frac{X_i - X_j}{p}\| \to_p \sqrt{2}\sigma_1$ and $\frac{\|X_i - X_j\|}{\sqrt{p}} = \sqrt{2}\sigma_1 + O_p\left(\frac{1}{\sqrt{p}}\right)$. Thus, $\frac{\|X_i - X_j\|^2}{p} = 2\sigma_1^2 + O_p(\frac{1}{\sqrt{p}}) + O_p(\frac{1}{p})$. So,

$$NR_{\mu_{FF}} = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{1}{8\xi_{ij}^{\frac{3}{2}}} \left(\sqrt{N} 2\sigma_1^2 - \sqrt{N} 2\sigma_1^2 + O_p(\frac{\sqrt{N}}{\sqrt{p}}) + O_p(\frac{\sqrt{N}}{p}) \right)^2 .$$

Since $\frac{N}{p} \to 0$ then, $NR_{\mu_{FF}} \to_p 0$ □

1.4 Ratio-Consistent Estimators

To make BF-test useful in high-dimensional settings, we need to find ratio-consistent estimators of trace (Σ), trace (Σ^2) under $\Sigma_1 = \Sigma_2 = \Sigma$. In other words, we need estimators such that

$$\frac{\widehat{\text{trace}\,(\Sigma)}}{\text{trace}\,(\Sigma)} \to_p 1 \qquad\qquad \frac{\widehat{\text{trace}\,(\Sigma^2)}}{\text{trace}\,(\Sigma^2)} \to_p 1.$$

Since $\frac{\text{trace}(\Sigma)}{p}$ and $\frac{\text{trace}\left(\Sigma^2\right)}{p^2}$ are bounded for all $p > 0$, we only need find consistent estimators for them. We can use the following two estimators

$$\frac{\widehat{\text{trace}\,(\Sigma)}}{p} = \frac{m}{m+n} \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \|\frac{X_i - X_j}{p}\|^2 + \frac{n}{m+n} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \|\frac{Y_i - Y_j}{p}\|^2$$

$$\frac{\widehat{\text{trace}\,(\Sigma^2)}}{p^2} = \frac{m}{m+n} \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \|\frac{X_i - X_j}{p}\|^4 + \frac{n}{m+n} \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \|\frac{Y_i - Y_j}{p}\|^4$$

Theorem 1.4.1. *Under the assumptions A(1)-A(3), and assume H_0 is true, $\frac{\widehat{trace(\Sigma)}}{p}$ and $\frac{\widehat{trace(\Sigma^2)}}{p}$ are consistent unbiased estimators.*

Proof. It is clear that each estimator is a sumation of two $U-$statistics. So we only need to show that the $\mathbb{E}\left(\|\frac{X_i - X_j}{p}\|^2\right) < \infty$ and $\mathbb{E}\left(\|\frac{X_i - X_j}{p}\|^4\right) < \infty$ because $U-$ statistics converge to its mean almost surely if $\mathbb{E}\left(|h|\right) < \infty$, where h is the kernel function of a $U-$statistic. Thus, we only need show the follwing. Without the loss of generality, we can assume $\mathbb{E}X_i = 0$.

$$\mathbb{E}\left(\|\frac{X_i - X_j}{p}\|^4\right) = \frac{\mathbb{E}\left(\left(X_i'X_i\right)^2 + \left(X_j'X_j\right)^2 + 4\left(X_i'X_j\right)^2 + 2\left(X_i'X_i\right)\left(X_j'X_j\right)\right)}{p^2}$$

$$\mathbb{E}\left(\frac{X'_iX_i}{p}\right)^2 = \mathbb{E}\left(\frac{\sum_{s=1}^p \sum_{t=1}^p x_{is}^2 x_{it}^2}{p^2}\right) < \frac{\sum_{s=1}^p \sum_{t=1}^p \sqrt{\mathbb{E}x_{is}^4 \mathbb{E}x_{it}^4}}{p^2} < \infty$$

The rest can be proven by a similar way. So,

$$\mathbb{E}\left(\|\frac{X_i - X_j}{p}\|^4\right) < \infty$$

□

According to the above analysis, we can also discuss the power properties of the BF-test. It is clear that BF-test won't work, if $\sigma_1 = \sigma_2$ and $v^2 = 0$, but X_i's and Y_j's from different distributions. That is because, in such case BF-test converge to the same distribution as it does under the null hypothesis. Even when we assume that $v^2 = 0$ and $\sigma_1 \neq \sigma_2$, the BF-test will still perform poorly in the cases where $v^2 \leq |\sigma_1^2 - \sigma_2^2|$, according to Biswas and Ghosh (2014)'s simulation results.

Chapter 2

Behavior of the BG-test as $m, n, p \to \infty$

In this chapter, we are going first to find the main contributors of $\frac{N}{p}T_{mn}^{BG}$ by using the Taylor method, where $N = m + n$. Secondly, we present the asymptotic distribution of $\frac{N}{p}T_{mn}^{BG}$ in high dimensional settings. Last, we can show its power properties by analyzing its Taylor approximation.

2.1 Taylor Approximation of $\frac{N}{p}T_{mn}^{BG}$

$\frac{N}{p}T_{mn}^{BG}$ can be written as below,

$$\frac{N}{p}T_{mn}^{BG} = \frac{1}{2}\left(\left(\frac{\sqrt{N}}{\sqrt{p}}\left(2\hat{\mu}_{FG} - \hat{\mu}_{FF} + \hat{\mu}_{GG}\right)\right)^2 + \left(\sqrt{N}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}})\right)^2\right)$$

According to the analysis from the last chapter, we know that, under the null hypothesis

$$\frac{\sqrt{N}}{\sqrt{p}}\left(2\hat{\mu}_{FG} - \hat{\mu}_{FF} + \hat{\mu}_{GG}\right) = \frac{T_{mn}^{BF*}}{p\sqrt{N}} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

So, we only need deal with

$$\left(\sqrt{N}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}})\right)^2.$$

Now, let's find out the Taylor approximation of $\sqrt{N}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}})$.

$$\begin{aligned}\sqrt{N}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}}) &= \sqrt{N}\left(\sqrt{2}\sigma_1 - \sqrt{2}\sigma_2\right) + \frac{\sqrt{N}}{m\sqrt{2}\sigma_1}\sum_{i=1}^{m}\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \sigma_1^2\right)\\ &- \frac{\sqrt{N}}{n\sqrt{2}\sigma_2}\sum_{i=1}^{n}\left(\frac{(Y_i-\mu_1)^{'}(Y_i-\mu_2)}{p} - \sigma_2^2\right)\\ &- \frac{2\sqrt{N}}{m(m-1)\sqrt{2}\sigma_1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{(X_i-\mu_1)^{'}(X_j-\mu_1)}{p}\\ &+ \frac{2\sqrt{N}}{n(n-1)\sqrt{2}\sigma_2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{(Y_i-\mu_2)^{'}(Y_j-\mu_2)}{p}\end{aligned}$$

We know that

$$\mathbb{V}(\frac{2\sqrt{N}}{m(m-1)\sqrt{2}\sigma_1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{(X_i-\mu_1)^{'}(X_j-\mu_1)}{p}) = \frac{2N}{m(m-1)\sigma_1}\frac{\mathrm{tr}(\Sigma^2)}{p^2} \to 0$$

so

$$\frac{2\sqrt{N}}{m(m-1)\sqrt{2}\sigma_1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{(X_i-\mu_1)^{'}(X_j-\mu_1)}{p} \to_p 0.$$

Thus, the main contributor of $\sqrt{N}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}})$ is, under H_0

$$\frac{\sqrt{N}}{m\sqrt{2}\sigma_1}\sum_{i=1}^{m}\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \sigma_1^2\right) - \frac{\sqrt{N}}{n\sqrt{2}\sigma_2}\sum_{i=1}^{n}\left(\frac{(Y_i-\mu_2)^{'}(Y_i-\mu_2)}{p} - \sigma_2^2\right).$$

2.2 Asymptotic distribution of $\frac{N}{p}T_{mn}^{BG}$

Theorem 2.2.1. *Under the assumptions A(1)-A(3),*

$$\frac{\sum_{i=1}^{m}\left(\frac{(\mathbf{X}_i-\mu_1)^{'}(\mathbf{X}_i-\mu_1)}{p} - \frac{trace\Sigma_1}{p}\right)}{\sqrt{m}\sqrt{\mathbb{V}\left(\frac{(\mathbf{X}_i-\mu_1)^{'}(\mathbf{X}_i-\mu_1)}{p}\right)}} \to_d N(0,1),$$

as $m, p \to \infty$.

Proof. According to the theorem 1.4.1, we know that

$$\mathbb{V}\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p}\right) < \infty \text{ for all } p > 0.$$

And,

$$\langle \left[\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}\right]^2 \rangle$$

is uniformly integrable over p, since

$$\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}\right)^2 \rightarrow_p 0$$

and according to the property of ρ-mixing, as $p \to \infty$

$$\mathbb{E}\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}\right)^2 \rightarrow 0$$

So, the Lindeberge Condition is met, because

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}\right)^2 ; |\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}| > \sqrt{m}\epsilon\right] \\
&= \mathbb{E}\left[\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}\right)^2 ; |\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}| > \sqrt{m}\epsilon\right] \\
&\leq \sup_{p>0}\mathbb{E}\left[\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p}\right)^2 ; (\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \frac{\text{trace}\Sigma_1}{p})^2 > m\epsilon\right] \rightarrow 0
\end{aligned}$$

Thus, the result follows. □

Let's denote

$$S = \frac{\sqrt{N}}{m\sqrt{2}\sigma_1}\sum_{i=1}^{m}\left(\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)}{p} - \sigma_1^2\right) - \frac{\sqrt{N}}{n\sqrt{2}\sigma_2}\sum_{i=1}^{n}\left(\frac{(Y_i-\mu_2)^{'}(Y_i-\mu_2)}{p} - \sigma_2^2\right).$$

Lemma 2.2.2. *Under A(1)- A(3), and assume H_0 is true, as* $\min(m,n), p \to \infty$,

$$\frac{S}{\sqrt{\mathbb{V}(S)}} \to_d N(0,1)$$

Proof. Under H_0, $\sigma = \sigma_1 = \sigma_2$ and $\Sigma = \Sigma_1 = \Sigma_2$. It is easy to check that

$$\begin{aligned} S = & \frac{\sqrt{N}}{m\sqrt{2}\sigma} \sum_{i=1}^{m} \left(\frac{(X_i - \mu_1)^{'} (X_i - \mu_1)}{p} - \frac{\text{trace}\Sigma}{p} \right) \\ & - \frac{\sqrt{N}}{n\sqrt{2}\sigma} \sum_{i=1}^{n} \left(\frac{(Y_i - \mu_2)^{'} (Y_i - \mu_2)}{p} - \frac{\text{trace}\Sigma}{p} \right) \end{aligned}$$

So, according to the theorem 2.2.1, the result follows. □

Based on the above analysis, we can now present the following theorem.

Theorem 2.2.3. *Under A(1)- A(3), and assume H_0 is true,*

$$\frac{2NT_{m,n}^{BG}}{p\mathbb{V}(S)} \to_d \chi_1^2,$$

where $N = m + n$, as $\min(m,n), p \to \infty$, *and* $\min(m,n) = o(p)$.

2.3 A ratio consistent estimator for $\mathbb{V}(\frac{N}{p}T_{mn}^{BG})$

During our carefully studying, we found that it is hard to find a ratio-consistent estimator of

$$\mathbb{V}\left(X_i^{'}X_i\right),$$

which is the main part of $\mathbb{V}(S)$. Thus, we will find an ratio-consistent estimator of

$$\mathbb{V}\left(\sqrt{N}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}})\right)$$

under H_0 in the following way.

Since

$$\sqrt{N}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}}) = \sqrt{N}\binom{m}{2}^{-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{\|X_i - X_j\|}{\sqrt{p}} - \sqrt{N}\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{\|Y_i - Y_j\|}{\sqrt{p}}$$

it is clearly a sum of two U-statistics. Let

$$U_x = \binom{m}{2}^{-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{\|X_i - X_j\|}{\sqrt{p}}$$

and

$$U_y = \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{\|Y_i - Y_j\|}{\sqrt{p}}.$$

We only need to find the limiting distribution of U_x. The limiting distribution of U_y can be derived in the same way.

The H*á*jek projection of U_x can be expressed as

$$\hat{U}_x - \theta = \frac{2}{m}\sum_{i=1}^{m}\tilde{h}_1(X_i),$$

where

$$\theta_1 = \mathbb{E}(\frac{\|X_1 - X_2\|}{\sqrt{p}})$$

$$\tilde{h}_1(X_i) = \mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i) - \theta_1.$$

Theorem 2.3.1. *Follow the assumptions A(1)-A(3),*

$$\sqrt{m}\frac{(\hat{U}_x - \theta_1)}{\delta_1/\sqrt{p}} \rightarrow_d N(0,4),$$

where $\frac{\delta_1^2}{p} = \mathbb{V}(\tilde{h}_1(X_i))$.

Proof. It is easy to check

$$\frac{\delta_1^2}{p} = \mathbb{V}(\tilde{h}_1(X_i)) = \mathbb{E}(\mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i)^2) - \theta^2 \leq \mathbb{E}(\frac{\|X_i - X_j\|^2}{p}) = \frac{2\mathrm{tr}(\Sigma_1)}{p}.$$

According to Theorem 1.4.1, we know that

$$\mathbb{E}(\mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i)^4) \leq \mathbb{E}(\frac{\|X_i - X_j\|^4}{p^2}) < \infty, \text{ for all } p > 0.$$

Thus,

$$\left\langle \mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i)^s \right\rangle$$

is uniformly integrable over p for all $0 < s < 4$ since

$$\sup_{p>0} \mathbb{E}\left(\mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i)^s; \mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i)^s > m\right)$$

$$\leq \frac{1}{m^{4/s-1}} \sup_{p>0} \mathbb{E}\left[\mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i)^4; \mathbb{E}(\frac{\|X_i - X_j\|}{\sqrt{p}}|X_i)^s > m\right] \to 0, \text{ as } m \to \infty.$$

It is clear that $\left\langle \tilde{h}_1^2(X) \right\rangle$ is also uniform integrable over p. Since for all $\epsilon > 0$

$$\begin{aligned}
&\frac{1}{\delta_1^2/p} \sum_{i=1}^{m} \mathbb{E}\left(\frac{\tilde{h}_1^2(X_i)}{m}; |\tilde{h}_1(X_i)| > \sqrt{m}\epsilon\right) \\
&= \frac{1}{\delta_1^2/p} \mathbb{E}\left(\tilde{h}_1^2(X_i); |\tilde{h}_1(X_i)| > \sqrt{m}\epsilon\right) \\
&\leq \frac{1}{\delta_1^2/p} \sup_{p>0} \mathbb{E}\left(\tilde{h}_1^2(X_i); \tilde{h}_1^2(X_i) > m\epsilon\right) \to 0 \text{ as } n \to \infty
\end{aligned}$$

by the Lindeberg - Feller theorem,

$$\sqrt{m}\frac{(\hat{U}_x - \theta_1)}{\delta_1/\sqrt{p}} \to_d N(0, 4)$$

□

In order to show

$$\sqrt{m}\frac{U_x - \theta_1}{\delta_1/\sqrt{p}} \to N(0, 4),$$

we need to show that

$$e = U_x - \hat{U}_x$$

is negligible. In other words, we need to show that

$$\frac{\sqrt{N}e}{\delta_1/\sqrt{p}} \to_p 0.$$

It is clear that $\mathbb{E}(e) = 0$. So, we need show that

$$\mathbb{V}(\frac{\sqrt{N}e}{\delta_1/\sqrt{p}}) \to 0.$$

It can be shown in the following.

$$\begin{aligned}
\mathbb{V}(\frac{\sqrt{N}e}{\delta_1/\sqrt{p}}) &= \frac{N}{\delta_1^2/p}\mathbb{E}(U_x - \theta_1 - \hat{U}_x)^2 \\
&= \frac{N}{\delta_1^2/p}[\mathbb{E}(U_x - \theta_1)^2 + \mathbb{E}\hat{U}_x^2 - 2\mathbb{E}((U_x - \theta_1)\hat{U}_x)] \\
&= \frac{N}{\delta_1^2/p}[\mathbb{E}(U_x - \theta_1)^2 + \mathbb{E}\hat{U}_x^2 - 2\sum_{i=1}^{m}\mathbb{E}\mathbb{E}(U_x h_1(X_i)|X_i)] \\
&= \frac{N}{\delta_1^2/p}[\mathbb{E}(U_x - \theta_1)^2 + \mathbb{E}\hat{U}_x^2 - 2\mathbb{E}\hat{U}_x^2] \\
&= \frac{N}{\delta_1^2/p}[\mathbb{E}(U_x - \theta_1)^2 - \mathbb{E}\hat{U}_x^2] \\
&= \frac{N}{\delta_1^2/p}(\frac{4}{m}\delta_1^2/p + o(n^{-1}) - \frac{4}{m}\delta_1^2/p) \\
&= o(1)
\end{aligned}$$

Now, we can present the following theorem.

Theorem 2.3.2. *Follow the assumptions A(1)-A(3), under H_0, let $\frac{\delta^2}{p} = \mathbb{V}(\tilde{h}_1(X_i)) = \mathbb{V}(\tilde{h}_1(Y_j))$, for all $0 < i, j < m, n$*

$$\frac{NT_{BF}}{2\delta^2\left(N/m + N/n\right)} \to_d \chi_1^2,$$

as $\min(n, m), p \to \infty$ and $\min(m, n) = o(p)$.

Proof. First, let's denote $\lambda = \lim \frac{m}{N}$. Since under H_0, let's assume $\theta = \mathbb{E}(U_x) = \mathbb{E}(U_y)$. From the analysis above, we know that

$$\frac{\sqrt{N}}{2\delta/\sqrt{p}}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}}) = \frac{\sqrt{N}}{\sqrt{m}}\left(\sqrt{m}\frac{U_x - \theta}{2\delta/\sqrt{p}}\right) - \frac{\sqrt{N}}{\sqrt{n}}\left(\sqrt{n}\frac{U_y - \theta}{2\delta/\sqrt{p}}\right)$$

$$\rightarrow_d N(0, \frac{1}{\lambda} + \frac{1}{1-\lambda})$$

So, we have the following

$$\frac{\frac{\sqrt{N}}{2\delta/\sqrt{p}}(\frac{\hat{\mu}_{FF}}{\sqrt{p}} - \frac{\hat{\mu}_{GG}}{\sqrt{p}})}{\sqrt{N/m + N/n}} \rightarrow_d N(0,1)$$

Thus,

$$\frac{NT_{BF}}{2\delta_1^2\,(N/m + N/n)} \rightarrow_d \chi_1^2$$

□

According to the analysis above, we know that Biswas and Ghosh(2014) suggested to estimate $\delta_1^2 = p\mathbb{V}(\tilde{h}_1(X_i))$ and $\delta_2^2 = p\mathbb{V}(\tilde{h}_1(Y_i))$ by using the following two estimators.

$$S_1 = \left[\binom{m}{3}^{-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{k=j+1}^{m}\|X_i - X_j\|\|X_i - X_k\|\right] - \left[\binom{m}{2}^{-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\|X_i - X_j\|\right]^2$$

$$S_2 = \left[\binom{n}{3}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n}\|Y_i - Y_j\|\|Y_i - Y_k\|\right] - \left[\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\|Y_i - Y_j\|\right]^2.$$

According to our simulation, these two estimators often give negative values. Thus, we recommend the estimator proposed by Sen(1960),

$$S_1^* = \frac{1}{m-1}\sum_{i=1}^{m}\left[\widehat{h}_1(X_i) - U_m\right]^2$$

where

$$\widehat{h}_1(X_i) = \frac{1}{m-1}\sum_{j=1}^{m}\|X_i - X_j\|$$

and

$$U_m = \binom{m}{2}^{-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\|X_i - X_j\|.$$

Later, in our simulation studies, we employ the pooled estimator

$$S = \frac{mS_1^* + nS_2^*}{m+n}$$

where

$$S_2^* = \frac{1}{n-1}\sum_{i=1}^{n}\left[\widehat{h}_2(Y_i) - U_n\right]^2$$

$$\widehat{h}_2(Y_i) = \frac{1}{n-1}\sum_{j=1}^{n}\|Y_i - Y_j\|$$

and

$$U_n = \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\|Y_i - Y_j\|.$$

Based on our previous analysis, it is clearly that

$$\frac{S_i^*}{p} \to_p \lim_{p\to\infty}\frac{\sigma_i}{p}, i = 1, 2$$

because

$$\mathbb{V}\left(\frac{S_i^*}{p}\right) \to 0, i = 1, 2.$$

2.4 Power properties of the BG-test

Theorem 2.4.1. *Under the assumptions A(1)-A(3), the power of BG-test of level α tend to 1 if either $\sigma_1 \neq \sigma_2$ or $v^2 \geq 0$.*

Proof. If $\sigma_1 \neq \sigma_2$ or $v^2 \geq 0$, most of terms of the Taylor approximation of

$$\frac{\sqrt{N}}{\sqrt{p}}\left(2\hat{\mu}_{FG} - \hat{\mu}_{FF} + \hat{\mu}_{GG}\right)$$

converges to zero in probability except

$$\sqrt{(m+n)}\left(2\sigma_3 - \sigma_1 - \sigma_2\right) + \sqrt{(m+n)}\frac{1}{2\sigma_3}\left(\frac{\|\mu_1 - \mu_2\|^2}{p} - v^2\right)$$

,

$$\sqrt{N}\left[\left(\frac{1}{m\sigma_3}-\frac{1}{m\sqrt{2}\sigma_1}\right)\sum_{i=1}^{m}\frac{(X_i-\mu_1)^{'}(X_i-\mu_1)-p\sigma_1^2}{p}\right]$$

and,

$$\sqrt{N}\left[\left(\frac{1}{m\sigma_3}-\frac{1}{m\sqrt{2}\sigma_1}\right)\sum_{i=1}^{n}\frac{(Y_i-\mu_2)^{'}(Y_i-\mu_2)-p\sigma_2^2}{p}\right].$$

According to the previous analysis, we know that

$$\left(\frac{\sqrt{N}}{\sqrt{p}}\left(2\hat{\mu}_{FG}-\hat{\mu}_{FF}+\hat{\mu}_{GG}\right)\right)^2=\left[\sqrt{(m+n)}\left(2\sigma_3-\sigma_1-\sigma_2\right)+\sqrt{(m+n)}\frac{1}{2\sigma_3}\left(\frac{\|\mu_1-\mu_2\|^2}{p}-v^2\right)\right]^2$$
$$+O_p(1)$$

And,

$$N(\frac{\hat{\mu}_{FF}}{\sqrt{p}}-\frac{\hat{\mu}_{GG}}{\sqrt{p}})^2=N\left(\sqrt{2}\sigma_1-\sqrt{2}\sigma_2\right)^2+O_p\left(1\right)$$

Thus, $\frac{N}{p}T_{BG}\to_p\infty$. □

From above analysis, it is easy to see if $\sigma_1=\sigma_2$ and $v^2=0$, BG-test does not work because in such case the test would converge to the same distribution as it does under the null hypothesis.

Chapter 3

Simulation Results

In this chapter, we report results from simulation studies designed to evaluate the power of the two test in high dimensional case.

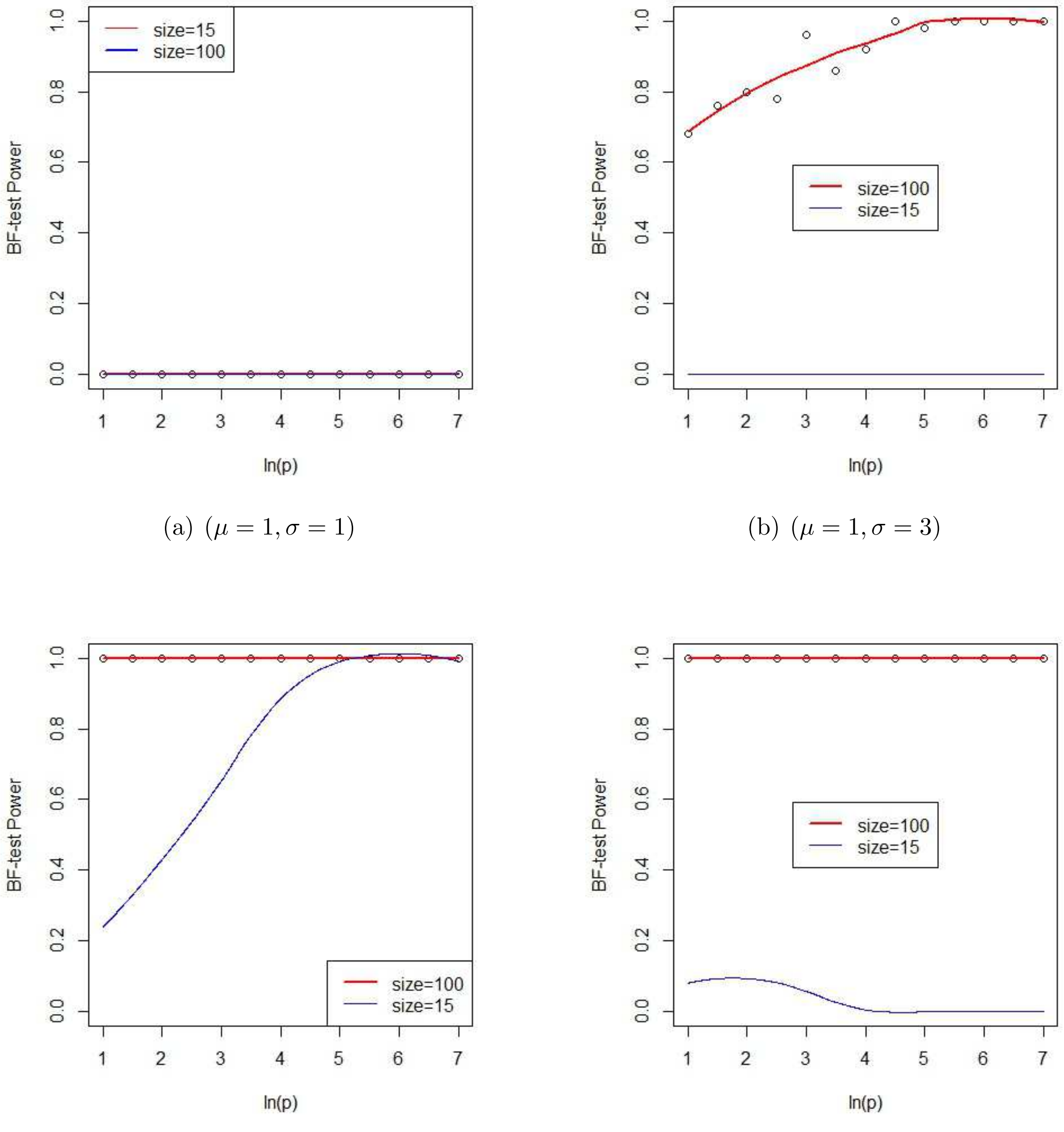
3.1 Simulation of Power curves

First, we will estimate the power of BF-test. We set F distributed as $N_p((\mu,\ldots,\mu)',\sigma I_p)$ where I_p stands for identity matrix. And, G is distributed as $(Exp(1),\ldots,Exp(1))' \in R^p$. We consider four cases, namely $(\mu=1,\sigma=1)$,$(\mu=1,\sigma=3)$,$(\mu=0,\sigma=1)$ and$(\mu=0,\sigma=2)$. In each case, we generate n observations from each distribution to test $H_0 : F = G$. We choose $n = 15$ and 100. And, the experiment is repeated 200 times, and the proportion of times a test rejected H_0 was considered as an estimate of its power. Since

$$\frac{T^{BF*}_{m,n}}{2\sqrt{\frac{\widehat{\mathrm{tr}(\Sigma^2)}}{\widehat{\mathrm{tr}(\Sigma)}}}\sqrt{(\frac{m+n}{m-1})^2+(\frac{m+n}{n-1})^2+(\frac{m+n}{\sqrt{2mn}})^2}}$$

converges to $N(0,1)$, the null hypothesis would be rejected if test is larger than 1.96.

From the plots below, we see that BF-test does not work when $\mu_1=\mu_2$, $\sigma_1=\sigma_2$. In the other cases, BF-test works fine when $\min(m,n), p \to \infty$.



(a) $(\mu = 1, \sigma = 1)$

(b) $(\mu = 1, \sigma = 3)$

(c) $(\mu = 0, \sigma = 1)$

(d) $(\mu = 0, \sigma = 2)$

Figure 3.1: Power curves of BF-test

Next, we would like to estimate the power of BG-test. We would follow the same procedures. Additionally, we would like to estimate the power of BG-test in the case of HDLSS. We also consider the same four cases, namely $(\mu = 1, \sigma = 1)$,$(\mu = 1, \sigma = 3)$,$(\mu = 0, \sigma = 1)$ and$(\mu = 0, \sigma = 2)$. Since BG-test

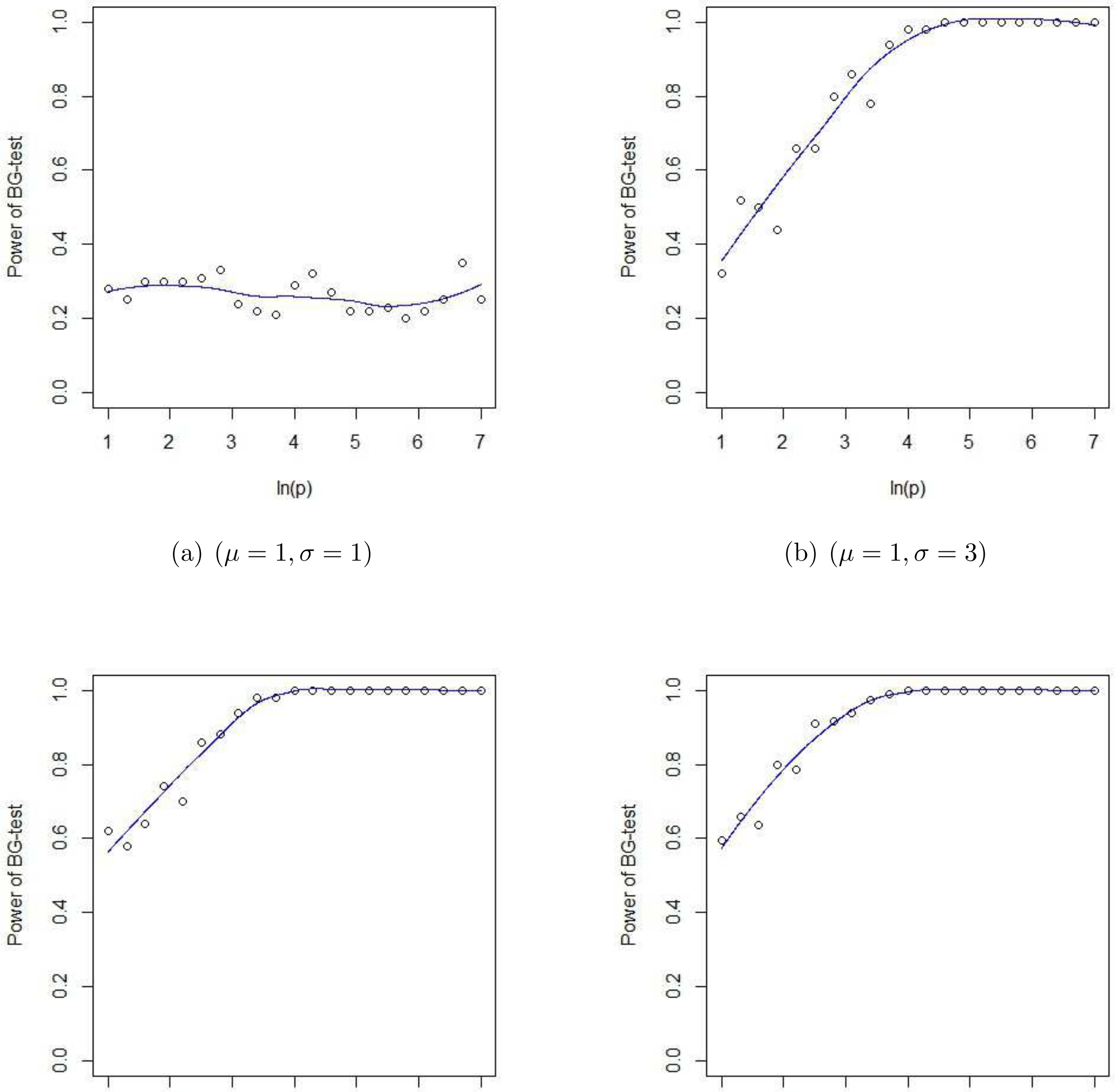
$$\frac{NT_{BF}}{2\hat{\delta}_1^2 \left(N/m + N/n\right)}$$

converges to Chi-square distribution with degree freedom 1, the null hypothesis will be rejected if it is larger than 3.841.

According to the following plots 3.2, it is easy to see that BG-test also does not work when $\mu_1 = \mu_2$, $\sigma_1 = \sigma_2$. In other case, BG-test works very well even when sample size equals 4. Biswas and Ghosh (2014) suggested to use permutation test if sample size is small. It may not be necessary. This problem is beyond the scope of this work but it could be one of my future research topics.

3.2 Conclusions of Simulation

According to our analysis, we can say both tests are mean-variance tests. In other words, they both test $H_0 : \mathbb{E}X = \mathbb{E}Y$ and $\mathbb{V}X = \mathbb{V}Y$ rather than $H_0 : F = G$. It is because two distributions sharing the same mean and variance may be different in many other ways. Thus, further studies are needed to propose a more sophisticated test.



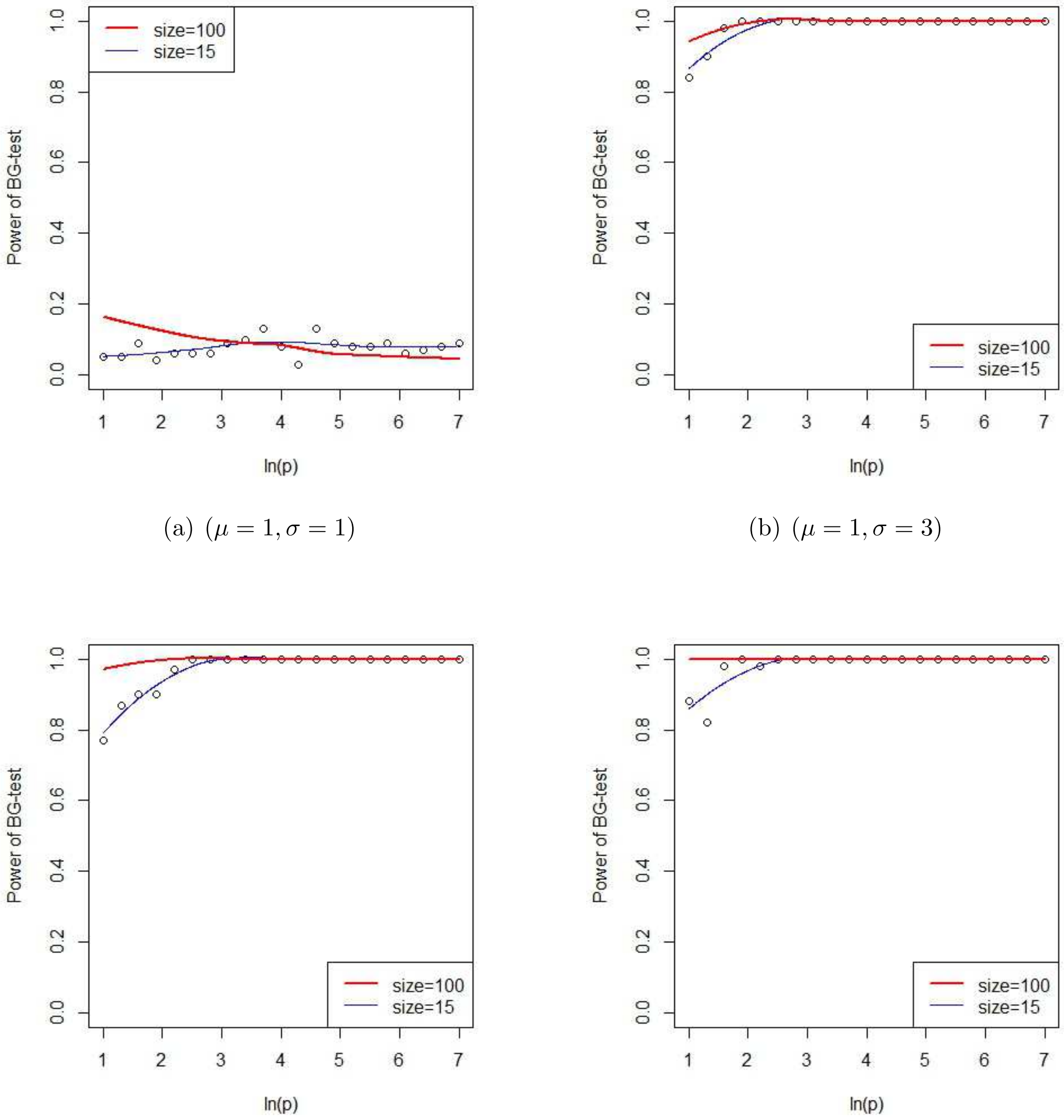
(a) $(\mu = 1, \sigma = 1)$

(b) $(\mu = 1, \sigma = 3)$

(c) $(\mu = 0, \sigma = 1)$

(d) $(\mu = 0, \sigma = 2)$

Figure 3.2: Power curves of BG-test in the case of HDLSS



(a) $(\mu = 1, \sigma = 1)$

(b) $(\mu = 1, \sigma = 3)$

(c) $(\mu = 0, \sigma = 1)$

(d) $(\mu = 0, \sigma = 2)$

Figure 3.3: Power curves of BG-test

REFERENCES

1. Bai and Saranadasa, H. (1996) Effect of high-dimension: by an example of a two-sample problem. *Statistica Sinica*, 311-329.

2. Biswas, M. and Ghosh, A. (2014). A nonparametric two-sample test applicable to high dimensional data. *J. Multivariate Anal.* **123**, 160-171.

3. Chen, Song Xi; Qin, Ying-Li. (2010) A two-sample test for high-dimensional data with applications to gene-set testing. Ann. Statist. 38 (2010), no. 2, 808–835. doi:10.1214/09-AOS716

4. Fan, Jianqing; Fan, Yingying. (2008) High-dimensional classification using features annealed independence rules. Ann. Statist. 36 , no. 6, 2605–2637. doi:10.1214/07-AOS504. http://projecteuclid.org/euclid.aos/1231165181.

5. Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7**, 697-717.

6. Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16**, 772-783.

7. Hall, P., Marron, J. S. and Neeman, A. (2005), Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67: 427-444. doi: 10.1111/j.1467-9868.2005.00510.x

8. Rauf Ahmad (2014) A U-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens-Fisher setting *Annals of the Institute of Statistical Mathematics* 0020-3157 66:1

9. Rauf Ahmad, M., von Rosen, D. and Singull, M. (2013), A note on mean testing for high dimensional multivariate data under non-normality. Statistica Neerlandica, 67: 81-99. doi: 10.1111/j.1467-9574.2012.00533.x

10. Serfling, R. J. (1980). Approximation theorems of mathematical statistics. New York: Wiley

11. Schucany, W. R. and Bankson, D. M. (1989), Small sample variance estimators for U-statistics. Australian Journal of Statistics, 31: 417426. doi: 10.1111/j.1467-842X.1989.tb00986.x

12. Sen, P.K. (1960). On some convergence properties of U-statistics. Cal. Statist. Assoc. Bull. 10 1-18

13. Sungkyu Jung, Arusharka Sen, J.S. Marron, Boundary behavior in High Dimension, Low Sample Size asymptotics of PCA, Journal of Multivariate Analysis, Volume 109, August 2012, Pages 190-203, ISSN 0047-259X.

14. Tao, T. (2012) Topics in Random Matrix Theory, isbn: 9780821885079, American Mathematical Soc.

15. van der Vaart, A.W. (2000). Asymptotic statistics. Cambridge University Press