More powerful two-sample tests for univariate and high-dimensional data

by

Huaiyu Zhang

M.S., Dongbei University of Finance and Economics, China, 2014

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

Comparing the means of two populations is a common task in scientific studies. In this dissertation, we consider more powerful tests for testing the equality of means for univariate and high-dimensional settings. In the univariate case, the classical two-sample t-test is not robust to skewed population, and the large-sample test has low accuracy for finite sample sizes. The first part of this dissertation proposes two new types of tests, the TCFU, and the TT tests, for comparing means with unequal-variance populations. The TCFU test uses Welch's t-statistic as the test statistic and the Cornish-Fisher expansion as its critical values. The TT tests transform Welch's t-statistic and use the normal percentiles as critical values. Four types of monotone transformations are considered for the TT tests. Power and type I error rate comparison of different tests are conducted theoretically and numerically. Analytical conditions are derived to help practitioners choose a powerful test. Two real-data examples are presented to illustrate the application of the new tests.

The second part considers a more challenging situation: testing the equality of two high-dimensional means. When the sample sizes are much smaller than the dimensionality, it is not viable to construct a uniformly most powerful test. Here we propose a new test based on the average squared component-wise t-statistic. Our new test shares some similarity with the generalized component test (GCT) proposed by Gregory et al. (2015), but it differs from the latter test in the following aspects: (i) our new test constructs a different scaling parameter that can be directly estimated from the data instead of from the t-statistics sequence. (ii) it does not require the stationarity condition implicitly assumed in the GCT test; (iii) the new variance estimator guarantees non-negativeness as it is supposed to have; (iv) the test works well even when components of the data vector have high correlations, as long as such correlation reduces suitably fast as the separation of the component indices increases (at least with polynomial rate). The limiting distribution of the test statistic and the power function are derived. The new test is also compared with several other existing tests through Monte Carlo experiments. With acute lymphoblastic leukemia gene expression data, we demonstrated how the new test can be used to give more consistent results in detecting differently expressed Gene Ontology terms than competing tests.

In the last part of the dissertation, we consider power adjustments to address a question of how to fairly compare the power of competing methods in simulation studies when

they have different empirical type I error rates. After discussing some existing methods and their drawbacks, we introduce a new power adjustment method. The new power adjustment method is used to compare the simulation results in the previous two parts of the dissertation.

More powerful two-sample tests for univariate and high-dimensional data

by

Huaiyu Zhang

Dongbei University of Finance and Economics, China, 2014

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Haiyan Wang

Copyright

Abstract

Comparing the means of two populations is a common task in scientific studies. In this dissertation, we consider more powerful tests for testing the equality of means for univariate and high-dimensional settings. In the univariate case, the classical two-sample t-test is not robust to skewed population, and the large-sample test has low accuracy for finite sample sizes. The first part of this dissertation proposes two new types of tests, the TCFU, and the TT tests, for comparing means with unequal-variance populations. The TCFU test uses Welch's t-statistic as the test statistic and the Cornish-Fisher expansion as its critical values. The TT tests transform Welch's t-statistic and use the normal percentiles as critical values. Four types of monotone transformations are considered for the TT tests. Power and type I error rate comparison of different tests are conducted theoretically and numerically. Analytical conditions are derived to help practitioners choose a powerful test. Two real-data examples are presented to illustrate the application of the new tests.

The second part considers a more challenging situation: testing the equality of two high-dimensional means. When the sample sizes are much smaller than the dimensionality, it is not viable to construct a uniformly most powerful test. Here we propose a new test based on the average squared component-wise t-statistic. Our new test shares some similarity with the generalized component test (GCT) proposed by Gregory et al. (2015), but it differs from the latter test in the following aspects: (i) our new test constructs a different scaling parameter that can be directly estimated from the data instead of from the t-statistics sequence. (ii) it does not require the stationarity condition implicitly assumed in the GCT test; (iii) the new variance estimator guarantees non-negativeness as it is supposed to have; (iv) the test works well even when components of the data vector have high correlations, as long as such correlation reduces suitably fast as the separation of the component indices increases (at least with polynomial rate). The limiting distribution of the test statistic and the power function are derived. The new test is also compared with several other existing tests through Monte Carlo experiments. With acute lymphoblastic leukemia gene expression data, we demonstrated how the new test can be used to give more consistent results in detecting differently expressed Gene Ontology terms than competing tests.

In the last part of the dissertation, we consider power adjustments to address a question of how to fairly compare the power of competing methods in simulation studies when

they have different empirical type I error rates. After discussing some existing methods and their drawbacks, we introduce a new power adjustment method. The new power adjustment method is used to compare the simulation results in the previous two parts of the dissertation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Comparing the means of two populations is a common task in scientific studies, which often requires statistical evidence. This research is interested in testing the equality of means for two independent populations. Chapter 2 focuses on the univariate case and Chapter 3 is on the high-dimensional case.

The traditional approach for univariate two-sample mean comparison is the two-sample t-test. If the variances are unequal for the two populations, Welch's t-statistic is frequently used. When the underlying populations follow normal distribution, the Welch's t-statistic follows t-distribution whose degrees of freedom can be approximated by the Satterthwaite method. This distributional result, however, is not robust to the violation of the normality assumption, especially when the two populations have different skewness parameters. A possible solution is to approximate the sampling distribution by normal distribution based on the central limit theorem when the sample size is large. The accuracy of approximation though is still affected by the skewness, kurtosis and other high-order moments. Via the expansion theory, we know that the cumulative distribution function (CDF) of the Welch's t-statistic admits an Edgeworth expansion. The leading term in the expansion is the CDF of the normal distribution, and the second term, of order $O(n^{-1/2})$, is affected by the relative skewness of the two populations. If the sample size is finite and the skewness is large, the effect due to the $O(n^{-1/2})$ term could be significant.

In Chapter 2, we will consider two types of adjustments for the two-sample mean test with unequal-variance populations. The first is to improve the critical values by Cornish-Fisher expansion, i.e., the percentiles of the Edgeworth expansion. This type of tests will be called TCFU tests. We will consider two versions of TCFU tests: one with accurate $O(n^{-1/2})$ term in the Cornish-Fisher expansion and the other one with accurate terms for both the $O(n^{-1/2})$ and the $O(n^{-1})$ orders. The second type, called the TT tests, is to transform the non-normal Welch's t-statistic by some asymmetric functions so that the transformed statistic can achieve a faster converging speed to the normal distribution than the Welch's t-statistic. The transformation essentially reduces the skewness effect from the sampling distribution. The transformed statistic is then compared with the percentiles of normal distribution to make the decision. We consider four transformations in our study. For all the new proposed tests, we will derive their theoretical power function and conduct extensive comparisons on their power and type I error, both theoretically and numerically. Two real-data examples are given in the end to show possible applications of our new tests.

Chapter 3 extends the two-sample mean test to high-dimensional scenario. The demand for analyzing high-dimensional data surges with technology advances. High-throughput data from life sciences, health metric data from electronic devices, user profiles from social media, etc., provide abundant sources of high-dimensional data. Analyzing these high-dimensional data poses challenges to classical statistics. The traditional method for multivariate two-sample test is the Hotelling's T^2 test, whose statistic depends on the inverse of the sample covariance matrix. In high-dimensional case when the dimensionality is larger than the sample size, the covariance matrix is singular. The core of the Hotelling's T^2 is a sample version Mahalanobis distance measuring the separation of the group means. To avoid estimating the inverse of the covariance matrix, other distances were adopted in the literature, for example, Euclidean distance, diagonalize Mahalanobis distance, and Lomonosov distance. Another popular approach in the literature is random projection: the data are projected to a low-dimensional space through a random matrix, then the Hotelling's T^2 is valid to be applied.

We propose a new test for testing the equality of high-dimensional means. The proposed test statistic is based on the average squared univariate t-statistics, denoted by T_n. Gregory et al. (2015) also used T_n to construct their GCT statistic. The scaling parameter in the GCT statistic builds on the autocorrelations of component-wise t^2, but there is no replication available for it. Instead, we directly construct a scaling parameter based on replications in the sample. Simulation study also shows that our new test performs better in controlling the type I error and has more power relative to the GCT test. In a real-data example, our new test and other existing tests were used to test the Gene Ontology terms for different phenotypes. The new test shows good control in type I error and more statistical power. Notably, on different datasets, our test can provide good consistency in identifying important Gene Ontology terms.

The rest of the dissertation is organized as follows: Section 2.1 provides the background and introduction to our new univariate two-sample tests; Section 2.2 and 2.3 present our new TCFU and TT tests, derive the theoretical power functions, and conduct extensive comparison for their power and type I error; Section 2.4 carries out the Monte Carlo experiments and justifies the theoretical results; Section 2.5 gives real-data examples of the new tests; Section 2.6 discusses the potential use of the new tests in high-dimensional variable screening; Section 2.7 summarizes the chapter with some final remarks. Section 3.1 gives the motivation of the new high-dimensional two-sample mean test; Section 3.2 reviews the existing studies relevant to the two-sample mean test in high dimension; 3.3 introduces our new scaling parameter and the new test statistics; Section 3.4 presents theoretical results including the sampling distribution for the new test under both the null and alternative hypotheses; Section 3.5 compares type I error rates and power of our test and several other tests through Monte Carlo experiments; Section 3.6 applies the new test to detecting differentially expressed Gene Ontology terms for acute lymphoblastic leukemia; Section 3.7 concludes the chapter. Chapter 4 presents a new power adjustment method for fairly comparing competing tests when their type I error rates differ. The new adjustment method is applied to the power comparisons in Chapter 2 and 3. Most of the proofs, some notations, and regularity conditions are deferred to the appendices.

Chapter 2

New two-sample tests with heterogeneous variance and their theoretical power

2.1 Introduction

Testing the equality of two population means is a commonly used technique in many scientific studies. Suppose $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ are two simple random samples drawn from two populations $\mathcal{P}_1$ and $\mathcal{P}_2$ with mean μ_1, μ_2 and variance σ_1^2, σ_2^2, respectively. The hypotheses of the two-sample mean testing problem are $H_0 : \mu_1 - \mu_2 = \mu_{10} - \mu_{20}$, and $H_a : \mu_1 - \mu_2 = \mu_{10} - \mu_{20} + \delta_n$, where μ_{10} and μ_{20} are the population means under H_0. The δ_n is the departure from the H_0. When $\sigma_1^2 \neq \sigma_2^2$, a classic method is the Welch's (or unequal variances) two-sample t-test, of which the test statistic is defined as

$$T_n = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_{10} - \mu_{20})}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \tag{2.1}$$

where $\bar{X}_i$ is the sample mean and S_i^2 is the sample variance of the i^{th} sample for $i = 1, 2$. If $\mathcal{P}_1$ and $\mathcal{P}_2$ follow normal distribution, T_n follows t-distribution. When the normality assumption does not hold but the sample sizes are large, T_n approximately follows normal distribution by applying the Central Limit Theorem (CLT). When the sample sizes are small and the data deviate from normality, neither the t-distribution nor the normal approximation can provide a reliable distribution of T_n.

In asymptotic theory, the Edgeworth expansion technique provides higher-order approximation for the cumulative distribution function (CDF) of some statistic. It is usually more accurate than CLT in that it takes into account of high-order moments information such as skewness and kurtosis. Xu et al. (2009) gives the second-order Edgeworth expansion of T_n defined in (2.1) under H_0 as

$$P_{H_0}(T_n \leq x) = \Phi(x) + n^{-1/2} p_{10}(x)\phi(x) + n^{-1} p_{20}(x)\phi(x) + O(n^{-3/2}), \tag{2.2}$$

where $n = n_1 + n_2$ is the total sample size, $\phi(x)$ and $\Phi(x)$ are the probability density function (PDF) and CDF of the standard normal distribution, respectively, and $p_{10}(x)$ and $p_{20}(x)$ are polynomials whose coefficients depend on first four moments of the data. The CLT only keeps $\Phi(x)$ as the approximation of the CDF of T_n and omits the remaining high-order terms. In contrast, the Edgeworth expansion in (2.2) can capture more complicated behaviors of the limiting distribution by taking into account of some skewness and kurtosis in $p_1(x)$ and $p_2(x)$. The influence of $p_1(x)$ and $p_2(x)$ diminish along with increasing sample size, but when the sample size is moderate or small, those terms can improve the approximation to the true CDF. The second-order expansion in (2.2) refines the first-order approximation given by Zhou and Dinh (2005). In particular, the $n^{-1/2}$ term corrects the skewness at the first order, and the n^{-1} term corrects kurtosis at the first order and skewness at the second order. The correction for kurtosis is not trivial when handling heavy-tailed data, for example, in computer science (Gong et al., 2001; Psounis et al., 2005). Although higher-order correction is attractive in theory, the application of the approximation is limited by the accuracy of moment estimators.

Under H_0, if T_n admits the Edgeworth expansion in (2.2), then the $100\alpha^{th}$ percentile of T_n has the following Cornish-Fisher expansion

$$\xi_\alpha = z_\alpha + n^{-1/2} q_{10}(z_\alpha) + n^{-1} q_{20}(z_\alpha) + O(n^{-3/2}) \tag{2.3}$$

where z_α is the $100\alpha^{th}$ percentile, $q_{10}(.)$ and $q_{20}(.)$ are some polynomials whose details will be given in Section 2.2.

Once the Edgeworth expansion for the two-sample t-statistic is derived, the expansion for the percentiles, namely the Cornish-Fisher expansion, follows directly. Since the Edgeworth expansion is more accurate, it is natural to use the Cornish-Fisher expansion as the critical value for testing. Tong (2016) proposed first-order Cornish-Fisher expansion based tests for both T_n and the equal variance (pooled) two-sample t-statistic

$$\tilde{T}_n = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_{10} - \mu_{20})}{\sqrt{(1/n_1 + 1/n_2)[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)}}. \tag{2.4}$$

He derived their theoretical type I errors and power functions and carried out comparisons at the order of $O(n^{-1/2})$. Wang et al. (2017) proposed a second-order Cornish-Fisher expansion based test for (2.4) and improved the type I error and power function accuracy up to the order $O(n^{-1})$. In Section 2.2, we propose a second-order Cornish-Fisher expansion based two-sample test for (2.1) and derive its theoretical power function to the accuracy of $O(n^{-1})$. Xu et al. (2009) and Xu (2010) derived the second-order Edgeworth expansion for T_n under H_0 and H_a, respectively. We will relate their results in our later discussion.

Another related direction of literature is to use transformations to handle skewed data problem. A skewed random variable Y can be thought of as the result of imposing an asymmetric transformation, say $G(.)$, on a random variable Z following normal distribution, i.e., $Y = G(Z)$. Applying the inverse transformation $Z = G^{-1}(Y)$, one can recover Z and use normal distribution for further inference, and hence the effect of skewness is reduced. The form of the transformations and their theoretical properties are often motivated by the

Edgeworth expansions. For example, Hall (1992a) proposed two monotone transformations to refine confidence intervals for skewed population. Zhou and Dinh (2005) proposed a simpler transformation for the one- and two-sample confidence intervals based on t-statistics. Their theoretical results have errors of order $O(n^{-1})$ and only apply to the null hypotheses. In Section 2.3, we will use the second-order Edgeworth expansion to define four tests based on transformations. The transformations include the two functions proposed by Hall (1992b), one proposed by Zhou and Dinh (2005), and one we propose. We apply these transformations to two-sample test with T_n being the test statistic, and compare their type I error rates and power functions theoretically and numerically. Tong (2016) also studied these transformations for testing with $\tilde{T}_n$ and examined their properties based on the **first-order** Edgeworth expansion (accurate to the order of $O(n^{-1/2})$. One of our goals is to give theoretical conditions for more powerful tests. We compare the theoretical power and type I error rates for the tests, but some of them only differ at the $O(n^{-1})$ terms. The first-order expansion provided by Tong (2016) is not sufficient, so we develop the **second-order** Edgeworth expansion in which the $O(n^{-1})$ terms are given in detail.

This chapter has the following contributions:

(i) We extend the second-order Cornish-Fisher expansion based test to unequal variances case for test statistic T_n. We use the second-order Edgeworth expansion to quantify its theoretical power function and compare it with the first-order Cornish-Fisher expansion based test proposed by Tong (2016). As suggested by Wang et al. (2017), the Cornish-Fisher expansion based tests are asymptotically equivalent to the Bootstrap-t tests. It provides an approach to studying the theoretical power for the Bootstrap-t tests. The Cornish-Fisher expansion based tests have comparable performance as their bootstrap counterparts but save computational time considerably. The computational efficiency is especially important for testing a large number of hypotheses, for example, in gene expression data.

(ii) Xu (2010) studied the second-order Edgeworth expansion of T_n, but the critical value is not clearly specified for the testing procedure, and hence he did not give a real power function. Moreover, his result only allows for computing the power at some fixed value, but in our Cornish-Fisher expansion based test, the critical value is a random variable because the Cornish-Fisher expansion needs to be estimated in practice. We explicitly define the transformation tests and derive their theoretical power functions to take care of the issue of random critical values.

(iii) Zhou and Dinh (2005) compared three transformations by Monte Carlo experiments, but we provide a more rigorous comparison in theory. Moreover, as shown in Section 3, the Cornish-Fisher expansion based tests and two of the transformation tests are equivalent in that they both have an error of order $O(n^{-1})$. We use the second-order expansion to study their difference. Our results provide practical guidelines for choosing the transformation and benefit the power analysis and sample size calculation.

(iv) The T_2-transformation proposed by Hall (1992b) will lead to an approximately normally distributed statistic, but it cannot achieve the claimed accuracy of $O(n^{-1})$. We modify it and propose a new transformation G_4, which improves the normal approximation with error order $O(n^{-3/2})$.

(iv) Hall (1992a) and Zhou and Dinh (2005) studied the transformations in confidence interval framework, which only considered the properties of the statistic under H_0. Our extension to hypotheses testing makes it possible to establish the properties under H_a.

The rest of the paper is organized as follows: Section 2.2 proposes a second-order Cornish-Fisher expansion based test for T_n; Section 2.3 considers four transformations on T_n and uses normal percentiles as the critical values. Theoretical power functions are given and used to compare the powers and type I error rates among the transformation tests and Cornish-Fisher expansion based tests. The results of simulation studies and real-data applications are reported in Sections 2.4 and 2.5. Section 2.6 discusses the potential use of the tests in high-dimensional variable screening. Section 2.7 concludes the paper with some final remark. Technical proofs are deferred to the appendices.

2.2 A new test based on second-order Cornish-Fisher expansion and its power function

The classical two-sample t-test with the test statistic T_n, defined in (2.1), requires the underlying population to be normally distributed or the sample sizes to be large. Skewed or heavy-tailed data with finite sample sizes may cause poor performance. The Edgeworth expansion more accurately approximates the distribution by taking high-order moments into account, for example, skewness and kurtosis. The corresponding asymptotic expansion of the percentiles, namely the Cornish-Fisher expansion, is a potential way of improving the critical value. Following this idea, Wang et al. (2017) proposed a Cornish-Fisher expansion based (TCF) test for the **equal-variance** scenario. In this section, we will establish a Cornish-Fisher expansion based testing procedure for the **unequal-variance** scenario and study its theoretical properties.

Xu (2010) derived the second-order Edgeworth expansion for T_n under H_a. We directly cite the result to facilitate our further derivation. Assume $\lambda_i = n_i/n = O(1)$ for $i = 1, 2$, i.e., neither sample size of the two groups would be dominating when the total sample size grows. Suppose the i^{th} population has skewness $\gamma_i = E[(X_{i,j} - \mu_i)^3]/\sigma_i^3$ and kurtosis $\tau_i = E[(X_{i,j} - \mu_i)^4]/\sigma_i^4$, $i = 1, 2$. Also assume Pitman alternative where $\delta_n = \mu_1 - \mu_2 - (\mu_{10} - \mu_{20}) = O(n^{-1/2})$. Suppose the regularity conditions in Appendix A.2 hold, the second-order Edgeworth expansion of the cumulative density function (CDF) of T_n under H_a can be expressed as

$$F_{T,w}(x) = \Phi(x - w) + n^{-1/2}p_1(x - w)\phi(x - w) + n^{-1}p_2(x - w)\phi(x - w) + O(n^{-3/2}) \quad (2.5)$$

where $\phi(x)$ and $\Phi(x)$ are respectively the PDF and CDF of the standard normal distribution, $w = \delta_n(\sigma_1^2/n_1 + \sigma_2^2/n_2)^{-1/2}$, $p_1(x)$ and $p_2(x)$ are polynomials whose coefficients depend on the first four moments of X_{11} and X_{21}. The details of the forms of $p_1(x)$ and $p_2(x)$ are deferred to Appendix A.1 for brevity. Letting $\delta_n = 0$, we naturally get $w = 0$ and the second-order Edgeworth expansion under H_0:

$$F_{T,w=0}(x) = \Phi(x) + n^{-1/2}p_{10}(x)\phi(x) + n^{-1}p_{20}(x)\phi(x) + O(n^{-3/2}),$$

where $p_{10}(x)$ and $p_{20}(x)$ are given in Appendix A.1. This result coincides with the one obtained by Xu et al. (2009).

Based on the Edgeworth expansion given by equation (2.5), the Cornish-Fisher expansion follows directly from Hall (1992a) as given below:

Proposition 1. *Under regularity conditions in Appendix A.2, the $(100\alpha)^{th}$ percentile ζ_α of the distribution $F_{T,w}(x)$ in (2.5) admits the following Cornish-Fisher expansion*

$$\zeta_\alpha = z_\alpha + w + n^{-1/2}q_1(z_\alpha) + n^{-1}q_2(z_\alpha) + O(n^{-3/2})$$

uniformly in $\epsilon < \alpha < 1 - \epsilon$ for each $\epsilon > 0$, where z_α is the $(100\alpha)^{th}$ percentile of standard normal distribution, $q_1(x) = -p_1(x)$, and $q_2(x) = p_1(x)p_1'(x) - xp_1(x)^2/2 - p_2(x)$.

In particular, under H_0, the Cornish-Fisher expansion is

$$\xi_\alpha = z_\alpha + n^{-1/2}q_{10}(z_\alpha) + n^{-1}q_{20}(z_\alpha) + O(n^{-3/2})$$

where $q_{10}(x) = -p_{10}(x)$ and $q_{20}(x) = p_{10}(x)p_{10}'(x) - xp_{10}(x)^2/2 - p_{20}(x)$.

We can approximate ξ_α at the first order by

$$\xi_{1,\alpha} = z_\alpha + n^{-1/2}q_{10}(z_\alpha), \tag{2.6}$$

or at the second order

$$\xi_{2,\alpha} = z_\alpha + n^{-1/2} q_{10}(z_\alpha) + n^{-1} q_{20}(z_\alpha), \tag{2.7}$$

so that $\xi_\alpha = \xi_{1,\alpha} + O(n^{-1})$ and $\xi_\alpha = \xi_{2,\alpha} + O(n^{-3/2})$. In our new test, we will use the approximated Cornish-Fisher expansion as the critical values, but the population moments involved in $q_{10}(.)$ and $q_{20}(.)$ are unknown when working with real data. A natural strategy is to substitute the population moments by their sample estimates. Consider the following estimators:

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2, \ \hat{\gamma}_i = \frac{n_i}{S_i^3 (n_i - 1)(n_i - 2)}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^3 \tag{2.8}$$

$$\text{and } \hat{\tau}_i = \frac{n_i \sum_{i=1}^{n_i}(X_{ij} - \bar{X}_i)^4}{[\sum_{i=1}^{n_i}(X_{ij} - \bar{X}_i)^2]^2} + \frac{6}{(n_i + 1)}, \quad i = 1, 2. \tag{2.9}$$

That is, the sample variance estimates the population variance. The skewness estimator slightly adjusts the sample skewness so that the term $\hat{\gamma}_i S_i^3$ is an unbiased estimate for $E(X_{ij} - \mu_i)^3$. One can show that $\hat{\gamma}_i - \gamma_i = O_p(n_i^{-1/2})$ and $E(\hat{\gamma}_i) = \gamma_i + O(n_i^{-1})$. The kurtosis estimator $\hat{\tau}$ was given by An and Ahmed (2008) and was shown to perform better than several other estimators for skewed or heavy-tailed data in their simulation studies. Replacing the population moments in q_{10} and q_{20} by above estimators leads to the following estimated version of first- and second-order Cornish-Fisher expansions:

$$\hat{\xi}_{1,\alpha} = z_\alpha + n^{-1/2} \hat{q}_{10}(z_\alpha), \tag{2.10}$$

and

$$\hat{\xi}_{2,\alpha} = z_\alpha + n^{-1/2} \hat{q}_{10}(z_\alpha) + n^{-1} \hat{q}_{20}(z_\alpha). \tag{2.11}$$

Although $\hat{\xi}_{2,\alpha}$ is a higher-order approximation than $\hat{\xi}_{1,\alpha}$, its accuracy is restricted by the estimation of the population parameters. Indeed, because $\hat{\xi}_{1,\alpha} - \xi_{1,\alpha} = O_p(n^{-1})$, and $n^{-1}\left[\hat{q}_{20}(z_\alpha) - q_{20}(z_\alpha)\right] = O_p(n^{-3/2})$, it follows that $\hat{\xi}_{2,\alpha} - \xi_{2,\alpha} = O_p(n^{-1})$. Both $\xi_{1,\alpha}$ and

$\xi_{2,\alpha}$ approximate the true percentile ξ_α, and $\xi_{2,\alpha} - \xi_{1,\alpha} = O(n^{-1})$. Therefore, both $\hat{\xi}_{1,\alpha}$ and $\hat{\xi}_{2,\alpha}$ estimate ξ_α with error of order $O_p(n^{-1})$.

Next, we will state the testing procedure of the Cornish-Fisher expansion based two-sample test in unequal-variance setting (called TCFU in the following). The test statistic T_n is defined as (2.1). The TCFU$_i$ ($i = 1$ or 2) test rejects H_0 if

(i) $T_n \leq \hat{\xi}_{i,\alpha/2}$ or $T_n \geq \hat{\xi}_{i,1-\alpha/2}$ for the two-sided alternative hypothesis $H_a : \mu_1 - \mu_2 \neq \mu_{10} - \mu_{20}$;

(ii) $T \geq \hat{\xi}_{i,1-\alpha}$ for the upper-tailed alternative hypothesis $H_a : \mu_1 - \mu_2 > \mu_{10} - \mu_{20}$;

(iii) $T \leq \hat{\xi}_{i,\alpha}$ for the lower-tailed alternative hypothesis $H_a : \mu_1 - \mu_2 < \mu_{10} - \mu_{20}$.

2.2.1 Power function for the TCFU test

The power function is an important aspect of evaluating the performance of a test. When using t-distribution or large-sample normal approximation, the power function could be computed directly through the CDF in equation (2.5) because the critical values are fixed percentiles of t distribution or normal distribution. This is the case for the power comparison conducted by Xu (2010). But for the TCFU tests, the critical values $\hat{\xi}_{1,\alpha}$ or $\hat{\xi}_{2,\alpha}$ are random variables. We cannot directly plug these random values in (2.5), as implied in Xu (2010), to compute the powers. The difficulty of deriving the power functions lies in dealing with the extra uncertainty brought by the estimators. In the following theorem, we account for this extra uncertainty by including extra terms and present the power function of the lower-tailed TCFU$_i$ test, i.e., $P(T \leq \hat{\xi}_{i,\alpha})$. For the upper-tailed and the two-sided tests, the power function can be obtained by computing $1 - P(T \leq \hat{\xi}_{i,1-\alpha})$ and $P(T \leq \hat{\xi}_{i,\alpha/2}) + P(T \geq \hat{\xi}_{i,1-\alpha/2})$, respectively. The proof does not recalculate the Edgeworth expansion but instead follow the approach from Section 3.5.2 of Hall (1992a), which applies the Delta method and focuses on the changes in cumulants. More details of the proof are deferred to Appendix A.3.1.

Theorem 1. *Under regularity conditions in Appendix A.2, the power function for the lower-tailed TCFU$_i$ test is*

$$P(T \leq \hat{\xi}_{i,\alpha}) = F_{T,w}(\xi_{i,\alpha}) - n^{-1}c_\alpha(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2}),\ i = 1, 2,$$

where $F_{T,w}(\cdot)$ is given in (2.5), $\xi_{i,\alpha}$ and $\hat{\xi}_{i,\alpha}$ are given by (2.6)-(2.11), and $c_\alpha = (B/6 - A^2/4)(2z_\alpha^2 + 1)$, with

$$A = \left(\frac{\sigma_1^2}{\lambda_1} + \frac{\sigma_2^2}{\lambda_2}\right)^{-3/2}\left(\frac{\sigma_1^3\gamma_1}{\lambda_1^2} - \frac{\sigma_2^3\gamma_2}{\lambda_2^2}\right),\ \textit{and}\ B = \left(\frac{\sigma_1^2}{\lambda_1} + \frac{\sigma_2^2}{\lambda_2}\right)^{-2}\left(\frac{\sigma_1^4(\tau_1 - 3)}{\lambda_1^3} + \frac{\sigma_2^4(\tau_2 - 3)}{\lambda_2^3}\right).$$

Remark We can interpret the coefficient A as the relative skewness between the two populations and B as the pooled kurtosis. The term $n^{-1}c_\alpha(\xi_{i,\alpha} - w)\phi(\xi_{i,\alpha} - w)$ is due to the extra uncertainty introduced by the estimation of $\xi_{i,\alpha}$. If we use the first-order Edgeworth expansion, this term will go to the remainder $O(n^{-1})$. This power function allows for random critical values. The same technique used here will also be applied to computing the powers of transformation tests later. To compute the sample size for a predetermined power, one can set this power function equal to the desired power and solve the equation.

To determine which test is more powerful, we can compare the power difference between TCFU$_1$ and TCFU$_2$ tests as shown in the following corollary. The proof is deferred to Appendix A.3.2. Let "PD$_{test_1,test_2,H_a}$" denote the power of test 1 minus that of test 2 under H_a. For example, the power difference between lower-tailed TCFU$_2$ and TCFU$_1$ test is denoted by

$$\text{PD}_{TCFU_2,TCFU_1,lower} = P(T \leq \hat{\xi}_{2,\alpha}) - P(T \leq \hat{\xi}_{1,\alpha}).$$

Corollary 1. *The power difference for the lower-tailed TCFU$_2$ and TCFU$_1$ tests at significance level α is*

$$PD_{TCFU_2,TCFU_1,lower} = n^{-1}q_{20}(z_\alpha)\phi(z_\alpha - w) + O(n^{-3/2}) \tag{2.12}$$

The two lower-tailed TCFU tests have powers that only differ at the $O(n^{-1})$ order. The sign of the power difference only depends on $q_{20}(z_\alpha)$ since the normal density function evaluated at $z_\alpha - w$ is always nonnegative. Note that

$$q_{20}(x) = (-x^5/9 + x^3/18 + 83x/72)A^2 + (x^3/12 - x/4)B_1 + (-x^3/2 + x/2)B_2, \qquad (2.13)$$

where

$$B_1 = \left(\frac{\sigma_1^2}{\lambda_1} + \frac{\sigma_2^2}{\lambda_2}\right)^{-2} \left(\frac{\sigma_1^4\tau_1}{\lambda_1^3} + \frac{\sigma_2^4\tau_2}{\lambda_2^3}\right),\ B_2 = \left(\frac{\sigma_1^2}{\lambda_1} + \frac{\sigma_2^2}{\lambda_2}\right)^{-2} \left(\frac{\sigma_1^4}{\lambda_1^3} + \frac{\sigma_2^4}{\lambda_2^3}\right).$$

If $q_{20}(z_\alpha) > 0$, the lower-tailed TCFU$_2$ will be more powerful than the TCFU$_1$ test. Using the fact that $q_{20}(x)$ is an odd function of x, we have the power difference for the upper-tailed alternative as

$$PD_{TCFU_2,TCFU_1,upper} = n^{-1}q_{20}(z_\alpha)\phi(z_{1-\alpha} - w) + O(n^{-3/2}),$$

and the two-sided alternative as

$$PD_{TCFU_2,TCFU_1,two-sided} = n^{-1}q_{20}(z_{\alpha/2})[\phi(z_{\alpha/2} - w) + \phi(z_{1-\alpha/2} - w)] + O(n^{-3/2}).$$

Hence, the condition for the TCFU$_2$ test being more powerful than TCFU$_1$ is still $q_{20}(z_\alpha) > 0$ for the upper-tailed H_a and $q_{20}(z_{\alpha/2}) > 0$ for the two-sided H_a.

It may be noted that the power comparisons above are also valid under H_0 in which $w = 0$. But $w = 0$ will not affect the sign of power difference as well as $q_{20}(.)$. Therefore, the condition that guarantees more power would also make the type I error rate inflate. For example, letting "T1ED" denote the difference between type I errors for two tests, we have

$$\text{T1ED}_{TCFU_2,TCFU_1,lower} = n^{-1}q_{20}(z_\alpha)\phi(z_\alpha) + O(n^{-3/2}).$$

It is clear that $q_{20}(z_\alpha) > 0$ leads to higher power for TCFU_2 test relative to TCFU_1, but its type I error rate is also higher than that of the latter. This is a common trade-off for most of the testing procedures. Practitioners need to balance them to choose a powerful test with a controlled type I error rate.

In the next section, we will establish the theoretical power for transformation based tests. The power comparison can be extended to a wider range.

2.3 Transformation tests

Another popular approach to handling the skewed population is through transformations. If the skewness is induced by some asymmetric transformation on a symmetric random variable, applying a proper inverse transformation to the skewed statistic will remove the skewness and recover symmetry. In this section, we will construct two-sample tests based on the transformations discussed by Zhou and Dinh (2005), and introduce a new transformation. We will also investigate their theoretical power and type I error rates and identify conditions leading to high power.

Let $U = T_n/\sqrt{n}$. We consider the following transformations:

$$\begin{aligned}
G_1(U) &= U + \hat{A}U^2/3 + \hat{A}^2U^3/27 + \hat{A}/(6n) \\
G_2(U) &= (2/3n^{-1/2}\hat{A})^{-1}[\exp(2/3n^{-1/2}\hat{A}U) - 1] + \hat{A}/(6n) \\
G_3(U) &= U + U^2 + U^3/3 + \hat{A}/(6n) \\
G_4(U) &= (2/3\hat{A})^{-1}[\exp(2/3\hat{A}U) - 1] + \hat{A}/(6n)
\end{aligned} \tag{2.14}$$

where $\hat{A}$ is obtained by substituting the population moments in A defined in Theorem 1 by their sample estimates. These functions are the inverse transformations mentioned above. They can help us to convert the skewed T_n to symmetric statistics.

The transformations $G_1(.)$ and $G_2(.)$ were given by Hall (1992b) to remove the skewness for confidence intervals. Zhou and Dinh (2005) added $G_3(.)$ as a simpler alternative to

$G_1(.)$. Via Monte Carlo experiments, they showed that G_3 outperforms G_2, and that it has competitive performance as G_1 and Bootstrap-t in terms of coverage probabilities and interval length of confidence intervals. The new transformation G_4 was motivated by G_2 because the normal approximation to G_2 cannot achieve the accuracy claimed by Hall (1992b). We will explain this more rigorously later.

Both Hall (1992b) and Zhou and Dinh (2005) focused on confidence intervals. Adapting the transformations to hypotheses testing framework can provide more insight into the properties under the alternative hypothesis, for example, power analysis and sample size calculation. Xu et al. (2009) proposed a test procedure based on the transformation that coincides with G_1 and derived the second-order Edgeworth expansion for the transformation under H_0. Xu (2010) further derived the second-order expansion under H_a and computed the power function. He considered three testing strategies: (i) use T_n as test statistic and normal distribution as the null distribution; (ii) use T_n as test statistic and the second-order Edgeworth expansion as the null distribution; (iii) use $\sqrt{n}G_1(T_n/\sqrt{n})$ as the test statistic and normal distribution as the null distribution. A problem with his power comparison is that all the tests need to use the common critical value. The null distributions are different for (ii) and (iii), so it is not reasonable to assume the same critical value.

We explicitly define the transformation tests (called TT tests below) and compare their theoretical power functions. We also extend the comparison to TCFU tests. In some scenarios, the terms of order $O(n^{-1/2})$ vanish in the power differences, so we need to give explicit forms of order $O(n^{-1})$. A complexity associated with it is that the parameter estimation will also affect the $O(n^{-1})$ terms. Therefore, we need to take this extra source of uncertainty into account.

Suppose the null hypothesis $H_0 : \mu_1 - \mu_2 = \mu_{10} - \mu_{20}$ is true. Then T_n and $\sqrt{n}G_i(T_n/\sqrt{n})$ (i = 1, 2, 3, 4) are approximately normal (Hall, 1992b; Zhou and Dinh, 2005). The percentiles of normal distribution can be used as critical values. If one takes T_n as the test statistic, it reduces to the large-sample t test.

The TT$_i$ test ($i = 1, 2, 3, 4$) uses $\sqrt{n}G_i(T_n/\sqrt{n})$ as the test statistic and rejects H_0 if

(i) $\sqrt{n}G_i(T_n/\sqrt{n}) < z_{\alpha/2}$ or $\sqrt{n}G_i(T_n/\sqrt{n}) > z_{1-\alpha/2}$ for the two-sided alternative hypothesis $H_a : \mu_1 - \mu_2 \neq \mu_{10} - \mu_{20}$;

(ii) $\sqrt{n}G_i(T_n/\sqrt{n}) > z_{1-\alpha}$ for the upper-tailed alternative hypothesis $H_a : \mu_1 - \mu_2 > \mu_{10} - \mu_{20}$;

(iii) $\sqrt{n}G_i(T_n/\sqrt{n}) < z_{\alpha}$ for the lower-tailed alternative hypothesis $H_a : \mu_1 - \mu_2 < \mu_{10} - \mu_{20}$.

Although the transformed statistics are all approximately normal, their convergence rates are not all identical. It was stated in Hall (1992b) that with $U = T_n/\sqrt{n}$

$$P_{H_0}(\sqrt{n}G_i(U) \leq x) = \Phi(x) + O(n^{-1}),\ i = 1, 2.$$

However, close examination reveals that

$$P_{H_0}(\sqrt{n}G_2(U) \leq x) = \Phi(x) + O(n^{-1/2}). \tag{2.15}$$

If we instead use our modified version G_4, we can show that

$$P_{H_0}(\sqrt{n}G_4(U) \leq x) = \Phi(x) + O(n^{-1}). \tag{2.16}$$

Furthermore,

$$\begin{aligned} P_{H_0}(\sqrt{n}G_3(U) \leq x) &= P_{H_0}(T_n + \frac{T_n^2}{\sqrt{n}} + \frac{T_n^3}{3n} + \frac{\hat{A}}{6\sqrt{n}} \leq x) \\ &= P_{H_0}(T_n + \frac{\hat{A}}{6\sqrt{n}}(2T_n^2 + 1) + O_p(n^{-1/2}) \leq x) \\ &= \Phi(x) + O(n^{-1/2}). \end{aligned} \tag{2.17}$$

The last line follows from $P_{H_0}(T_n + \hat{A}(2T_n^2 + 1)/(6\sqrt{n}) \leq x) = \Phi(x) + o(n^{-1/2})$ (Abramovitch and Singh, 1985, Theorem 1) and Delta method (Hall, 1992a, Section 2.7). According to the discussion above, G_1 and G_4 transformations have faster convergence rate relative to G_2 and G_3. The following theorem gives the power function for the lower-tailed TT tests. The power function for the upper-tailed and two-sided tests can be obtained by computing $1 - P(\sqrt{n}G_i(T_n/\sqrt{n}) \leq z_\alpha)$ and $P(\sqrt{n}G_i(T_n/\sqrt{n}) \leq z_{\alpha/2}) + P(\sqrt{n}G_i(T_n/\sqrt{n}) \geq z_{1-\alpha/2})$, respectively. The proof of the theorem can be found in Appendix A.3.3.

Theorem 2. *Suppose regularity conditions in Appendix A.2 hold. Then*

$$P(\sqrt{n}G_i(T/\sqrt{n}) \leq z_\alpha) = F_{T,w}(\eta_{i,\alpha}) - n^{-1}c_{i,\alpha}(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2}),$$

where $F_{T,w}(\cdot)$ *is defined in (2.5),* $c_{1,\alpha} = c_{4,\alpha} = (B/6 - A^2/4)(2z_\alpha^2 + 1)$, $c_{2,\alpha} = c_{3,\alpha} = B/6 - A^2/4$,

$$\begin{aligned}
\eta_{1,\alpha} &= z_\alpha - n^{-1/2}(2z_\alpha^2 + 1)A/6 + n^{-1}(5z_\alpha^3 + 3z_\alpha)A^2/27, \\
\eta_{2,\alpha} &= z_\alpha - n^{-1/2}A/6 - n^{-1}Az_\alpha^2/3, \\
\eta_{3,\alpha} &= z_\alpha - n^{-1/2}(6z_\alpha^2 + A)/6 + n^{-1}(5z_\alpha^3 + Az_\alpha)/3 \\
\eta_{4,\alpha} &= z_\alpha - n^{-1/2}(2z_\alpha^2 + 1)A/6 + n^{-1}(4z_\alpha^3 + 3z_\alpha)A^2/27.
\end{aligned}$$

From this theorem, we observe that G_1 and G_4 lead to very similar power functions. Letting $w = 0$, one can easily verify (2.15)-(2.17) for the discussion under H_0.

2.3.1 Power comparison among the TT tests

Using the power functions in Theorem 2 we can compare the theoretical powers among the TT tests by looking at their power differences. The result is given in the following corollary.

Corollary 2. *The power differences for the lower-tailed TT_i tests are given below:*

$$PD_{TT_4,TT_1,lower} = n^{-1}\phi(z_\alpha - w)(-A^2 z_\alpha^3/27) + O(n^{-3/2})$$

$$PD_{TT_2,TT_i,lower} = n^{-1/2}\phi(z_\alpha - w)z_\alpha^2 A/3 + O(n^{-1}),\ \textit{for } i = 1 \textit{ or } 4,$$

$$PD_{TT_3,TT_i,lower} = n^{-1/2}\phi(z_\alpha - w)z_\alpha^2(A-3)/3 + O(n^{-1}),\ \textit{for } i = 1 \textit{ or } 4,$$

$$PD_{TT_2,TT_3,lower} = n^{-1/2}\phi(z_\alpha - w)z_\alpha^2 + O(n^{-1}).$$

Remark Since both $\phi(z_\alpha - w)$ and z_α^2 are positive, the only source that determines the sign of the power difference between TT$_1$ and TT$_2$ or TT$_3$ is the value of A, the adjusted relative skewness between two populations. Noting that z_α usually takes negative values, the lower tailed TT$_4$ test is more powerful than TT$_1$, and TT$_2$ is more powerful than TT$_3$, regardless of the value of A. When $A > 0$, the TT$_2$ test has more power than TT$_1$ and TT$_4$. When $A > 3$, the TT$_3$ test is more powerful than TT$_1$ and TT$_4$.

The conditions for the lower-tailed tests will be reversed for the upper-tailed alternative. To see this, we can compute the power differences as follows:

$$\text{PD}_{TT_4,TT_1,upper} = n^{-1}\phi(z_{1-\alpha} - w)(A^2 z_{1-\alpha}^3/27) + O(n^{-3/2}),$$

$$\text{PD}_{TT_2,TT_i,upper} = -n^{-1/2}\phi(z_{1-\alpha} - w)z_{1-\alpha}^2 A/3 + O(n^{-1}),\ \text{for } i = 1 \textit{ or } 4,$$

$$\text{PD}_{TT_3,TT_i,upper} = -n^{-1/2}\phi(z_{1-\alpha} - w)z_{1-\alpha}^2(A-3)/3 + O(n^{-1}),\ \text{for } i = 1 \textit{ or } 4,$$

$$\text{PD}_{TT_2,TT_3,upper} = -n^{-1/2}\phi(z_{1-\alpha} - w)z_{1-\alpha}^2 + O(n^{-1}).$$

In the upper-tailed case, TT$_4$ test is still more powerful than TT$_1$, but the TT$_3$ test becomes more powerful than TT$_2$. When $A > 0$, TT$_1$ and TT$_4$ have more power than TT$_2$. When $A > 3$, TT$_1$ and TT$_4$ are more powerful than TT$_3$. Analogously, we can further write the power differences for the two-sided alternative as follows:

$$\mathrm{PD}_{TT_4,TT_1,two-sided} = n^{-1}[\phi(z_{\alpha/2}-w)+\phi(z_{1-\alpha/2}-w)]z^3_{1-\alpha/2}A^2/27+O(n^{-3/2}),$$

$$\mathrm{PD}_{TT_2,TT_i,two-sided} = n^{-1/2}[\phi(z_{\alpha/2}-w)-\phi(z_{1-\alpha/2}-w)]z^2_{\alpha/2}A/3+O(n^{-1}),\ \text{for } i=1 \text{ or } 4,$$

$$\mathrm{PD}_{TT_3,TT_i,two-sided} = n^{-1/2}[\phi(z_{\alpha/2}-w)-\phi(z_{1-\alpha/2}-w)]z^2_{\alpha/2}(A-3)/3+O(n^{-1}),\ \text{for } i=1 \text{ or } 4,$$

$$\mathrm{PD}_{TT_2,TT_3,two-sided} = n^{-1/2}[\phi(z_{\alpha/2}-w)-\phi(z_{1-\alpha/2}-w)]z^2_{\alpha/2}+O(n^{-1}).$$

Unlike the one-sided case, the sign of the power difference depends on both A and w. For example, $w > 0$ leads to $\phi(z_{\alpha/2}-w)-\phi(z_{1-\alpha/2}-w) < 0$. In this case, if $A > 0$, the two-sided TT$_1$ and TT$_4$ tests are more powerful than the TT$_2$ test. Table 2.1 summarizes the comparison results for the TT tests, in which "$\succ$" is used to denote the power relation. For example, "$TT_1 \succ TT_2$" represents that TT$_1$ test is more powerful than TT$_2$ test. When $A = 0$, TT$_1$, TT$_2$ and TT$_4$ are equivalent up to a remainder of order $O(n^{-1})$; TT$_1$ and TT$_4$ are equivalent up to a remainder of order $O(n^{-3/2})$. When $A = 3$, TT$_1$, TT$_2$ and TT$_4$ are equivalent up to a remainder of order $O(n^{-1})$. For brevity, we do not list their details on the table.

Table 2.1: *Power comparison among TT tests*

Condition	Lower-tailed or two-sided ($w < 0$)	Upper-tailed or two-sided ($w > 0$)
$A < 0$	$TT_4 \succ TT_1 \succ TT_2 \succ TT_3$	$TT_3 \succ TT_2 \succ TT_4 \succ TT_1$
$0 < A < 3$	$TT_2 \succ TT_4 \succ TT_1 \succ TT_3$	$TT_3 \succ TT_4 \succ TT_1 \succ TT_2$
$A > 3$	$TT_2 \succ TT_3 \succ TT_4 \succ TT_1$	$TT_4 \succ TT_1 \succ TT_3 \succ TT_2$

It can be observed that TT$_2$ has the highest power when $A > 0$ for the lower-tailed and the two-sided alternatives when $w < 0$, but it ranks the lowest when $A > 0$ and $w < 0$. TT$_3$ is the best when $A < 3$ for the upper tailed or two-sided alternatives when $w > 0$, but it has the lowest power for lower tailed and two-sided alternatives with $w < 0$.

Note that the comparison conditions only consider the dominating term in the power difference. We implicitly assume that the $O(n^{-1/2})$ term dominates the $O(n^{-1})$ term. This is reasonable when the sample size is not very small.

2.3.2 Power comparison between TT and TCFU tests

We can also compare the power performance between the TCFU tests with the TT tests. Combining Theorem 1, equation (2.12) and Corollary 2, it can be seen that the power functions of the TCFU, the TT$_1$, and the TT$_4$ tests differ at the order of $O(n^{-1})$. Using the result in Appendix A.3.4 and the fact that $q_{20}(.)$ is an odd function, it can be shown that

$$\begin{aligned}
&\mathrm{PD}_{TT_1,TCFU_2,lower} = n^{-1}\phi(z_\alpha - w)[q_{30}(z_\alpha) - q_{20}(z_\alpha)] + O(n^{-3/2}),\\
&\mathrm{PD}_{TT_1,TCFU_2,upper} = n^{-1}\phi(z_{1-\alpha} - w)[q_{30}(z_\alpha) - q_{20}(z_\alpha)] + O(n^{-3/2}),\\
&\mathrm{PD}_{TT_1,TCFU_2,two-sided} = n^{-1}[q_{30}(z_{\alpha/2}) - q_{20}(z_{\alpha/2})][\phi(z_{\alpha/2} - w) + \phi(z_{1-\alpha/2} - w)] + O(n^{-3/2}),
\end{aligned} \tag{2.18}$$

where $q_{20}(.)$ is defined in equation (2.13) and

$$q_{30}(x) = A^2(5x^3 + 3x)/27.$$

It is observed that the condition for TT$_1$ being more powerful than the TCFU$_2$ test is $q_{30}(z_\alpha) > q_{20}(z_\alpha)$ for the one-sided alternatives and $q_{30}(z_{\alpha/2}) > q_{20}(z_{\alpha/2})$ for the two-sided alternative. We can also compare the TCFU$_1$ test with the TT$_1$ test as

$$\begin{aligned}
&\mathrm{PD}_{TT_1,TCFU_1,lower} = n^{-1}\phi(z_\alpha - w)q_{30}(z_\alpha) + O(n^{-3/2}),\\
&\mathrm{PD}_{TT_1,TCFU_1,upper} = n^{-1}\phi(z_{1-\alpha} - w)q_{30}(z_\alpha) + O(n^{-3/2}),\\
&\mathrm{PD}_{TT_1,TCFU_1,two-sided} = n^{-1}q_{30}(z_{\alpha/2})[\phi(z_{\alpha/2} - w) + \phi(z_{1-\alpha/2} - w)] + O(n^{-3/2}),
\end{aligned} \tag{2.19}$$

Note that both $q_{30}(z_\alpha)$ and $q_{30}(z_{\alpha/2})$ are less than 0, so the TCFU$_1$ test is more powerful than the TT$_1$ test for all types of alternatives. Both (2.18) and (2.19) still hold if we substitute TT$_1$ and q_{30} by TT$_4$ and $q_{40}(x) = A^2(4x^3 + 3x)/27$, respectively.

These comparisons are summarized in Table 2.2. The result can be obtained with basic algebra and the fact that $0 > q_{40}(x) > q_{30}(x)$ for $x < 0$.

Table 2.2: *Power comparison among $TCFU_1$, $TCFU_2$, TT_1, and TT_4 tests*

Condition	Comparison
$q_{20}(z) > 0$	$TCFU_2 \succ TCFU_1 \succ TT_4 \succ TT_1$
$0 > q_{20}(z) > q_{40}(z)$	$TCFU_1 \succ TCFU_2 \succ TT_4 \succ TT_1$
$q_{40}(z) > q_{20}(z) > q_{30}(z)$	$TCFU_1 \succ TT_4 \succ TCFU_2 \succ TT_1$
$q_{40}(z) > q_{30}(z) > q_{20}(z)$	$TCFU_1 \succ TT_4 \succ TT_1 \succ TCFU_2$

$z = z_\alpha$ and $z = z_{\alpha/2}$ for one-sided and two-sided alternatives, respectively.

For real data, it is rare to know the population parameters. Practitioners can estimate the unknown population parameters from the sample or by a pilot study, check the conditions we give above, and choose an appropriate testing procedure to achieve higher power.

2.3.3 Type I error comparison for the TT tests

Since type I error is a special case of power function by setting $w = 0$, we can easily give the difference in type I error between the TT tests. For reference in later discussion, we give the differences among the type I errors for the lower-tailed TT tests here:

$$\begin{aligned}
\text{T1ED}_{TT_4,TT_1,lower} &= n^{-1}\phi(z_\alpha)(-A^2 z_\alpha^3/27) + O(n^{-3/2}),\\
\text{T1ED}_{TT_2,TT_i,lower} &= n^{-1/2}\phi(z_\alpha) z_\alpha^2 A/3 + O(n^{-1}),\ \text{for } i = 1 \text{ or } 4,\\
\text{T1ED}_{TT_3,TT_i,lower} &= n^{-1/2}\phi(z_\alpha) z_\alpha^2 (A-3)/3 + O(n^{-1}),\ \text{for } i = 1 \text{ or } 4,\\
\text{T1ED}_{TT_2,TT_3,lower} &= n^{-1/2}\phi(z_\alpha) z_\alpha^2 + O(n^{-1}).
\end{aligned}$$

Considering the facts that $z_\alpha = -z_{1-\alpha}$ and $\phi(z_\alpha) = \phi(z_{1-\alpha})$, the type I error differences for the upper-tailed case are

$$\begin{aligned}
\text{T1ED}_{TT_4,TT_1,upper} &= n^{-1}\phi(z_\alpha)(-A^2 z_\alpha^3/27) + O(n^{-3/2}),\\
\text{T1ED}_{TT_2,TT_i,upper} &= -n^{-1/2}\phi(z_\alpha) z_\alpha^2 A/3 + O(n^{-1}),\ \text{for } i = 1 \text{ or } 4,\\
\text{T1ED}_{TT_3,TT_i,upper} &= -n^{-1/2}\phi(z_\alpha) z_\alpha^2 (A-3)/3 + O(n^{-1}),\ \text{for } i = 1 \text{ or } 4,\\
\text{T1ED}_{TT_2,TT_3,upper} &= -n^{-1/2}\phi(z_\alpha) z_\alpha^2 + O(n^{-1}).
\end{aligned}$$

The conditions leading to higher power for a test would also result in higher type I error.

The two-sided type I error is an important case due to the connection with the coverage probability of confidence intervals. Using the same test statistic and sampling distribution, one can construct an equivalent transformation confidence interval for $\mu_1 - \mu_2$ as:

$$\mathcal{I}_i = \left[\overline{X}_1 - \overline{X}_2 - n^{1/2} G_i^{-1} \left(\frac{z_{1-\alpha/2}}{\sqrt{n}} \right) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 - n^{1/2} G_i^{-1} \left(\frac{z_{\alpha/2}}{\sqrt{n}} \right) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right].$$

Subtracting the type I error of the two-sided TT$_i$ test from 1 leads to the coverage probability of $\mathcal{I}_i$. The type I errors for the TT tests are all of the form $\alpha + O(n^{-1})$. We will elaborate the $O(n^{-1})$ terms in the following corollary. The proof is deferred to Appendix A.3.5.

Corollary 3. *The type I error for the two-sided TT_i test is $\alpha + n^{-1}\Lambda_i + O(n^{-3/2})$, and the coverage probability of $\mathcal{I}_i$ is $1 - \alpha - n^{-1}\Lambda_i + O(n^{-3/2})$, $i = 1, 2, 3, 4,$ where*

$$\begin{aligned}
\Lambda_1 =& 2\phi(z_{\alpha/2})[A^2(5z_{\alpha/2}^3 + 3z_{\alpha/2})/27 - (B/6 - A^2/4)(2z_{\alpha/2}^3 + z_{\alpha/2}) - q_{20}(z_{\alpha/2})],\\
\Lambda_2 =& 2\phi(z_{\alpha/2})[2z_{\alpha/2}^3 A^2/9 - (B/6 - A^2/4)z_{\alpha/2} - q_{20}(z_{\alpha/2}) - A^2 z_\alpha^5/18],\\
\Lambda_3 =& 2\phi(z_{\alpha/2})[2z_{\alpha/2}^3 A(A-3)/9 + (5z_{\alpha/2}^3 + Az_{\alpha/2})/3 - (B/6 - A^2/4)z_{\alpha/2} - q_{20}(z_{\alpha/2})\\
& - (A-3)^2 z_{\alpha/2}^5/18],\\
\Lambda_4 =& 2\phi(z_{\alpha/2})[A^2(4z_{\alpha/2}^3 + 3z_{\alpha/2})/27 - (B/6 - A^2/4)(2z_{\alpha/2}^3 + z_{\alpha/2}) - q_{20}(z_{\alpha/2})].
\end{aligned}$$

Although all the type I errors converge to α with large sample size, examining the n^{-1} terms is not trivial for the finite sample setting and for the theoretical study. Zhou and Dinh (2005) compared $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$ through Monte Carlo experiments. The first-order Edgeworth expansion they established cannot distinguish the three transformations in terms of coverage probability because Corollary 3 shows that the coverage probabilities are all $1 - \alpha + O(n^{-1})$. They recommended the $\mathcal{I}_3$ over $\mathcal{I}_1$ and $\mathcal{I}_2$ based on their simulation results. In fact, the better performance of $\mathcal{I}_3$ can be explained analytically with the second-order expansions. For example, in their simulation, $\mathcal{I}_3$ often outperforms $\mathcal{I}_2$ in the coverage probabilities. By checking Corollary 3, the conditions for $\mathcal{I}_3$ having larger coverage probability than $\mathcal{I}_2$ is $\Lambda_2 > \Lambda_3$. With some algebra, the condition reduces to

$$(3-2A)z_{\alpha/2}^4 + (4A-10)z_{\alpha/2}^2 - 2A < 0.$$

When $\alpha = 0.05$, $z_{\alpha/2} \approx -1.96$, then the condition can be simplified as $5.858 - 16.149A < 0$, which holds when $A > 0.363$. Observing the simulation results in Zhou and Dinh (2005), the cases where $\mathcal{I}_3$ has more coverage probability than $\mathcal{I}_2$ mostly satisfy $A > 0.363$. Using our derived condition, we can have a simple but rigorous interpretation of the empirical results.

2.4 Simulation study

We conducted Monte Carlo simulation to compare the type I error rates and power of the TCFU, TT and classical t test. Another goal of the simulation is to verify the theoretical power comparisons summarized in Table 2.1. Note that the computation of q_{20}, q_{30}, and q_{40} involves A^2 whose estimate is often biased. It is difficult to verify Table 2.2 through simulation with small sample sizes.

2.4.1 Simulation settings

All the tests are carried out at the significance level of 5%. The sample sizes are unbalanced and $\lambda = n_1/n = 0.6$. If the total sample size is moderate and λ takes extreme values such as 0.1 or 0.9, one of the sample sizes would be too small to give satisfactory estimates for the population parameters.

In order to obtain the empirical type I errors and powers, each setting contains 2000 runs. In each run, a data set is generated, and each test will produce a decision "reject H_0" or "do not reject H_0". If the setting is under H_0, the proportion of rejections out of the 2000 runs gives the empirical type I error rate. If the setting is under H_a, the proportion of rejections is the empirical power.

The data under H_0 are randomly generated from the following settings:

- Setting 1: $LN(0,1)$ vs. $N(\sqrt{e}, 0.25)$, and $A = 7.11$;

- Setting 2: $-LN(0,1)$ vs. $N(-\sqrt{e}, 0.25)$, and $A = -7.11$,

where "LN" represents the log-normal distribution and "N" is the normal distribution. The center of the normal distribution has been adjusted to match the center of the log-normal distribution. Observing the power functions in Theorem 1 and Theorem 2, the relative skewness A plays an important role in the theoretical power. We configure the settings to have different A values.

Throughout the theoretical derivation, the local alternative $\delta_n = O(n^{-1/2})$ was assumed. For the simulation, we follow the strategy suggested by Cohen (1988) and set the effect size as $\delta = 0.3\sigma_1$, where σ_1 is the standard deviation of the first population. In the upper-tailed and two-sided alternatives, $0.3\sigma_1$ is added to each observation in population 1. In the lower-tailed alternatives, $0.3\sigma_1$ is subtracted from each observation of population 1.

All the simulation studies are implemented in `R 3.5.1`. The Welch's two-sample t-test (use the t-distribution to obtain the p-value) is also conducted as a benchmark. It is through the `t.test` function with the option `var.equal = False`. In the tables, the t-test is simply denoted by "t".

2.4.2 Numerical results

Table 2.3 displays the empirical type I error rates and power of the **upper-tailed** tests under **setting 1** at significance level $\alpha = 0.05$. The last column provides the average of 2000 estimated A values for 2000 runs. The TCFU_1 has liberal type I error rates for smaller sample sizes. The t-test, on the other hand, is conservative for all the sample sizes we considered. The rest of the tests have well-controlled type I error rates even with small sample sizes. The TT_1 and TT_4 have empirical type I error rates most close to the nominal level. The TCFU_2

and all the TT tests exhibit higher power than the t-test, especially when the sample sizes are small. TT$_3$ achieves the highest power.

Table 2.3: *The empirical type I error and power for the* ***upper-tailed*** H_a *in setting 1 at level* $\alpha = 0.05$. *The power for* $TCFU_1$ *for smaller sample sizes is not displayed due to overly inflated type I error.*

	n_1	TCFU$_1$	TCFU$_2$	TT$_1$	TT$_2$	TT$_3$	TT$_4$	t	$\hat{A}$
Type I error	15	0.087	0.020	0.046	0.020	0.056	0.048	0.009	1.845
	25	0.077	0.024	0.040	0.022	0.043	0.042	0.011	2.307
	40	0.078	0.029	0.048	0.026	0.045	0.050	0.014	2.784
	50	0.080	0.033	0.051	0.024	0.044	0.053	0.017	2.977
	80	0.064	0.032	0.046	0.022	0.037	0.048	0.014	3.446
	120	0.059	0.034	0.045	0.026	0.037	0.048	0.021	3.827
	160	0.065	0.039	0.052	0.031	0.042	0.053	0.024	4.123
	250	0.056	0.037	0.044	0.026	0.036	0.044	0.024	4.540
Power	15	–	0.345	0.460	0.359	0.499	0.468	0.249	1.845
	25	–	0.545	0.634	0.524	0.650	0.638	0.424	2.307
	40	–	0.772	0.823	0.750	0.828	0.826	0.674	2.784
	50	–	0.841	0.868	0.815	0.872	0.871	0.770	2.977
	80	–	0.961	0.967	0.956	0.968	0.967	0.942	3.446
	120	0.994	0.992	0.993	0.991	0.994	0.993	0.989	3.827
	160	1.000	1.000	1.000	1.000	1.000	1.000	1.000	4.123
	250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	4.540

The true value of A is 7.11 under setting 1. All the sample estimates for A are much below the true value even when the sample size n_1 reaches 250. The reason is that A contains the skewness of the two populations, whose estimation requires relatively large sample sizes for the highly skewed log-normal distribution. Although A is difficult to be accurately estimated, the TT$_1$, TT$_3$ and TT$_4$ tests are still able to give high power when the sample sizes are moderate, and they all show improvements over the classical t-test.

Because $A > 3$ in this case, following the comparison displayed in Table 2.1, we should expect $TT_4 \succ TT_1 \succ TT_3 \succ TT_2$ in terms of power or type I error. Our comparison takes place at the order of $O(n^{-1/2})$ or even higher order, so the distinction can be very small when the sample size is growing. Because A determines the relative rank of TT$_3$. In many cases when the sample size is not big enough, A tends to be underestimated. In these cases, TT$_3$ tends to have a bigger type I error rate and power than other TT tests. On the other hand, there are Monte Carlo errors. For example, in our 2000 runs, if the true power is 0.9, the margin of error can be loosely computed as $2\sqrt{0.1 * 0.9/2000} \approx 0.013$. The differences among TT_1, TT_3 and TT_4 are mostly within this margin of error when $n_1 > 25$. Despite these issues, we can still find evidence in Table 2.3 supporting our statement on the ranking of power or type I error rates.

For the **lower-tailed** H_a in **setting 1**, we report the cases with sample size n_1 up to 1500 (see Table 2.4) because this setting is challenging in controlling type I error. Figure 2.1 panel (I) shows the density curves under H_0. We can observe that the mass of the log-normal distribution concentrates on the left of the mass of the normal distribution. In this case, a lower-tailed test tends to support the $H_a : \mu_{LN} < \mu_N$. All the type I error rates are inflated when $n_1 < 300$, and the ordinary t-test deviates the most. The type I error of TT_3 is the least inflated in this setting and is relatively stable along with increasing sample sizes. The type I error of TCFU_2, TT_1, and TT_4 shrink to a similar level as TT_3 when the sample size is greater than 120. Since none of the type I errors is under control for smaller sample sizes, it is not necessary to compare their power. For $n_1 = 1000$, the empirical power is plotted in Figure 2.2. Even with such large sample sizes, the type I error rates for TT_2 and t are still heavily inflated.

Table 2.4: *The empirical type I error for the* ***lower-tailed*** H_a *in setting 1*

n_1	TCFU_1	TCFU_2	TT_1	TT_2	TT_3	TT_4	t	$\hat{A}$
15	0.142	0.128	0.130	0.163	0.009	0.133	0.165	1.845
50	0.096	0.090	0.090	0.120	0.071	0.092	0.126	2.977
120	0.085	0.079	0.080	0.102	0.077	0.081	0.104	3.827
250	0.070	0.066	0.066	0.087	0.066	0.068	0.090	4.540
300	0.068	0.065	0.064	0.080	0.068	0.066	0.083	4.625
500	0.063	0.061	0.061	0.072	0.064	0.062	0.073	5.012
750	0.064	0.063	0.063	0.073	0.064	0.063	0.075	5.289
1000	0.061	0.058	0.058	0.072	0.062	0.058	0.073	5.521
1500	0.050	0.050	0.050	0.058	0.052	0.050	0.060	5.802

Under the **two-sided** alternative for **setting 1** (see Table 2.5 and Figure 2.3), the TT_3 performs better than the rest of the tests: it well controls the type I error and shows high

Figure 2.1: *The density curves under H_0 (I), upper-tailed H_a (II), and lower-tailed H_a (III) in setting 1. The solid line represents normal distribution, and the dashed line represents the log-normal distribution.*

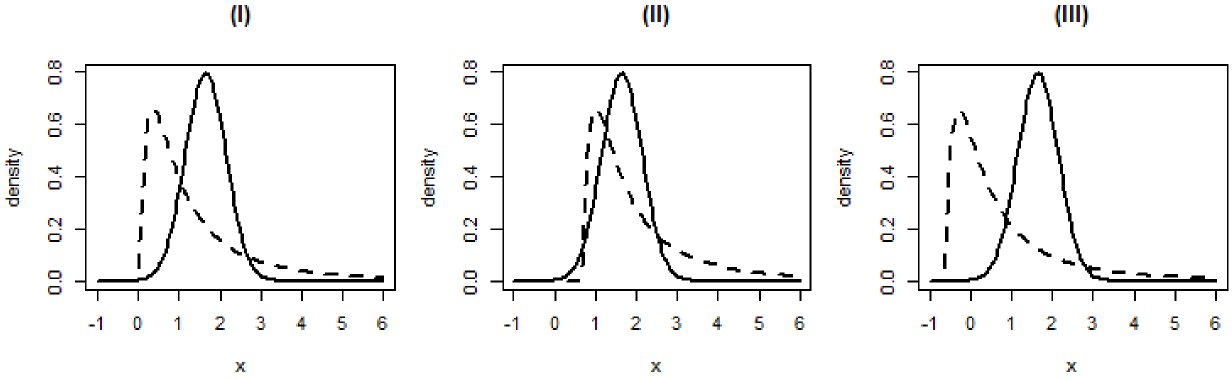
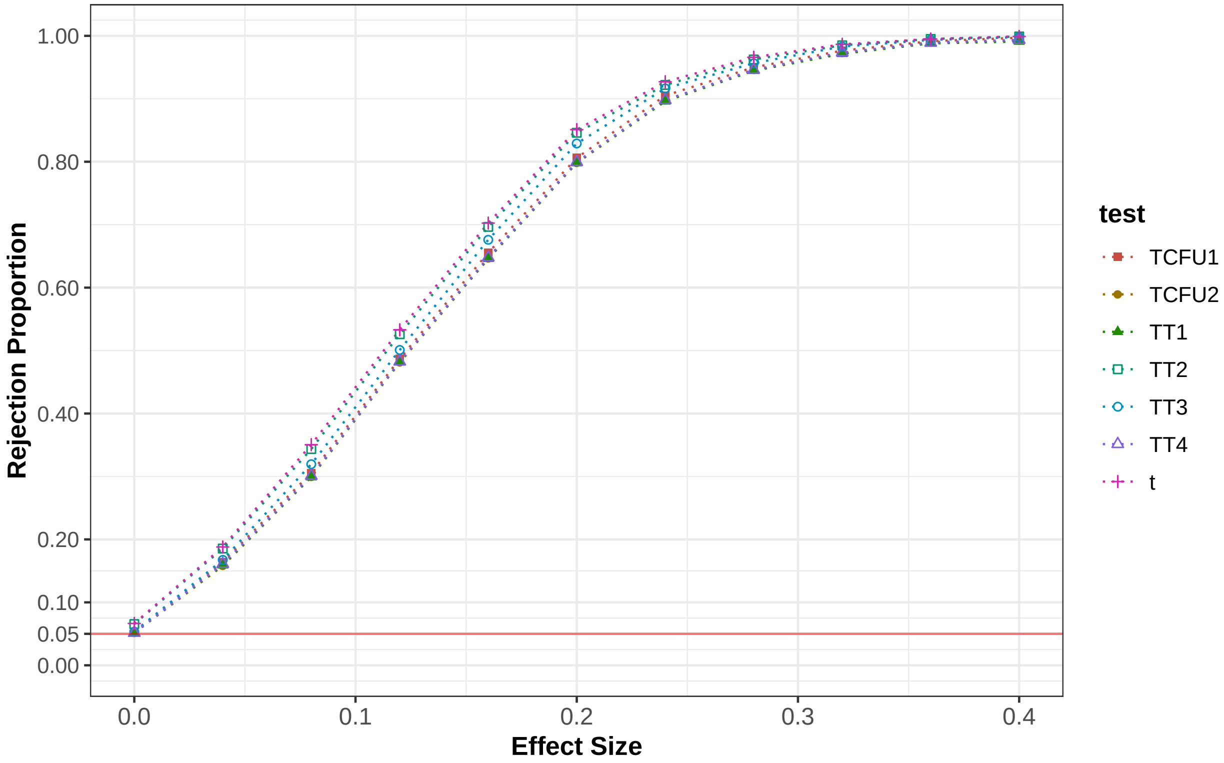


Figure 2.2: *The proportion of rejections for the* ***lower-tailed*** *H_a in setting 1 when $n_1 =$ 1000. "Effect Size" is the magnitude of δ in the alternative hypothesis.*



power. This finding agrees with the simulation done by Zhou and Dinh (2005) for confidence intervals. The power ranking based on Table 2.1 for the TT tests is $TT_4 \succ TT_1 \succ TT_3 \succ TT_2$.

If we account for the Monte Carlo error, this ranking loosely holds for $n_1 > 25$.

Table 2.5: *The empirical type I error for the* ***two-sided*** H_a *in setting 1 at level* $\alpha = 0.05$.

n_1	TCFU_1	TCFU_2	TT_1	TT_2	TT_3	TT_4	t	$\hat{A}$
15	0.167	0.095	0.110	0.131	0.028	0.117	0.122	1.845
25	0.141	0.076	0.091	0.114	0.036	0.093	0.114	2.307
40	0.118	0.062	0.078	0.083	0.055	0.082	0.087	2.784
50	0.119	0.071	0.080	0.090	0.057	0.085	0.091	2.977
80	0.094	0.059	0.069	0.078	0.056	0.071	0.077	3.446
120	0.095	0.064	0.071	0.076	0.059	0.073	0.075	3.827
160	0.083	0.058	0.066	0.068	0.056	0.067	0.066	4.123
250	0.078	0.057	0.063	0.061	0.054	0.065	0.059	4.540
300	0.073	0.064	0.064	0.068	0.063	0.066	0.066	4.625
350	0.068	0.058	0.059	0.058	0.056	0.060	0.053	4.744

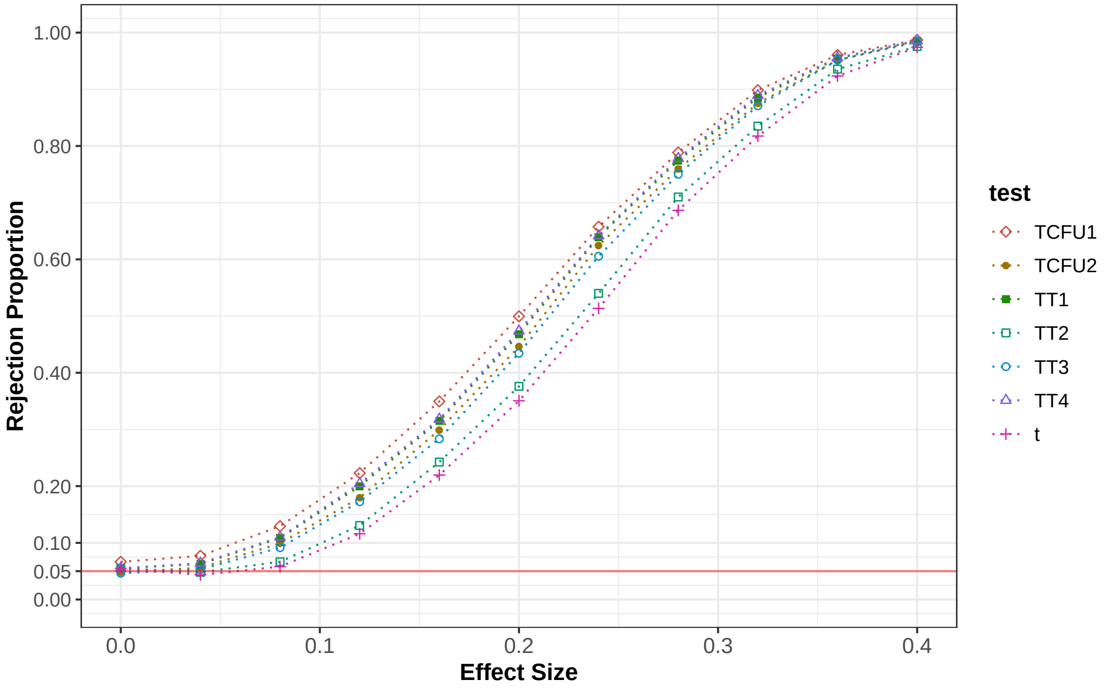


Figure 2.3: *The proportion of rejections for* ***two-sided*** H_a *in setting 1 when* $n_1 = 350$.

The readers may have an impression that TT_3 is the best among all the tests being compared. Revisiting the transformation definitions in (2.14), it is noted that G_3 is equivalent to G_1 when $\hat{A} = 3$. The transformation G_1 has its theoretical background rooted in the Edgeworth expansion, but G_3 forces $\hat{A} = 3$. This is an advantage when the actual value A is positive and the sample estimate $\hat{A}$ underestimates A. When A is negative, manually forcing

$\hat{A} = 3$ in G_3 transformation will become a disadvantage. To see this, we designed setting 2 in which $A = -7.11$. The results for **lower-tailed** alternative in **setting 2** are shown in Table 2.6. TT$_3$ gives extremely conservative type I error rates for small and medium sample sizes, and the power of it is even lower than the t-test. TT$_1$ and TT$_4$ have empirical type I error around 5%, and produce the highest power among others. Moreover, Table 2.1 states the power rank as $TT_4 \succ TT_1 \succ TT_2 \succ TT_3$, and it is confirmed by the results.

Table 2.6: *The empirical type I error and power for the* ***lower-tailed*** *H_a in setting 2 at level $\alpha = 0.05$.*

	n_1	TCFU$_1$	TCFU$_2$	TT$_1$	TT$_2$	TT$_3$	TT$_4$	t	$\hat{A}$
Type I error	15	0.083	0.024	0.043	0.024	0.000	0.045	0.012	-1.849
	25	0.075	0.021	0.045	0.019	0.001	0.049	0.013	-2.311
	40	0.073	0.032	0.044	0.020	0.002	0.046	0.009	-2.785
	50	0.078	0.032	0.050	0.022	0.002	0.051	0.013	-2.978
	80	0.068	0.036	0.046	0.025	0.007	0.048	0.019	-3.444
	120	0.066	0.038	0.050	0.026	0.010	0.052	0.019	-3.827
	160	0.059	0.038	0.047	0.022	0.013	0.048	0.018	-4.122
	250	0.058	0.040	0.047	0.027	0.020	0.048	0.025	-4.539
Power	15	-	0.332	0.459	0.345	0.002	0.465	0.227	-1.849
	25	-	0.548	0.653	0.527	0.141	0.657	0.435	-2.311
	40	-	0.757	0.810	0.733	0.449	0.813	0.652	-2.785
	50	-	0.845	0.874	0.817	0.609	0.876	0.767	-2.978
	80	-	0.962	0.970	0.948	0.882	0.971	0.934	-3.444
	120	-	0.995	0.996	0.995	0.985	0.996	0.993	-3.827
	160	1.000	1.000	1.000	0.999	0.998	1.000	0.998	-4.122
	250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-4.539

In both settings, $q_{20}(z_\alpha)$ and $q_{20}(z_{\alpha/2})$ are less than 0. According to Corollary 1 and the simulation results, the TCFU$_1$ is more powerful than TCFU$_2$. But very often, TCFU$_1$ cannot provide acceptable type I error control. This indicates that our further correction at the $O(n^{-1})$ term in the TCFU$_2$ test improves the accuracy of approximation. Taking the lower-tailed test for example, the TCFU$_2$ uses $z_\alpha + n^{-1/2}\hat{q}_{10}(z_\alpha) + n^{-1}\hat{q}_{20}(z_\alpha)$ as the critical value, while the TCFU$_1$ directly discards the $n^{-1}\hat{q}_{20}(z_\alpha)$. Despite the poor estimation of $\hat{q}_{20}(z_\alpha)$, keeping this term still improves the accuracy of approximation.

In summary, the TCFU and TT tests all improve the ordinary t-test to some degree under different settings. No test can provide globally best performance under all circumstances, including TT_3 advocated by Zhou and Dinh (2005). Our analytical conditions for high power are mostly verified by the simulation results. Practitioners can use these conditions as a guideline to choosing the best test.

2.5 Real-data examples

In this section, the TCFU and TT tests are applied to real-data examples. The first example is a direct application to a univariate two-sample mean testing problem. The second example comes from the genome study focusing on detecting differentially expressed genes.

2.5.1 Repair-person mileage

Ott and Longnecker (2008) exercise 6.17 describes that a cable TV company is interested in making the operation more efficient by reducing the distances between service calls while maintaining the service quality. The treatment group has 18 repair-persons. In this group, a dispatcher monitors all incoming repair requests and decides the service orders. The control group also has 18 persons who work in the regular routine, providing service roughly in a sequential order as requests coming in. The average mileage for each person is listed below:

Treatment	62.2, 79.3, 83.2, 82.2, 84.1, 89.3, 95.8, 97.9, 91.5, 96.6, 90.1, 98.6, 85.2, 87.9, 86.7, 99.7, 101.1, 88.6
Control	97.1, 70.2, 94.6, 182.9, 85.6, 89.5, 109.5, 101.7, 99.7, 193.2, 105.3, 92.9, 63.9, 88.2, 99.1, 95.1, 92.4, 87.3

The company wants to determine whether the new workflow successfully reduces the daily mileage of repair-persons. The estimates defined by (2.8) are $S_1 = 33.016$, $S_2 = 9.333$,

$\hat{\gamma}_1 = 2.112$, $\hat{\gamma}_2 = -1.221$, $\hat{\tau}_1 = 6.196$, and $\hat{\tau}_2 = 5.050$. $\hat{A} = 2.696$ by plugging in those moment estimates. At the 5% significance level, we apply the upper-tailed TCFU, TT, and t-test to this data set and display the results in Table 2.7.

Table 2.7: *The test results for repair-person mileage data.*

	TCFU_1	TCFU_2	TT_1	TT_2	TT_3	TT_4	t
Test statistic	1.705	1.705	2.253	1.855	2.310	2.300	1.705
Critical value	1.165	1.470	1.645	1.645	1.645	1.645	1.726*
Reject H_0	Yes	Yes	Yes	Yes	Yes	Yes	No

*The critical value is from t-distribution with degrees of freedom 19.7 based on the Satterthwaite approximation.

All the TCFU and TT tests reach consistent conclusion. The critical value of TCFU_2 differs from that of TCFU_1 by a significant amount $(1.470 - 1.165 = 0.305)$, which is due to the effect of order $O(n^{-1})$. All the procedures successfully reject the H_0 except for the classical t-test. Compared to the t-test, the TCFU tests correct the critical values to smaller values and the TT tests correct the test statistics to large values.

2.5.2 Detect differentially expressed genes

In biological studies, identifying differentially expressed genes helps understand the complex mechanism at the genetic level. A commonly used technique is the two-sample t-test. When the population does not follow normal distribution, has skewness, or the data have small sample sizes, the t-test cannot provide reliable test results. To illustrate the advantage of our new tests, we use the TCFU and TT tests to identify differentially expressed genes for certain phenotypes. Through a series of experiments, the new tests demonstrate good control in type I error and higher statistical power in the task.

The dataset we used is the acute lymphoblastic leukemia (ALL) dataset firstly introduced by Chiaretti et al. (2004). It contains 128 cell observations with 12,625 microarray expression measures and 21 phenotypes. For the two-sample test setting, we focus on two phenotypes: B-cell ALL with the BCR/ABL fusion (sample size $n_1 = 37$) and cytogenetically normal NEG B-cell ALL (sample size $n_2 = 42$). We obtained the data from the `ALL` (Li, 2009) package in R software, which contains the expression measures of genes preprocessed by three-step robust multichip average method and was subjected to base 2 logarithmic transformation. We follow the strategy of Dudoit et al. (2008) to filter the genes: first, retain the genes with expression measure greater than 100, in the absolute scale, in at least 25% of the 79 observations; second, retain the genes with expression measure having interquartile range greater than 0.5, in log base 2 scale. Finally, 2,391 genes were kept for further consideration after filtering. For each of these genes, we apply the t-test, TCFU, and TT tests to testing the equality of means of "BCR/ABL" group vs. "NEG" group. If the test rejects the null hypothesis of equal means, the gene is considered as a potentially differentially expressed gene. With this configuration, more liberal tests can always have more rejections. Therefore, we designed another set of experiments which simulates a pseudo null hypothesis by randomly splitting the 42 "NEG" observations into two groups, 21 observations for each, and conduct the tests to compare their means for each gene. In this setting, the tests are expected to not reject the null hypothesis. Since each test was applied to many genes, we further apply p-value adjustments to control the family-wise error rate or false discovery rate. The results are summarized in Table 2.8 where the significance level is set as $\alpha = 0.05$.

Table 2.8: *The number of rejections out of 2391 genes for detecting differentially expressed genes on the ALL data*

	Adjustment	TCFU$_1$	TCFU$_2$	TT$_1$	TT$_2$	TT$_3$	TT$_4$	t
NEG vs. NEG	Bonferroni	0	0	0	0	8	0	0
BCR/ABL vs. NEG	Bonferroni	24	15	28	26	99	28	14
NEG vs. NEG	BY	0	0	0	0	8	0	0
BCR/ABL vs. NEG	BY	53	26	50	52	142	50	24

"BY" is the FDR control method proposed by Benjamini and Yekutieli (2001)

Even though the Bonferroni method controlled the family-wise type I error at 5%, TT$_3$ test reported that some genes are significantly different in the NEG vs. NEG setting. All other tests rejected none of them. This indicates that the tests can control the type I error well. When comparing the means of BCR/ABL vs. NEG, TCFU and TT tests can identify more significantly different genes than t-test. The results show that the TCFU and TT tests provide more statistical power and controlled type I error rates.

2.6 Potential use in high-dimensional variable screening

In multi-class classification problems, it is of interest to compare one class with the rest of the classes. For example, in support vector machine classification, a typical way of handling the multiple classes is through one-versus-other scheme, i.e., comparing one class versus other classes. When the dimension is high, the support vector machine becomes computationally very extensive. Consequently, it is a common practice that variable screening is applied first to remove non-informative variables that do not help differentiate class labels. The one-versus-other scheme can be thought of as a two-sample comparison, in which one sample is from a pure population while the other is from a mixture distribution that contains multiple populations. Another example is the nearest shrunken centroid algorithm (Tibshirani et al., 2002), which shrinks the studentized comparison statistic toward zero via soft-thresholding to help select important variables. The studentized comparison statistic is equivalent to the pooled classical t-statistic of comparing the sample from one class with the sample from a mixture of other classes. The theory behind the soft-thresholding is based on normally distributed data (Donoho, 1995).

In the aforementioned screening for large-scale variable selection problems, t-test or t-statistic with large-sample normal approximation is often used. We have shown that it may not be robust to the skewness of the data. We conducted a Monte Carlo experiment to investigate the performance of the TCFU and TT tests under the one-vs-other setting. Population 1 is set as the standard Normal distribution, and Population 2 is a mixture of two distinct distributions: the random variable has 50% chance to come from a Gamma distribution with shape parameter 0.18 and rate parameter 0.3, and has 50% chance to come from a log-normal distribution with logarithm mean and standard deviation 0 and 0.5, respectively. The Gamma and log-normal distributions are shifted to have mean 0. The effect size $\delta = 0.3$ is added to Population 1 for upper-tailed and two-sided alternatives and subtracted from Population 1 for lower-tailed alternative.

The empirical type I error and power of all tests are given in Table 2.9. It can be seen that in one-sided case, TCFU$_1$, TT$_2$, TT$_3$ and classical t-test converge to their nominal type I error slower than TCFU$_2$, TT$_1$, and TT_4. In lower-tailed case, all TCFU and transformation tests are more powerful than the classical t-test except for TT$_3$. They can be used to help reduce false-positive rates in high-dimensional variable screening and increase the power to identify important features in marginal screening. An application of the proposed test in the nearest shrunken centroid algorithm could lead to an algorithm that will give variable-specific optimal thresholding parameters. This would allow different variables to have very different distributions as is often the case in high throughput genomic data. We will explore such applications in future work.

2.7 Summary and discussion

In this section, we proposed the TCFU and the TT tests for testing the equality of means of two independent univariate populations with unequal variances. The TCFU tests extended the TCF test (Wang et al., 2017) to the unequal-variance scenario. The TT tests considered four transformations including one correcting a transformation proposed by Hall (1992b). Through the theoretical power functions, we compared the new tests in terms of power and type I error and derived the analytical conditions leading to high power. These conditions, summarized in Tables 2.1 and 2.2, depend on the relative skewness, the pooled kurtosis, and the adjusted effect size. The theoretical result reveals that for one-sided alternatives, the TCFU$_2$, TT$_1$ and TT$_4$ tests have type I error converging to the nominal level faster than that of the TT$_2$ and TT$_3$ tests. The power ranking $TCFU_1 \succ TT_4 \succ TT_1$ always holds for all types of alternatives. We also presented the coverage probabilities accurate to $O(n^{-1})$ for two-sided transformation-based confidence intervals. Using these theoretical results, we can provide a more rigorous explanation of the simulation results obtained by Zhou and Dinh (2005). Monte Carlo simulation studies showed that no test can achieve the best performance for all scenarios.

Xu (2010) and Zhou and Dinh (2005) mentioned that higher power or coverage probability can always be achieved by managing the label to get a positive A value. This strategy is feasible for two-sided tests or confidence intervals since changing labels will not affect the hypothesis. But for one-sided cases, switching the labels will reverse the direction of the hypothesis and change the conditions associated with high power.

Although Bootstrap-t procedure amounts to the infinite-order Edgeworth expansion, where moments are replaced by plug-in estimators, it can only achieve the same order of accuracy as the TCFU procedure does. The asymptotic properties of the TCFU tests can provide an appealing alternative to estimate the power of the Bootstrap-t test. The power of the bootstrapping test could not be computed with simple bootstrap because it is always conducted under H_0. Moreover, the TCFU tests are much more efficient with computation, which is especially important when testing a large number of hypotheses.

Table 2.9: *The empirical type I error and power at level $\alpha = 0.05$ in the* ***mixture setting****. Due to overly inflated type I error for smaller sizes for two-sided and upper-tailed alternatives, the power was only given for large sample sizes.*

H_a		n_1	TCFU_1	TCFU_2	TT_1	TT_2	TT_3	TT_4	t	$\hat{A}$
Two sided	Type I error	15	0.132	0.061	0.089	0.082	0.141	0.091	0.063	-1.002
		25	0.117	0.066	0.081	0.072	0.109	0.083	0.064	-1.364
		40	0.099	0.058	0.074	0.068	0.092	0.075	0.064	-1.684
		50	0.103	0.072	0.080	0.074	0.087	0.083	0.068	-1.973
		80	0.084	0.060	0.068	0.062	0.079	0.069	0.059	-2.264
		120	0.074	0.055	0.061	0.060	0.072	0.064	0.059	-2.577
		250	0.058	0.050	0.050	0.048	0.052	0.051	0.045	-3.196
		500	0.062	0.059	0.060	0.056	0.061	0.060	0.056	-3.471
		1000	0.062	0.059	0.060	0.058	0.056	0.061	0.056	-3.953
	Power	250	0.723	0.705	0.699	0.785	0.828	0.707	0.794	-3.196
		500	0.934	0.928	0.926	0.956	0.965	0.929	0.959	-3.471
		1000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	-3.953
Upper tailed	Type I error	50	0.080	0.077	0.078	0.085	0.117	0.078	0.086	-1.973
		80	0.075	0.074	0.075	0.083	0.112	0.075	0.083	-2.264
		120	0.073	0.070	0.071	0.083	0.103	0.071	0.084	-2.577
		250	0.055	0.053	0.053	0.061	0.068	0.054	0.064	-3.196
		500	0.064	0.062	0.062	0.074	0.084	0.062	0.077	-3.471
		1000	0.058	0.057	0.058	0.064	0.069	0.058	0.064	-3.953
	Power	250	0.823	0.804	0.803	0.858	0.880	0.807	0.866	-3.196
		500	0.964	0.962	0.961	0.974	0.977	0.962	0.976	-3.471
		1000	0.999	0.999	0.999	1.000	1.000	0.999	1.000	-3.953
Lower tailed	Type I error	15	0.087	0.030	0.051	0.032	0.000	0.054	0.019	-1.002
		25	0.084	0.034	0.054	0.030	0.002	0.058	0.018	-1.364
		40	0.075	0.042	0.056	0.031	0.005	0.058	0.020	-1.684
		50	0.070	0.035	0.051	0.028	0.005	0.052	0.023	-1.973
		80	0.064	0.040	0.050	0.030	0.010	0.050	0.021	-2.264
		120	0.052	0.037	0.045	0.024	0.008	0.046	0.019	-2.577
		250	0.051	0.040	0.043	0.029	0.016	0.043	0.024	-3.196
		500	0.051	0.046	0.048	0.041	0.032	0.048	0.039	-3.471
		1000	0.053	0.051	0.052	0.042	0.034	0.052	0.039	-3.953
	Power	15	-	0.151	0.219	0.163	0.002	0.223	0.119	-1.002
		25	-	0.234	0.292	0.231	0.046	0.296	0.177	-1.364
		40	-	0.339	0.380	0.323	0.139	0.385	0.279	-1.684
		50	-	0.404	0.444	0.373	0.200	0.448	0.331	-1.973
		80	-	0.566	0.599	0.533	0.376	0.600	0.495	-2.264
		120	0.743	0.715	0.730	0.692	0.585	0.731	0.667	-2.577
		250	0.954	0.948	0.952	0.940	0.921	0.953	0.938	-3.196
		500	0.997	0.997	0.997	0.997	0.995	0.997	0.996	-3.471
		1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-3.953

Chapter 3

A high-dimensional two-sample test

3.1 Introduction

Testing the equality of multivariate means is a fundamental method in multivariate statistics and a useful tool in data analysis tasks, for example, in life science studies (Goeman and Bühlmann, 2007; Van De Ville et al., 2004). A classical procedure to test the equality of multivariate means is the Hotelling's T^2-test. Suppose random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^p$ have mean $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ be two groups of independent random samples from the distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The multivariate two-sample mean test focuses on the hypotheses $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. The Hotelling's T^2-statistic is defined by

$$T^2 = \frac{nm}{n+m}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^\top \mathbf{S}^{-1}(\overline{\mathbf{X}} - \overline{\mathbf{Y}}) \tag{3.1}$$

where $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ are the sample means for each group, $\mathbf{S}_1 = (n-1)^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top$, $\mathbf{S}_2 = (m-1)^{-1}\sum_{j=1}^{m}(\mathbf{Y}_j - \overline{\mathbf{Y}})(\mathbf{Y}_j - \overline{\mathbf{Y}})^\top$ are the sample covariance matrices for each group, and $\mathbf{S} = [(n-1)\mathbf{S}_1 + (m-1)\mathbf{S}_2]/(n+m-2)$ is the pooled sample covariance matrix. Assuming that $\mathbf{X}$ and $\mathbf{Y}$ follow multivariate normal distributions and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the null distribution of the T^2 is given by the F-distribution:

$$\frac{(n+m-p-1)}{p(n+m-2)} T^2 \sim F(p, n+m-p-1).$$

The “large-p-small-n” problem arises with new data collection technologies, for example, high-throughput data from life sciences, and hence challenges classical statistical methods including the Hotelling’s test. When $p > n+m-2$, the T^2 defined in (3.1) is not well defined because of the singularity of the sample covariance matrix $\mathbf{S}$ (Dempster, 1958) .

Blair et al. (1994) proposed a permutation test for paired two-sample multivariate means comparison. The p-dimensional difference vector was used to generate permutations. The result in their paper shows how Hotelling’s T^2 can fail in the large-p-small-n case. They considered three statistics for the permutation tests: the max of the component-wise $|t|$’s, the sum of the component-wise t’s, and the sum of the component-wise $|t|$’s. Their permutation tests can be applied to any p and n, large or small, even including the cases where the T^2 cannot be computed. Besides Blair et al. (1994), other existing studies mainly focus on two representative approaches. The first one tries to modify the distance metric in T^2. It is noted that the core part of T^2 is an empirical Mahalanobis distance that measures the separation of the group means. The challenge of high dimension comes from the singularity of the pooled covariance matrix $\mathbf{S}$. Other distances were adopted to avoid computing the pooled covariance matrix, for example, Euclidean distance (Bai and Saranadasa, 1996; Dempster, 1958), diagonalize Mahalanobis distance (Gregory et al., 2015; Srivastava and Du, 2008; Srivastava et al., 2013), and Lomonosov distance (Cai et al., 2014). Another popular approach is the random projection. Lopes et al. (2011) firstly introduced this technique to the area of high-dimensional tests. Through a random projection matrix, the data points are projected to a low-dimensional space, where the T^2 is well defined. To overcome the weak performance of a single random projection test, the implementation of these tests usually depends on an ensemble of a large number of repetitions. Variations of random projection approach can be found in Thulin (2014), Srivastava et al. (2016), and Zoh et al. (2018). More details of the related literature will be discussed in Section 3.2.

Here we propose a new test for testing the equality of high-dimensional means. Suppose $t_1, \ldots, t_p$ are the univariate two-sample Welch's t-statistics for each of the p components. Our new test is based on $T_n = \sum_{j=1}^{p} t_j^2$, i.e., the average of the component-wise squared t-statistics. With proper assumptions, it can be shown, by central limit theorem, that the standardized T_n asymptotically converges to normal distribution when the sample sizes and the number of dimensions both go to infinity. Standardizing T_n involves an unknown population scaling parameter $\sqrt{\mathrm{var}(T_n)}$. For real data, we propose to estimate $\mathrm{var}(T_n)$ by estimating its components $\mathrm{cov}(t_j^2, t_{j'}^2)$ using the sample information.

Gregory et al. (2015) proposed the generalized component test (GCT) which is also based on $T_n = \sum_{j=1}^{p} t_j^2$. They treat $\{t_1^2, \ldots, t_p^2\}$ as a univariate sequence and estimate $\mathrm{var}(T_n)$ by the summing up the autocovariance estimates. Our test has several advantages over the GCT test. First, the GCT test uses t_j^2 $(j = 1, \ldots, p)$ as the smallest unit to estimate the autocovariance. Consequently, there is no replication for each component t_j^2. Our test directly uses the original sample data as the units for estimation and hence gains replication from the data points. Second, we will show that the $\mathrm{cov}(t_j^2, t_{j'}^2)$ has a dominating term being nonnegative. Our scaling parameter can guarantee the nonnegativeness while the GCT test cannot. Third, in the GCT test, the estimations for autocovariance and spectral density require stationarity assumption, which makes the test restrictive. Our test avoids the stationarity condition by directly estimating the covariance for each pair of the squared t-statistics. It is noted that Srivastava et al. (2013) proposed a test that used a similar scaling parameter, but their theoretical properties were established under normality assumptions.

Through Monte Carlo experiments, we examined the impact of center correction, the dependence structure, and the innovation distributions on our new test and some existing tests. Our new test and Srivastava et al. (2013) test are more powerful under independent or weakly dependent cases. The new test is also robust to light skewness and heavy tails, even when the moment conditions are violated.

We also designed a series of experiments on the acute lymphoblastic leukemia (ALL) dataset (Chiaretti et al., 2004). Our new test and other existing tests are used to identify differentially expressed Gene Ontology terms. The new test shows good control in type I error and more statistical power. It is worth noting that, we do not have the ground truth of the important Gene Ontology terms, but on different (pseudo) datasets, our test provides much more consistent detection than other tests.

The remaining sections are organized as follows: Section 3.2 reviews the existing studies relevant to the two-sample test in high dimension; Section 3.3 introduces our new scaling parameter and new test statistics; Section 3.4 presents theoretical results including the sampling distribution for the new test under both the null and alternative hypotheses; Section 3.5 compares the empirical type I error and power of our new test and several existing tests via Monte Carlo experiments; Section 3.6 applies the new test to the acute lymphoblastic leukemia (ALL) dataset (Chiaretti et al., 2004); Section 3.7 concludes this chapter.

3.2 Related work

In this section, we will review some representative work in the area of testing the equality of high-dimensional means.

3.2.1 Tests with modified distances

The main reason why the Hotelling's T^2 fails in high-dimension lies in the singularity of the sample covariance matrix $\mathbf{S}$. In the core of T^2, there is a metric for the mean difference $(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^\top \mathbf{S}^{-1} (\overline{\mathbf{X}} - \overline{\mathbf{Y}})$ where a sample version of Mahalanobis distance is used. As it was pointed out by Hu and Bai (2016), the covariance matrix involved in the Mahalanobis distance demands estimating too many parameters relative to the limited sample size. Dempster (1958) suggested constructing a test based on the Euclidean distance so that the test is well defined for the high-dimensional cases. The statistic consists of a ratio of two mean squared distances. To derive the null distribution, it invokes a so-called χ^2- approximation technique which involves an unknown degrees of freedom parameter. The fact that a parameter needs to be estimated gives its name "non-exact test". Besides the feasibility in high dimension, Dempster's test also achieves higher power even when T^2 is well defined but p is large relative to the sample size. But the null distribution strongly relies on the normal assumption. The estimation for the degrees of freedom is very complicated.

The value of the early work by Dempster has been reemphasized with the growing research interest in DNA microarrays data. Following Dempster's test, Bai and Saranadasa (1996) considered the difference between two mean squared distances instead of the ratio:

$$T_{BS} = ||\overline{\mathbf{X}} - \overline{\mathbf{Y}}||^2 - \frac{n+m}{nm}\text{tr}(\mathbf{S}).$$

Further scaling by an estimate of standard deviation of T_{BS} makes the test statistic have asymptotic normal distribution. Chen and Qin (2010) observed that the tr($\mathbf{S}$) term used in T_{BS} is to offset the squared terms in $||\overline{\mathbf{X}} - \overline{\mathbf{Y}}||^2$. Those terms impose extra restriction on dimensionality but do not bring any benefit. They suggested to use the statistic

$$T_{\text{CQ}} = \frac{\sum_{i\neq j}^{n} \mathbf{X}_i'\mathbf{X}_j}{n(n-1)} + \frac{\sum_{i\neq j}^{m} \mathbf{Y}_i'\mathbf{Y}_j}{m(m-1)} - 2\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{X}_i'\mathbf{Y}_j}{nm}.$$

It removes some terms from the Euclidean distance without loss of information needed for testing, and it also relaxes the regularity conditions. Srivastava et al. (2013) pointed out that there is a natural variance estimator for T_{CQ} having uniformly minimum variance among all unbiased estimators, but Chen and Qin (2010) used a complicated estimator not possessing the property.

Srivastava and Du (2008) only retain the diagonal elements in the covariance matrix to avoid the singularity problem. They considered a statistic

$$T_{\text{SD}} = \frac{\frac{nm}{n+m}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})' D_S^{-1} (\overline{\mathbf{X}} - \overline{\mathbf{Y}}) - \frac{(n+m-2)p}{n+m-4}}{\sqrt{2\left(\text{tr}R^2 - \frac{p^2}{n+m-2}\right) c_p}},$$

where $D_S = \text{diag}(s_{11}, \ldots, s_{pp})$, and s_{ii} is the i-th diagonal element of the matrix $\mathbf{S}$, $R = D_S^{-1/2} \mathbf{S} D_S^{-1/2}$ is the sample correlation matrix, and $c_p = 1 + \text{tr}R^2 p^{-3/2}$ is an adjustment coefficient converging to 1 when p and sample sizes tend to infinity. Compared with the Euclidean distance, D_S^{-1} serves as a weighting matrix for each component accounting for heterogeneous variances. Their asymptotic distribution for T_{SD} is obtained under assumptions of normality and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Wu et al. (2006) constructed a test statistic based on the average of squared pooled component-wise t-statistic which is equivalent to $(\overline{\mathbf{X}} - \overline{\mathbf{Y}})' D_S^{-1} (\overline{\mathbf{X}} - \overline{\mathbf{Y}})$, but the distribution is given without any theoretical justification. Srivastava et al. (2013) extended the test to unequal covariance matrices situation by considering:

$$T_{\text{SKK}} = (\overline{\mathbf{X}} - \overline{\mathbf{Y}})' \hat{D}^{-1} (\overline{\mathbf{X}} - \overline{\mathbf{Y}}) - p,$$

where $\hat{D} = \hat{D}_1/n + \hat{D}_2/m$, $\hat{D}_1 = \text{diag}(S_{111}, \ldots, S_{1pp})$, $\hat{D}_2 = \text{diag}(S_{211}, \ldots, S_{2pp})$, S_{1ii} is the i-th diagonal element of the sample covariance $\mathbf{S}_1$, S_{2ii} is the i-th diagonal element of $\mathbf{S}_2$. The null distribution still relies on normality assumption. Gregory et al. (2015) proposed the GCT statistic based on the average of component-wise squared t-statistic $T_n = \sum_{j=1}^{p} t_j^2$, equivalent to $(\overline{\mathbf{X}} - \overline{\mathbf{Y}})' \hat{D}^{-1} (\overline{\mathbf{X}} - \overline{\mathbf{Y}})$. Our test statistic has some connections with Gregory et al. (2015). M ore details will be discussed in next section.

Cai et al. (2014) noted that the tests using Euclidean and Mahalanobis distance may not be powerful for sparse alternative. Instead, they proposed to use Lomonosov distance

(L_∞-norm) to establish the test by the strongest signal. The test statistic is

$$T_{\text{CLX}} = \frac{nm}{n+m} \max_{1\le i\le p} \frac{\hat{Z}_i^2}{\hat{\omega}_{i,i}}$$

where $\hat{\mathbf{Z}} = \hat{\mathbf{\Omega}}(\overline{\mathbf{X}} - \overline{\mathbf{Y}}) = (\hat{Z}_1, \ldots, \hat{Z}_p)'$ and $\hat{\mathbf{\Omega}} = (\hat{\omega}_{i,j})$ is a constrained ℓ-1 minimization estimator (Cai et al., 2011) for the precision matrix $\mathbf{\Sigma}^{-1}$. The limiting distribution of T_{CLX} is established based on the extreme value theory.

The test proposed by Cai et al. (2014) is a supremum based test, in contrast to the sum-of-squares based test such as Bai and Saranadasa (1996) and Srivastava et al. (2013). The supremum based test works well especially for sparse but strong signals. On the contrary, sum-of-squares based tests will average over all the signals, including very weak ones. If the data have sparse but strong signals, the strong ones will be dominated by the weak ones that are close to 0. Then the total effect will tend to be weak and not detectable because of the weak majority. Therefore, the sum-of-squares based test is appropriate for data having dense signals.

A critic on the modified distance based tests is that those metrics other than Mahalanobis distance essentially use the diagonal elements in the covariance structure, which ignores the dependency among variables (Lopes et al., 2011). But we should note that the distance is only to measure the separation of means. The dependency, however, can still be automatically incorporated into the test by a scaling parameter. Usually, this scaling parameter is a ratio-consistent estimator for the standard deviation of the distance.

3.2.2 Tests based on projection to subspaces

Another method of bypassing the singular covariance matrix is through dimension reduction. A popular one is random projection first proposed by Lopes et al. (2011) for two-sample mean testing problem. The rationale is to project the data to low-rank subspaces by multiplying a random pseudo-projection matrix $\mathbf{P} \in \mathbb{R}^{k\times p}$. Once the data are in low-dimension, we can compute the Hotelling's T^2 as $T^2 = nm(n+m)^{-1}[\mathbf{P}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})]'(\mathbf{PSP}')^{-1}[\mathbf{P}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})]$. The matrix $\mathbf{PSP}'$ is usually non-singular because it is of a low dimension k. A single random projection test usually has low power, so they suggested an ensemble algorithm: for each run, use a distinct projection matrix $\mathbf{P}_{(b)}$, then compute $T^2_{(b)}$. Repeat B times, then take the average $T_{\text{LJW}} = B^{-1}\sum_{b=1}^{B} T^2_{(b)}$ as the test statistic. Subject to normality assumption, Lopes et al. (2011) gave the critical value based on the F-distribution.

Thulin (2014) proposed an algorithm that randomly selects a subset of covariates for each run instead of using random projection. The p-value is obtained through a permutation procedure. It is noted that when a random projection matrix induces a random subset of covariates, Thulin's algorithm is a special case of Lopes et al. (2011) test. Srivastava et al. (2016) suggested using the average p-value of the random projected T^2 as the test statistic. Its null distribution does not depend on the parameters of the underlying populations. The critical value can be generated numerically. In the Bayesian hypotheses testing framework, Zoh et al. (2018) applied the random projection technique to overcome the singularity of the sample covariance matrix emerging from the Bayes factor.

3.3 Test statistic

Suppose random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^p$ have mean $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ and $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_m$ be two groups of independent identically distributed random samples, with sample sizes n and m, from the distributions of $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}_1 = (X_{11}, \ldots, X_{1p})^\top$ and $\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{1p})^\top$. The hypotheses to be tested are

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs. } H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

The sample sizes are assumed to have the same order so that neither of them will be dominant, that is, $n/(n+m) \to c \in (0,1)$ as $n, m \to \infty$.

The squared univariate two-sample t-statistic for the j^{th} component is $t_j^2 = (\overline{X}_j -$

$\overline{Y}_j)^2/(S_{1j}^2/n + S_{2j}^2/m)$, for $j = 1, 2, \ldots, p$, where $\overline{X}_j$, $\overline{Y}_j$ are the sample means, and S_{1j}^2, S_{2j}^2 are sample variances for the j^{th} component. Averaging over all the t_j^2 gives

$$T_n = p^{-1} \sum_{j=1}^{p} t_j^2. \tag{3.2}$$

The T_n measures the separation of the two groups. The GCT test (Gregory et al., 2015) also constructs the test statistic based on T_n. Next, we will describe the GCT test and discuss its details.

3.3.1 The GCT test

Further centering and scaling on T_n leads to the GCT test statistic $G_n = \sqrt{p}(T_n - \hat{\xi}_n)/\hat{\zeta}_n$, where $\hat{\xi}_n$ estimates the center of T_n, and $\hat{\zeta}_n^2$ is an estimator for $\text{var}(\sqrt{p}T_n)$.

Gregory et al. (2015) showed that $E(T_n) = 1 + n^{-1}a_n + n^{-2}b_n + O(n^{-3})$ under some moment conditions, where $a_n = p^{-1}\sum_{j=1}^{p} c_{nj}$, $b_n = p^{-1}\sum_{j=1}^{p} d_{nj}$, c_{nj} and d_{nj} are functions of the first four moments of X_{1j} and Y_{1j} (see Appendix B.2 for more details). Based on the expression of $E(T_n)$, they proposed two versions of estimators for the center: $\hat{\xi}_n^{(M)} = 1$ and $\hat{\xi}_n^{(L)} = 1 + n^{-1}\hat{a}_n + n^{-2}\hat{b}_n$, where the terms $\hat{a}_n$ and $\hat{b}_n$ are estimators substituting the population parameters in a_n and b_n with their sample estimates. According to different versions for $\hat{\xi}_n$, they define two versions of test statistics, the moderate-p test statistic $G_n^{(M)}$ using $\hat{\xi}_n^{(M)}$, and the large-p version $G_n^{(L)}$ using $\hat{\xi}_n^{(L)}$.

Under α-mixing condition and moment conditions, they showed that

$$\text{var}(\sqrt{p}T_n) = p^{-1} \sum_{j=1}^{p} \sum_{j'=1}^{p} \text{cov}(t_j^2, t_{j'}^2) \to \tau_\infty^2,$$

when n and p go to infinity, where $\tau_\infty^2 = \sum_{k=-\infty}^{\infty} \gamma(k) < \infty$, and

$$\gamma(k) = \lim_{n\to\infty}(p - |k|)^{-1} \sum_{j=1}^{p-|k|} \text{cov}(t_j^2, t_{j+|k|}^2). \tag{3.3}$$

Therefore, the goal of the scaling parameter is to estimate the square root of τ_∞^2. They presented an estimator $\hat{\zeta}_n$ such that

$$\hat{\zeta}_n^2 = \sum_{|k|<L} w\left(\frac{k}{L}\right)\hat{\gamma}(k),$$

where

$$\hat{\gamma}(k) = (p-|k|)^{-1}\sum_{j=1}^{p-|k|}(t_j^2 - T_n)(t_{j+|k|}^2 - T_n) \tag{3.4}$$

is an autocovariance estimator for the sequence of $\{t_1^2, \ldots, t_p^2\}$; $w(.)$ is a window function chosen as the Parzen window (Brockwell and Davis, 1991)

$$w\left(\frac{k}{L}\right) = \begin{cases} 1 - 6\left|\frac{k}{L}\right|^2 + 6\left|\frac{k}{L}\right|^3, \ \left|\frac{k}{L}\right| < 1/2 \\ 2(1-\left|\frac{k}{L}\right|)^3, \ 1/2 \leq \left|\frac{k}{L}\right| \leq 1 \\ 0, \ \left|\frac{k}{L}\right| > 1 \end{cases}$$

or the trapezoid window (Politis and Romano, 1995)

$$w\left(\frac{k}{L}\right) = \begin{cases} 1, \ \left|\frac{k}{L}\right| < 1/2 \\ 2(1-\left|\frac{k}{L}\right|), \ 1/2 \leq \left|\frac{k}{L}\right| \leq 1 \\ 0, \ \left|\frac{k}{L}\right| > 1 \end{cases}$$

and L is a predetermined window width.

Adapting the big-block-little-block technique, the authors briefly showed that $\sup_{x\in\mathbb{R}} |P(T_n - 1) - \Phi\{\sqrt{p}(x - n^{-1}a_n - n^{-2}b_n)/\tau_\infty\}| = o(1)$ where $\Phi(.)$ is the cumulative density function of standard normal distribution.

3.3.2 On the estimator $\hat{\gamma}(k)$

In equation (3.4), $\{t_1^2, \ldots, t_p^2\}$ is treated as a univariate process. The form of $\hat{\gamma}(k)$ coincides with the autocovariance estimation for a stationary process. By comparing $\hat{\gamma}(k)$ and $\gamma(k)$, it is noted that the sample average $(t_j^2 - T_n)(t_{j+|k|}^2 - T_n)$ is used to estimate the average of $\mathrm{cov}(t_j^2, t_{j+|k|}^2)$ for each j. In fact, we can estimate each $\mathrm{cov}(t_j^2, t_{j+|k|}^2)$ individually by directly making use of the sample. The following lemma expresses the covariance for any two squared component-wise t-statistics in terms of the variances and covariances of the data points. The proof of the lemma is given in Appendix B.1.1.

Lemma 1. *Let $N = n + m$, and assume $n/N \to c \in (0,1)$ as $N \to \infty$, $\sup_j E(X_{1j}^4) < \infty$, $\sup_j E(Y_{1j}^4) < \infty$, $\min\{\sigma_{1j}^2, \sigma_{2j}^2\} > 0$ for all $j = 1, \ldots, p$. Then under H_0, $cov(t_j^2, t_{j'}^2) = \gamma_{j,j'} + O(N^{-1/2})$, where*

$$\gamma_{j,j'} = \frac{2(\sigma_{1jj'}/\lambda_1 + \sigma_{2jj'}/\lambda_2)^2}{(\sigma_{1j}^2/\lambda_1 + \sigma_{2j}^2/\lambda_2)(\sigma_{1j'}^2/\lambda_1 + \sigma_{2j'}^2/\lambda_2)}, \tag{3.5}$$

$\lambda_1 = n/N$, $\lambda_2 = m/N$, $\sigma_{1jj'} = cov(X_{1j}, X_{1j'})$, $\sigma_{2jj'} = cov(Y_{1j}, Y_{1j'})$, $\sigma_{1j}^2 = var(X_{1j})$, $\sigma_{2j}^2 = var(Y_{1j})$.

It is worth noting that, $\gamma_{j,j'}$ is always a nonnegative value, which forces the $\gamma(k)$ defined in equation (3.3) to be nonnegative. However, because $(t_j^2 - T_n)(t_{j+|k|}^2 - T_n)$ can take either negative or positive values, one cannot guarantee that $\hat{\gamma}(k) \geq 0$. Naturally, we define a new estimator for $\mathrm{cov}(t_j^2, t_{j'}^2)$ by replacing the covariances and variances in $\gamma_{j,j'}$ with their consistent estimators :

$$\tilde{\gamma}_{j,j'} = \frac{2(S_{1jj'}/\lambda_1 + S_{2jj'}/\lambda_2)^2}{(S_{1j}^2/\lambda_1 + S_{2j}^2/\lambda_2)(S_{1j'}^2/\lambda_1 + S_{2j'}^2/\lambda_2)}, \tag{3.6}$$

where $S_{1jj'} = (n-1)^{-1}\sum_{i=1}^n (X_{ij} - \overline{X}_j)(X_{ij'} - \overline{X}_{j'})$, $S_{2jj'} = (m-1)^{-1}\sum_{i=1}^m (Y_{ij} - \overline{Y}_j)(Y_{ij'} - \overline{Y}_{j'})$, $S_{1j}^2 = (n-1)^{-1}\sum_{i=1}^n (X_{ij} - \overline{X}_j)^2$ and $S_{2j}^2 = (m-1)^{-1}\sum_{i=1}^m (Y_{ij} - \overline{Y}_j)^2$. It is obvious that $\tilde{\gamma}_{j,j'}$ always takes nonnegative values as it is expected to.

3.3.3 On the scaling parameter

In the course of establishing $G_n \xrightarrow{d} N(0,1)$, the consistency of $\hat{\zeta}_n^2$ is needed but not shown by Gregory et al. (2015). The lag-window construction of $\hat{\zeta}_n$ implicitly uses the spectral density estimation, whose consistency is well studied under stationarity (Brockwell and Davis, 1991). With the assumptions given by Gregory et al. (2015), lacking stationary, the proof of consistency is not trivial but was not provided. In fact, there are some potential problems with $\hat{\zeta}_n$. Gregory et al. (2015) did not discuss the choice of L, but $\hat{\zeta}_n \xrightarrow{p} \tau_\infty$ is not valid for arbitrary L. To illustrate this, note that

$$\begin{aligned}\hat{\zeta}_n^2 - \tau_\infty^2 &= \sum_{|k|<L} w(k/L)\hat{\gamma}(k) - \sum_{k=-\infty}^{\infty} \gamma(k) \\ &= \underbrace{\sum_{|k|<L} w(k/L)[\hat{\gamma}(k) - \gamma(k)]}_{\text{(I)}} + \underbrace{\sum_{|k|<L} [w(k/L) - 1]\gamma(k)}_{\text{(II)}} - \underbrace{\sum_{|k|\geq L} \gamma(k)}_{\text{(III)}}.\end{aligned} \tag{3.7}$$

If L is finite, (I) $\xrightarrow{p} 0$ by the asymptotic normality of $\hat{\gamma}(k)$ (see details in Appendix B.1.5). Recall that $\gamma_{j,j'}$ is non-negative, so is $\gamma(k)$. Therefore, (II) will always be a non-positive value since $w(k/L) \leq 1$ by the definition of the window functions, and (III) is a finite value not equal to 0, (II) – (III) will not converge to 0 unless $\gamma(k) = 0$ for all k.

When L increases with p, the term (III) in equation (3.7) goes to 0 as it was shown by Gregory et al. (2015). But term (II) is not beneficial because it induces bias. Since $\gamma(k)$ is non-negative for all k, the bias will accumulate as L grows.

In sight of the undesirable properties of $\hat{\zeta}_n$, we define a new scaling parameter $\tilde{\zeta}_n$ such that

$$\tilde{\zeta}_n^2 = p^{-1} \sum_{|j-j'|\leq L} \tilde{\gamma}_{j,j'}, \tag{3.8}$$

where j and $j' = 1, \ldots, p$, $L = \lfloor p^\epsilon \rfloor$ for some $0 < \epsilon < 1$. Let $\tilde{\gamma}(k) = p^{-1} \sum_{j=1}^{p-|k|} \tilde{\gamma}_{j,j+|k|}$, it is easy to see $\tilde{\zeta}_n^2 = \sum_{|k|<L} \tilde{\gamma}(k)$. Hence, $\tilde{\zeta}_n^2$ amounts to a modification on $\hat{\zeta}_n^2$ by replacing $\hat{\gamma}(k)$ with $\tilde{\gamma}(k)$ and letting $w(k/L) = 1$ for $k \leq L$. Then the new scaling parameter can benefit

from the good properties of $\tilde{\gamma}(k)$ and avoid the biases induced by term (II) in equation (3.7).

3.3.4 New test statistic

Now we can define our new test statistic. For small sample size, we suggest to use the test statistic

$$J_{1n} = p^{1/2}(T_n - 1)/\tilde{\zeta}_n, \tag{3.9}$$

and for large sample size, we suggest

$$J_{2n} = p^{1/2}[T_n - (1 + n^{-1}\hat{a}_n + n^{-2}\hat{b}_n)]/\tilde{\zeta}_n. \tag{3.10}$$

where T_n is defined in equation (3.2), $\tilde{\zeta}_n$ is defined by equation (3.8), $\hat{a}_n = p^{-1}\sum_{j=1}^{p}\hat{c}_{nj}$, $\hat{b}_n = p^{-1}\sum_{j=1}^{p}\hat{d}_{nj}$, $\hat{c}_{nj}$ and $\hat{d}_{nj}$ are given in Appendix B.2 for brevity.

The distinction between two test statistics is made similarly as the moderate-p and large-p versions of the GCT test. Gregory et al. (2015) suggested to use J_{1n} for the case when $p = o(n^2)$ and J_{2n} for $p = o(n^6)$. Theoretically, J_{2n} also works for the small sample sizes. However, J_{2n} contains the high-order correction for the center, which requires sample estimators for the third and fourth moments. Estimating those moments often requires relatively large sample sizes.

3.4 Main results

This section will show some technical results relevant to the new test statistic. Dependency potentially exists among the variables. Here we consider the α-mixing condition, a weak dependence structure widely used in studies (Gregory et al., 2015; Xu et al., 2016).

3.4.1 α-mixing condition and covariance inequalities

For a sequence of random variables $\{z_j : j = 1, 2, \ldots\}$, let $\mathcal{F}_a^b$ be the σ-field generated by $\{z_j : a \leq j \leq b\}$. Define the strong mixing coefficient as

$$\alpha(r) = \sup_{t \geq 1}\{|P(A \cup B) - P(A)P(B)| : A \in \mathcal{F}_1^t, B \in \mathcal{F}_{t+r}^{\infty}\}. \tag{3.11}$$

A sequence is α-mixing if $\alpha(r) \to 0$ as $r \to \infty$. $\alpha(r)$ evaluates the strength of dependence between two σ-fields. When $z_1, z_2, \ldots$ are independent with each other, $\alpha(r) = 0$. Two variables in an α-mixing sequence tend to become independent as the distance between them increases to infinity. A common technique for α-mixing sequence is the covariance inequalities. Suppose $\{z_j : j = 1, 2, \ldots\}$ is α-mixing, $E(z_t^s) < \infty$, $E(z_{t+r}^q) < \infty$, and $1/s + 1/q < 1$ for $s > 0$, $q > 0$. Davydov (1968) showed that

$$|\text{cov}(z_t, z_{t+r})| \leq 12[\alpha(r)]^{1-1/s-1/q}(Ez_t^s)^{1/s}(Ez_{t+r}^q)^{1/q}. \tag{3.12}$$

Billingsley (1995, pp. 365, Lemma 3) showed that if $E(z_t^4) \leq C_1$ and $E(z_{t+r}^4) \leq C_2$ for some $C_1, C_2 < \infty$, then

$$|\text{cov}(z_t, z_{t+r})| \leq 8(1 + C_1 + C_2)[\alpha(r)]^{1/2}. \tag{3.13}$$

Let $s = q = 4$, then the inequality (3.12) reduces to

$$|\text{cov}(z_t, z_{t+r})| \leq 12(Ez_t^4)^{1/4}(Ez_{t+r}^4)^{1/4}[\alpha(r)]^{1/2},$$

which provides similar upper-bound and requires finite fourth moments as inequality (3.13).

These inequalities are useful in finding the upperbound for the variance of partial sums of the α-mixing sequence. For example, the term $\sum_{|k|\geq L}\gamma(k)$ in equation (3.7) is a sum of covariances. If $\sup_{1\leq j\leq p} E(t_j^8) < \infty$, then by (3.12) or (3.13) we have

$$\gamma(k) = \lim_{n\to\infty}(p-|k|)^{-1}\sum_{j=1}^{p-|k|}\operatorname{cov}(t_j^2, t_{j+|k|}^2) \leq \lim_{n\to\infty}\sup_{1\leq j\leq p}|\operatorname{cov}(t_j^2, t_{j+|k|}^2)| \leq C[\alpha(|k|)]^{1/2},$$

for some constant C. If we further impose $\sum_{k=1}^{\infty}[\alpha(k)]^{1/2} < \infty$, then $\sum_{k=-\infty}^{\infty}\gamma(k) < \infty$ and $\sum_{|k|\geq L}\gamma(k) \to 0$ as $L \to \infty$. This property will be used later.

3.4.2 Results under the null hypothesis

We state several lemmas before establishing the sampling distribution for our proposed test statistics. Lemma 2 states that the sequence of vectors binding α-mixing sequences is also an α-mixing sequence with the same mixing coefficient. It is a special case of Lemma 4.4.2 in Wang (2004), so the proof is omitted here.

Lemma 2. *Suppose $\{X_{ij}; j = 1, 2\ldots\}$ for $i = 1, \ldots n$ and $\{Y_{ij}; j = 1, 2, \ldots\}$ for $i = 1, \ldots, m$ are α-mixing sequences of random variables having strong mixing coefficient $\alpha(r)$. Let $\mathbf{Z}_j$ be the vector binding all the observations at component j, i.e., $\mathbf{Z}_j = \{X_{1j}, \ldots, X_{nj}, Y_{1j}, \ldots, Y_{mj}\}'$. Then $\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots\}$ is an α-mixing sequence of random vectors having strong mixing coefficient $\alpha(r)$.*

Noting that $t_j^2 = (\overline{X}_j - \overline{Y}_j)^2/(S_{1j}^2/n + S_{2j}^2/m)$ is a measurable function of $\mathbf{Z}_j$ in Lemma 2, we have the following lemma. The proof is deferred to Appendix B.1.2.

Lemma 3. *Suppose $\{X_{ij}; j = 1, 2\ldots\}$ for $i = 1, \ldots n$ and $\{Y_{ij}; j = 1, 2, \ldots\}$ for $i = 1, \ldots, m$ are α-mixing sequences of random variables having strong mixing coefficient $\alpha(r)$. Then $\{t_1^2, t_2^2, \ldots\}$is an α-mixing sequence having strong mixing coefficient $\alpha(r)$.*

The lemma below, shown by Wang and Akritas (2010), is a central limit theorem for the partial sum of an α-mixing sequence.

Lemma 4 (Wang and Akritas, 2010)**.** *Suppose that* $z_1, z_2, \ldots$ *is* α*-mixing with* $\alpha(r) = O(r^{-5})$, $E(z_j) = 0$, *and* $\limsup_j E(z_j^{16}) < \infty$. *Let* $S_p = \sum_{j=1}^p z_j$. *If* $\lim_{p\to\infty} var(S_p)/p$ *exists and is greater than 0, then* $S_p/\sqrt{var(S_p)} \xrightarrow{p} \mathcal{N}(0,1)$.

We will use Lemma 4 to show the asymptotic normality of T_n by letting $z_j = t_j^2 - E(t_j^2)$. Then the partial sum S_p corresponds to $p(T_n - ET_n)$, and $\text{var}(S_p)/p$ corresponds to $\text{var}(\sqrt{p}T_n)$. So the condition on variance is converted to $0 < \lim_{N,p\to\infty} \text{var}(\sqrt{p}T_n) < \infty$, which is justified by the following lemma. The proof can be found in Appendix B.1.3.

Lemma 5. *Under* H_0*, suppose* $\limsup_j E(t_j^{4+2\nu}) < \infty$ *and* $\sum_{r=1}^\infty [\alpha(r)]^{\nu/(2+\nu)} < \infty$ *for some* $\nu > 0$, $\min\{\sigma_{1j}^2, \sigma_{2j}^2\} > 0$ *for all* $j = 1, \ldots, p$. *Then* $0 < \lim_{N,p\to\infty} var(\sqrt{p}T_n) < \infty$.

Regularity conditions (C.1)-(C.3) on the strong mixing coefficient and moments of data are stated below.

(C.1) Both $\{X_{ij}; j = 1, 2\ldots\}$ and $\{Y_{ij}; j = 1, 2, \ldots\}$ have strong mixing coefficient $\alpha(r)$ defined by (3.11) satisfies $\sum_{r=1}^\infty r[\alpha(r)]^{\nu/(2+\nu)} < \infty$ for some $\nu > 0$.

(C.2) $t_j = (\overline{X}_j - \overline{Y}_j)/(S_{1j}^2/n + S_{2j}^2/m)^{1/2}$, for $j = 1, 2, \ldots, p$, $\limsup_j E(t_j^{12+6\nu}) < \infty$ for some $\nu > 0$.

(C.3) $\min\{\sigma_{1j}^2, \sigma_{2j}^2\} > 0$ for all $j = 1, \ldots, p$.

The following theorem gives the limiting distribution of T_n. The proof is deferred to Appendix B.1.4.

Theorem 3. *Under* H_0*, suppose* $\{X_{ij}; j = 1, 2\ldots\}$ *for* $i = 1, \ldots n$ *and* $\{Y_{ij}; j = 1, 2, \ldots\}$ *for* $i = 1, \ldots, m$ *are sequences of random variables satisfying condition (C.1)-(C.3) for the same* ν. *If* $T_n = p^{-1}\sum_{j=1}^p t_j^2$, *then*

$$\frac{\sqrt{p}[T_n - E(T_n)]}{[var(\sqrt{p}T_n)]^{1/2}} \xrightarrow{d} \mathcal{N}(0,1), \ \ as\ p \to \infty. \tag{3.14}$$

Remark 1.

(i) A sufficient condition for $\sum_{r=1}^{\infty} r[\alpha(r)]^{\nu/(2+\nu)} < \infty$ in (C.1) is $\alpha(r) = O(r^{-c})$ for some $c > 2+4/\nu$. For example, taking $\nu = 2$, a set of sufficient conditions are $\limsup_j E(t_j^{24}) < \infty$ and $\alpha(r) = O(r^{-c})$ for $c > 4$. If taking $\nu = 1$, a set of sufficient conditions is $\limsup_j E(t_j^{18}) < \infty$ and $\alpha(r) = O(r^{-c})$ for $c > 6$.

(ii) There is a trade-off between the moment condition (C.2) and the convergence rate of the mixing coefficient $\alpha(r)$ in (C.1). Smaller ν will relax the moment condition, but it will require faster convergence of $\alpha(r)$. To see this, note that $\alpha(r) < 1$ and $\sum_{r=1}^{\infty} r[\alpha(r)]^{\nu_1/(2+\nu_1)} > \sum_{r=1}^{\infty} r[\alpha(r)]^{\nu_2/(2+\nu_2)}$ for $\nu_1 < \nu_2$. For example, taking $\nu = 2/3$, a set of sufficient conditions is $\limsup_j E(t_j^{16}) < \infty$ and $\alpha(r) = O(r^{-c})$ for $c > 8$. Meanwhile, as pointed out in (i), $\nu = 2$ corresponds to $c > 4$, where the convergence can be slower.

(iii) The moment condition (C.2) on t_j^2 can be converted to conditions directly on the observations X_j and Y_j. This is established in the following proposition. The proof can be found in Appendix B.1.6.

Proposition 2. *If* $\sup_j EX_j^{2k+2} < \infty$ *and* $\sup_j EY_j^{2k+2} < \infty$*, then* $\sup_j Et_j^{2k} < \infty$ *for any* $k > 1$.

(iv) The central limit theorem claimed in Theorem 3 can also be proved with other assumptions. For example, following the Theorem 16.3.5 in Athreya and Lahiri (2006), one also can show (3.14) if $\sup_{t\geq 1}(E|t_j^2|^{2+\nu})^{1/(2+\nu)} < \infty$ and $\sum_{r=1}^{\infty}[\alpha(r)]^{\nu/(2+\nu)} < \infty$ for some $\nu > 0$, and there exists $M_0 \in (0,\infty)$ and a function $\tau(.) : (M_0,\infty) \to (0,\infty)$ such that for all $M > M_0$,

$$\sup_j \left| p^{-1}\text{var}\left(\sum_{j=k}^{k+p-1} t_j^2 I(|t_j^2| < M)\right) - \tau(M)\right| \to 0 \text{ as } p \to \infty. \tag{3.15}$$

If further assume that the $\{t_1^2, t_2^2, \ldots\}$ sequence is stationary, the assumption in (3.15) can be replaced by $\lim_{p\to\infty} \text{var}(\sqrt{p}T_n) > 0$. A similar asymptotic normality of T_n was presented by Gregory et al. (2015) without proof.

The central limit theorem established above involves population quantities. In our test statistic, the center parameter $E(T_n)$ is estimated by 1 or $1+n^{-1}\hat{a}_n+n^{-2}\hat{b}_n$, and the scaling parameter $[\text{var}(\sqrt{p}T_n)]^{1/2}$ is estimated by $\tilde{\zeta}_n$. If the estimators for the center and scaling parameters are close enough to the true values, then the asymptotic normality of the test statistic follows. We will show the consistency of $\tilde{\zeta}_n$ in Theorem 4 and further discuss the center estimator in the proof of Corollary 4.

Recall that $\tilde{\zeta}_n^2 = p^{-1}\sum_{|j-j'|\leq L}\tilde{\gamma}_{j,j'}$. We first study the consistency of $\tilde{\gamma}_{j,j'}$ in the following lemma, whose proof is given in Appendix B.1.7.

Lemma 6. *Suppose $\gamma_{j,j'}$ and $\tilde{\gamma}_{j,j'}$ are defined as equations (3.5) and (3.6). Let $N=m+n$, $0<m/n<\infty$, $n/N\to c\in(0,1)$ as $N\to\infty$. Assume $\sup_j E(X_{1j}^2)<\infty$, $\sup_j E(Y_{1j}^2)<\infty$, and $\min\{\sigma_{1j}^2,\sigma_{2j}^2\}>0$ for all $j=1,\ldots,p$. Then under H_0, $\tilde{\gamma}_{j,j'}=\gamma_{j,j'}+O_p(N^{-1/2})$ for any given $j,\ j'\in\{1,\ldots,p\}$.*

Theorem 4 shows the consistency of $\tilde{\zeta}_n$. The proof is shown in Appendix B.1.8.

Theorem 4. *Suppose $\{X_j; j=1,\ldots,p\}$ and $\{Y_j; j=1,\ldots,p\}$ are sequences having the strong mixing coefficient $\alpha(r)$, condition (C.3) holds, $\sup_j E(t_j^{4+2\nu})<\infty$ for some $\nu>0$, and $L=\lfloor p^{\epsilon}\rfloor$ for some $0<\epsilon<1$.*

(i) If condition (C.1) holds, then $\tilde{\zeta}_n^2 - var(\sqrt{p}T_n) = O_p(N^{-1/2})+O(L^{-1})$ as N and p go to infinity.

(ii) If $\alpha(r)=O(r^{-h})$ for some $h>1$ and $\nu>2/(h-1)$, then $\tilde{\zeta}_n^2 - var(\sqrt{p}T_n) = O_p(N^{-1/2})+O(p^{1-h\nu/(2+\nu)})$ as N and p go to infinity.

Now, we are ready to show the sampling distribution of the test statistics in the following corollary. The proof is shown in Appendix B.1.9.

Corollary 4. *Suppose $\{X_{ij}; j=1,2\ldots\}$ for $i=1,\ldots n$ and $\{Y_{ij}; j=1,2,\ldots\}$ for $i=1,\ldots,m$ are sequences of random variables satisfying conditions (C.1)-(C.3) for the same ν. J_{n1} and J_{n2} are defined as (3.9) and (3.10). $L=\lfloor p^{\epsilon}\rfloor$ for some $0<\epsilon<1$. Under H_0, as N and p go to infinity, if $p=o(N^2)$ then $J_{1n}\xrightarrow{d}\mathcal{N}(0,1)$; if $p=o(N^6)$ then $J_{2n}\xrightarrow{d}\mathcal{N}(0,1)$.*

3.4.3 Results under the local alternative hypothesis

The CLT stated in Corollary 4 concerns about the property of the test statistic under H_0. In the following, we will give the relevant results under H_a. We will only consider the test statistic $J_{2n} = \sqrt{p}(T_n - \hat{\xi}_n)/\tilde{\zeta}_n$ because J_{1n} shares much similarity with it. Suppose the population means have relationship $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}$, where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)'$. The hypotheses can be written in terms of $\boldsymbol{\delta}$: $H_0 : \boldsymbol{\delta} = 0$ vs $H_a : \boldsymbol{\delta} \neq 0$. To simplify the derivation, we assume that all the δ_j values for $j = 1, \ldots, p$ are of the same order in terms of N, and $\max_j |\delta_j| = O(N^{-k})$ for some k. Recalling that the condition for the test statistic J_{2n} being valid is $p = o(N^6)$, so we assume $p = O(N^a)$ for $0 < a < 6$.

Suppose X_{ij} and Y_{ij} were observed under H_a. Define $X_{ij}^{(0)} = X_{ij} - EX_{1j}$ for $i = 1, \ldots, n$, $Y_{ij}^{(0)} = Y_{ij} - EY_{1j}$ for $i = 1, \ldots, m$, $\overline{X}_j^{(0)} = n^{-1}\sum_{i=1}^{n} X_{ij}^{(0)}$, and $\overline{Y}_j^{(0)} = m^{-1}\sum_{i=1}^{m} Y_{ij}^{(0)}$. Noting that $EX_{1j} - EY_{1j} = \delta_j$, then T_n under H_a, denoted by $T_n^{(1)}$, can be expressed as

$$T_n^{(1)} = \frac{1}{p}\sum_{j=1}^{p} \frac{(\overline{X}_j - \overline{Y}_j)^2}{\dfrac{S_{1j}^2}{n} + \dfrac{S_{2j}^2}{m}} = \frac{1}{p}\sum_{j=1}^{p} \frac{(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)} + \delta_j)^2}{\dfrac{S_{1j}^2}{n} + \dfrac{S_{2j}^2}{m}} = T_n^{(0)} + \Gamma_{1,n,p} + \Gamma_{2,n,p},$$

where $T_n^{(0)} \equiv p^{-1}\sum_{j=1}^{p}[t_j^{(0)}]^2$,

$$t_j^{(0)} \equiv \frac{\overline{X}_j^{(0)} - \overline{Y}_j^{(0)}}{\left(\dfrac{S_{1j}^2}{n} + \dfrac{S_{2j}^2}{m}\right)^{\frac{1}{2}}}, \ \Gamma_{1,n,p} \equiv \frac{1}{p}\sum_{j=1}^{p} \frac{2n(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})\delta_j}{S_{1j}^2 + S_{2j}^2(n/m)}, \ \Gamma_{2,n,p} \equiv \frac{1}{p}\sum_{j=1}^{p} \frac{n\delta_j^2}{S_{1j}^2 + S_{2j}^2(n/m)}. \tag{3.16}$$

In fact, $T_n^{(0)}$ is equivalent to T_n under H_0 because $\overline{X}_j^{(0)} - \overline{Y}_j^{(0)}$ has zero mean, and S_{1j}^2 and S_{2j}^2 are not affected by the centers of X_{ij} and Y_{ij}. This fact ensures that the results for T_n under H_0 are also valid for $T_n^{(0)}$ under H_a.

Due to the fact that $\sqrt{p}[\hat{\xi}_n - E(T_n^{(0)})] \xrightarrow{p} 0$ and $\tilde{\zeta}_n \xrightarrow{p} \zeta^{(0)} \equiv [\mathrm{var}(\sqrt{p}T_n^{(0)})]^{1/2}$, in conjunction with the Delta method, the cumulative distribution function for J_{2n} can be expressed as

$$\begin{aligned}
P_{H_a}(J_{2n} < x) =\ & P_{H_a}(\sqrt{p}(T_n^{(1)} - \hat{\xi}_n)/\tilde{\zeta}_n < x) \\
=\ & P_{H_a}(\sqrt{p}(T_n^{(1)} - E(T_n^{(0)}))/\zeta^{(0)} < x) + o(1) \qquad (3.17) \\
=\ & P_{H_a}\left(\frac{\sqrt{p}[T_n^{(1)} - E(T_n^{(1)})]}{\zeta^{(1)}} \frac{\zeta^{(1)}}{\zeta^{(0)}} + \frac{\sqrt{p}[E(T_n^{(1)}) - E(T_n^{(0)})]}{\zeta^{(0)}} < x\right) + o(1) \\
=\ & P_{H_a}\left(\frac{\sqrt{p}[T_n^{(1)} - E(T_n^{(1)})]}{\zeta^{(1)}} \frac{\zeta^{(1)}}{\zeta^{(0)}} + \frac{\sqrt{p}[E(\Gamma_{1,n,p}) + E(\Gamma_{2,n,p})]}{\zeta^{(0)}} < x\right) + o(1) \qquad (3.18)
\end{aligned}$$

where $\zeta^{(1)} \equiv [\mathrm{var}(\sqrt{p}T_n^{(1)})]^{1/2}$.

According to (3.18), the asymptotic behavior of the CDF relies on the magnitudes of $\zeta^{(1)}$, $E(\Gamma_{1,n,p})$ and $E(\Gamma_{2,n,p})$. Recall that the variance of $\sqrt{p}T_n$ is $p^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathrm{cov}(t_j^2, t_{j'}^2)$. Lemma 1 has shown the expression for the covariance term under H_0. With some algebra (details are deferred to Appendix B.3), we can show that

$$\begin{aligned}
(\zeta^{(1)})^2 = (\zeta^{(0)})^2 &+ 4p^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}\delta_j c_{jj'}\left[\frac{E[X_{1j}^{(0)}(X_{1j'}^{(0)})^2]}{\lambda_1^2} - \frac{E[Y_{1j}^{(0)}(Y_{1j'}^{(0)})^2]}{\lambda_2^2}\right] \\
&+ 4Np^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}\delta_j\delta_{j'}c_{jj'}\left[\frac{\sigma_{1jj'}}{\lambda_1} + \frac{\sigma_{2jj'}}{\lambda_2}\right],
\end{aligned} \qquad (3.19)$$

where $c_{jj'} = (\sigma_{1j}^2/\lambda_1 + \sigma_{2j}^2/\lambda_2)^{-1}(\sigma_{1j'}^2/\lambda_1 + \sigma_{2j'}^2/\lambda_2)^{-1}$. It follows that

$$\begin{aligned}
|(\zeta^{(1)})^2 - (\zeta^{(0)})^2| \leq & 4\max_j|\delta_j|p^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}c_{jj'}\left|\frac{E[X_{1j}^{(0)}(X_{1j'}^{(0)})^2]}{\lambda_1^2} - \frac{E[Y_{1j}^{(0)}(Y_{1j'}^{(0)})^2]}{\lambda_2^2}\right| \\
&+ 4N\max_j\delta_j^2 p^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}c_{jj'}\left|\frac{\sigma_{1jj'}}{\lambda_1} + \frac{\sigma_{2jj'}}{\lambda_2}\right|.
\end{aligned} \qquad (3.20)$$

By α-mixing condition and inequality (3.12), we have $|E[X_{1j}^{(0)}(X_{1j'}^{(0)})^2]| = |\mathrm{cov}[X_{1j}^{(0)}, (X_{1j'}^{(0)})^2]| \leq K[\alpha(j-j')]^{1/2}$, and $|\sigma_{1jj'}| \leq K[\alpha(j-j')]^{1/2}$ for some constant K. Similar property applies to $|E[Y_{1j}^{(0)}(Y_{1j'}^{(0)})^2]|$ and $|\sigma_{2jj'}|$. Hence the two double summations in (3.20) is of order $O(p)$.

Since $(\zeta^{(0)})^2 < \infty$ by Lemma 5, we have $(\zeta^{(1)})^2 = (\zeta^{(0)})^2 + O(N \max_j \delta_j^2)$, and

$$\zeta^{(1)} = \zeta^{(0)} + O(N^{1/2} \max_j |\delta_j|) = \zeta^{(0)} + O(N^{1/2-k}). \tag{3.21}$$

In the calculation of $\sqrt{p}E(\Gamma_{1,n,p})$ and $\sqrt{p}E(\Gamma_{2,n,p})$, it is common to encounter computing expectation of random variable $z_N = O_p(N^b)$. Lemma 7, together with some proper moment constraint, allows us to assert that $E(z_N) = O(N^b)$. Proof is shown in Appendix B.1.10.

Lemma 7. *Suppose X_n is a sequence of random variables such that $X_n \xrightarrow{d} X$. If for some $b > 0$, $\limsup_n E|X_n|^b < \infty$. Then $E(X_n^k) \to E(X^k)$ for $0 < k < b$.*

Applying Taylor expansion and Lemma 7, we can write

$$\begin{aligned}
& \sqrt{p}E(\Gamma_{1,n,p}) \\
= \ & \frac{2n}{p^{1/2}} \sum_{j=1}^{p} E \tfrac{\delta_j (\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})}{\sigma_{1j}^2 + \sigma_{2j}^2 (n/m) + [S_{1j}^2 - \sigma_{1j}^2 + (S_{2j}^2 - \sigma_{2j}^2)(n/m)]} \\
= \ & \frac{2n}{p^{1/2}} \sum_{j=1}^{p} E \tfrac{\delta_j (\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})}{\sigma_{1j}^2 + \sigma_{2j}^2 (n/m)} \left[1 - \tfrac{S_{1j}^2 - \sigma_{1j}^2 + (S_{2j}^2 - \sigma_{2j}^2)(n/m)}{\sigma_{1j}^2 + \sigma_{2j}^2 (n/m)} \right] + O(N^{-1} p^{1/2} \max_j |\delta_j| D_2) \\
= \ & \frac{-2n}{p^{1/2}} \sum_{j=1}^{p} E \tfrac{\delta_j (\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})(S_{1j}^2 - \sigma_{1j}^2 + (S_{2j}^2 - \sigma_{2j}^2)(n/m))}{(\sigma_{1j}^2 + \sigma_{2j}^2 (n/m))^2} + O(N^{-1} p^{1/2} \max_j |\delta_j| D_2) \\
= \ & p^{1/2} \max_j |\delta_j| D_1 + O(N^{-1} p^{1/2} \max_j |\delta_j|) \qquad (3.22)
\end{aligned}$$

where

$$D_1 \equiv \frac{-2}{p} \sum_{j=1}^{p} \frac{(\delta_j / \max_j |\delta_j|)[E(X_{1j}^{(0)})^3 - E(Y_{1j}^{(0)})^3 (n/m)^2]}{(\sigma_{1j}^2 + \sigma_{2j}^2 (n/m))^2}, \tag{3.23}$$

$$D_2 \equiv \frac{2N}{np} \sum_{j=1}^{p} E \left(\tfrac{(\delta_j / \max_j |\delta_j|)(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})}{\sigma_{1j}^2 + \sigma_{2j}^2 (n/m)} \left[\tfrac{S_{1j}^2 - \sigma_{1j}^2 + (S_{2j}^2 - \sigma_{2j}^2)(n/m)}{\sigma_{1j}^2 + \sigma_{2j}^2 (n/m)} \right]^2 \right) = \frac{2N}{np} \sum_{j=1}^{p} \frac{(\delta_j / \max_j |\delta_j|) g_j}{(\sigma_{1j}^2 + \sigma_{2j}^2 (n/m))^3},$$

and

$$
\begin{aligned}
g_j &\equiv n^2 E\left[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})(S_{1j}^2 - \sigma_{1j}^2 + (S_{2j}^2 - \sigma_{2j}^2)(n/m))^2\right] \\
&= n^2(n-1)^{-2}[(-6+16n^{-1}-10n^{-2})\sigma_{1j}^2 E(X_{1j}^{(0)})^3 + (1-2n^{-1}+n^{-2})E(X_{1j}^{(0)})^5] \\
&\quad -n^2(m-1)^{-2}[(-6+16m^{-1}-10m^{-2})\sigma_{2j}^2 E(Y_{1j}^{(0)})^3 + (1-2m^{-1}+m^{-2})E(Y_{1j}^{(0)})^5].
\end{aligned}
$$

Note that $p^{1/2}\max_j|\delta_j|D_1 = O(N^{a/2-k})$ and $N^{-1}p^{1/2}\max_j|\delta_j| = O(N^{a/2-k-1})$. This fact will be used in later discussion. We also have

$$
\begin{aligned}
\sqrt{p}E(\Gamma_{2,n,p}) &= \frac{n}{p^{1/2}}\sum_{j=1}^{p} E\frac{\delta_j^2}{\sigma_{1j}^2+\sigma_{2j}^2(n/m)+[S_{1j}^2-\sigma_{1j}^2+(S_{2j}^2-\sigma_{2j}^2)(n/m)]} \\
&= \frac{n}{p^{1/2}}\sum_{j=1}^{p}\frac{\delta_j^2}{\sigma_{1j}^2+\sigma_{2j}^2(n/m)}E[1+O_p(S_{1j}^2-\sigma_{1j}^2+(S_{2j}^2-\sigma_{2j}^2)(n/m))] \\
&= \frac{n\max_j\delta_j^2}{p^{1/2}}\sum_{j=1}^{p}\frac{(\delta_j/\max_j\delta_j)^2}{\sigma_{1j}^2+\sigma_{2j}^2(n/m)}(1+o(1)) \qquad (3.24) \\
&= O\left(p^{1/2}N\max_j\delta_j^2\right) \\
&= O(N^{1+a/2-2k}),
\end{aligned}
$$

where the equality in (3.24) is due to Lemma 7 as well.

The values of k and a will affect the limiting behavior of the CDF. Indeed, k reflects the magnitude of the effect size, and a bigger k corresponds to a smaller effect size. In the following, we will discuss the CDF of J_{2n} separately for $k \geq 1/2$ and $k < 1/2$.

- **Case 1:** $k \geq 1/2$.

  In this case, $\zeta^{(1)} = \zeta^{(0)} + O(N^{1/2-k}) < \infty$ as $N \to \infty$ by (3.21). Invoking the CLT for $T_n^{(1)}$, (3.18) becomes

$$
P_{H_a}(J_{2n} < x) = P_{H_a}\left(\frac{\zeta^{(1)}}{\zeta^{(0)}}Z + \frac{\sqrt{p}[E(\Gamma_{1,n,p})+E(\Gamma_{2,n,p})]}{\zeta^{(0)}} < x\right) + o(1)
$$

where Z is a random variable following standard normal distribution. Recall from (3.22) that the magnitude of $\sqrt{p}E(\Gamma_{1,n,p})$ is affected by whether D_1 in (3.23) vanishes or not. Specifically, when $D_1 \neq 0$, $\sqrt{p}E(\Gamma_{1,n,p}) = O(N^{a/2-k})$; when $D_1 = 0$ but $D_2 \neq 0$, $\sqrt{p}E(\Gamma_{1,n,p}) = O(N^{a/2-k-1})$.

(i) Suppose $D_1 \neq 0$.

- If $k > \max(a/2, 1/2 + a/4)$, then $\sqrt{p}E(\Gamma_{1,n,p}) \to 0$, $\sqrt{p}E(\Gamma_{2,n,p}) \to 0$ and $\zeta^{(1)}/\zeta^{(0)} \to 1$ (refer to equation (3.20) and $\zeta^{(0)} > 0$ by Lemma 5), and the CDF reduces to $\Phi(x) + o(1)$.
- If $1/2 \leq k < 1/2 + a/4$, then $\sqrt{p}E(\Gamma_{2,n,p})$ is the dominant term and goes to ∞. The CDF goes to 0.
- If $a > 2$ and $1/2 + a/4 \leq k < a/2$, then $\sqrt{p}E(\Gamma_{1,n,p})$ is the dominant term and goes to ∞ when $D_1 > 0$ and $-\infty$ when $D_1 < 0$, and the CDF goes to 0 and 1, respectively.
- If $k = \max(a/2, 1/2 + a/4)$, let $\Delta_1 = \lim_{N\to\infty} \sqrt{p}[E(\Gamma_{1,n,p}) + E(\Gamma_{2,n,p})]/\zeta^{(0)}$, then $P_{H_a}(J_{2n} < x) \to \Phi(x - \Delta_1)$.

(ii) Suppose $D_1 = 0$. Then $\sqrt{p}E(\Gamma_{1,n,p}) = O(N^{a/2-k-1})$ and $\sqrt{p}E(\Gamma_{2,n,p}) = O(N^{1+a/2-2k})$. Because $0 < a < 6$, we always have $a/4 + 1/2 > a/2 - 1$.

- If $k > a/4 + 1/2$, then $\sqrt{p}E(\Gamma_{1,n,p}) \to 0$, $\sqrt{p}E(\Gamma_{2,n,p}) \to 0$ and $\zeta^{(1)}/\zeta^{(0)} \to 1$, and the CDF goes to $\Phi(x)$.
- If $1/2 \leq k < 1/2 + a/4$, then $\sqrt{p}E(\Gamma_{2,n,p})$ is the dominant term and goes to ∞. The CDF goes to 0.
- If $k = 1/2 + a/4$, $\lim_{N\to\infty} \sqrt{p}E(\Gamma_{1,n,p}) \to 0$. Let $\Delta_2 = \lim_{N\to\infty} \sqrt{p}E(\Gamma_{2,n,p})/\zeta^{(0)}$. Then $P_{H_a}(J_{2n} < x) \to \Phi(x - \Delta_2)$.

- **Case 2:** $k < 1/2$.

In this case, $(\zeta^{(1)})^2$ possibly goes to infinity, and thus the CLT for $T_n^{(1)}$ does not necessarily hold. For example, when $\delta_j = N^{-1/2}p^b$ for some $b > 0$ for all $j = 1, \ldots, p$, and $\kappa \equiv p^{-1} \sum_{j=1}^p \sum_{j'=1}^p C_{jj'} (\sigma_{1jj'}/\lambda_1 + \sigma_{2jj'}/\lambda_2) > 0$, it is easy to see from (3.19) that

$$(\zeta^{(1)})^2 = (\zeta^{(0)})^2 + 4N^{-1/2}p^{b-1} \sum_{j=1}^p \sum_{j'=1}^p C_{jj'} \left[\frac{E[X_{j1}^{(0)}(X_{j'1}^{(0)})^2]}{\lambda_1^2} - \frac{E[Y_{j1}^{(0)}(Y_{j'1}^{(0)})^2]}{\lambda_2^2} \right] + 4p^{2b}\kappa$$

$$\to \infty, \text{as } N \text{ and } p \text{ go to infinity.}$$

We can rewrite the CDF in (3.17) as

$$\begin{aligned}
&P_{H_a}(J_{2n} < x) \\
=&P_{H_a}\left(\frac{\sqrt{p}[T_n^{(1)} - E(T_n^{(1)})]}{\zeta^{(1)}} + \frac{\sqrt{p}[E(T_n^{(1)}) - E(T_n^{(0)})]}{\zeta^{(1)}} + \frac{\zeta^{(1)} - \zeta^{(0)}}{\zeta^{(1)}} x < x \right) + o(1) \\
=&P_{H_a}\left(\frac{\sqrt{p}[T_n^{(1)} - E(T_n^{(1)})]}{\zeta^{(1)}} + \frac{\sqrt{p}[E(\Gamma_{1,n,p}) + E(\Gamma_{2,n,p})]}{\zeta^{(1)}} + \frac{\zeta^{(1)} - \zeta^{(0)}}{\zeta^{(1)}} x < x \right) + o(1).
\end{aligned}$$

Although the CLT does not hold, we still have $\sqrt{p}[T_n^{(1)} - E(T_n^{(1)})]/\zeta^{(1)} = O_p(1)$ by Theorem 14.4.1 in Bishop et al. (2007). $(\zeta^{(1)} - \zeta^{(0)})/\zeta^{(1)}$ converges to some constant $-\infty < C < \infty$. Recall that $\sqrt{p}E(\Gamma_{1,n,p}) = O(N^{a/2-k})$ and $\sqrt{p}E(\Gamma_{2,n,p}) = O(N^{1+a/2-2k})$. When $k < 1/2$, the dominant term is $\sqrt{p}E(\Gamma_{2,n,p})/\zeta^{(1)}$. From (3.21), we know that $\zeta^{(1)} = \zeta^{(0)} + O(N^{1/2}\max_j |\delta_j|) \leq \tilde{C}N^{1/2}\max_j |\delta_j|$ for some constant $\tilde{C}$ since $\zeta^{(0)} = O(1)$ and $N^{1/2}\max_j \delta_j \to \infty$. Thus we have

$$\begin{aligned}
\frac{\sqrt{p}E(\Gamma_{2,n,p})}{\zeta^{(1)}} &= (\zeta^{(1)})^{-1}p^{1/2}\frac{1}{p}\sum_{j=1}^p \frac{n\delta_j^2}{\sigma_{1j}^2 + \sigma_{2j}^2(n/m)}(1 + o(1)) \\
&\geq (\tilde{C}N^{1/2}\max_j |\delta_j|)^{-1}p^{1/2}\frac{1}{p}\sum_{j=1}^p \frac{n\delta_j^2}{\sigma_{1j}^2 + \sigma_{2j}^2(n/m)}(1 + o(1)) \\
&= \tilde{C}^{-1}nN^{-1/2}p^{1/2}\max_j |\delta_j| \frac{1}{p}\sum_{j=1}^p \frac{(\delta_j/\max_j \delta_j)^2}{\sigma_{1j}^2 + \sigma_{2j}^2(n/m)} \\
&\to \infty.
\end{aligned}$$

Then the CDF approaches $P_{H_a}(\infty < x) + o(1) \to 0$ as N and p go to infinity.

Our main focus here is the two-sided test at significance level α, hence the rejection rule is $|J_{2n}| > z_{\alpha/2}$, where z_α denote the α^{th} upper percentile of standard normal distribution. The power function is

$$P_{H_a}(|J_{2n}| > z_{\alpha/2}) = 1 - P(-z_{\alpha/2} < J_{2n} < z_{\alpha/2}) = 1 - [P(J_{2n} < z_{\alpha/2}) - P(J_{2n} < -z_{\alpha/2})].$$

Summarizing the CDF discussed above, the power function is given in the following theorem.

Theorem 5. *Suppose $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\delta_1, \ldots, \delta_p)'$, where all the δ_j values are of the same order in terms of N, $\max_j |\delta_j| = O(N^{-k})$ for some k, $p = O(N^a)$ for $0 < a < 6$. Assuming conditions (C.1) and (C.3) hold, and (C.2) holds for $t_j^{(0)}$ defined in (3.16). With D_1 defined in (3.23), the power function for the two-sided alternative is given below.*

(i) If $k \geq 1/2$ and $D_1 \neq 0$, $\Delta_1 \equiv \lim_{N\to\infty} \sqrt{p}[E(\Gamma_{1,n,p}) + E(\Gamma_{2,n,p})]/\zeta^{(0)}$, where $\sqrt{p}E(\Gamma_{1,n,p})$ and $\sqrt{p}E(\Gamma_{2,n,p})$ are given in (3.22) and (3.24), then

$$\text{power} \quad \to \begin{cases} \alpha, \ \text{if } k > \max(a/2, 1/2 + a/4), \\ 1, \ \text{if } 1/2 \leq k < \max(a/2, 1/2 + a/4), \\ 1 - [\Phi(z_{\alpha/2} - \Delta_1) - \Phi(-z_{\alpha/2} - \Delta_1], \ \text{if } k = \max(a/2, 1/2 + a/4), \end{cases}$$

as N and p go to infinity.

(ii) If $k \geq 1/2$ and $D_1 = 0$, $\Delta_2 \equiv \lim_{N\to\infty} \sqrt{p}E(\Gamma_{2,n,p})/\zeta^{(0)}$, then

$$\text{power} \quad \to \begin{cases} \alpha, \ \text{if } k > 1/2 + a/4, \\ 1, \ \text{if } 1/2 \leq k < 1/2 + a/4, \\ 1 - [\Phi(z_{\alpha/2} - \Delta_2) - \Phi(-z_{\alpha/2} - \Delta_2], \ \text{if } k = 1/2 + a/4, \end{cases}$$

as N and p go to infinity.

(iii) If $k < 1/2$, then power → 1, as N and p go to infinity.

3.5 Simulation studies

The theoretical properties were established in previous sections. To examine the convergence under real scenarios, we will conduct Monte Carlo experiments and compare the performance, including the type I error and the power, of the new test with the tests proposed by Chen and Qin (2010), Srivastava et al. (2013), and Gregory et al. (2015). At the end of this section, we will also compare the new test with four Bootstrap tests which are variations of the permutation test procedure proposed by Blair et al. (1994).

3.5.1 Simulation settings

We generated the samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from the model $\mathbf{X} = \boldsymbol{\mu}_x + \boldsymbol{\xi}_x$, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ from $\mathbf{Y} = \boldsymbol{\mu}_y + \boldsymbol{\xi}_y$, where $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y \in \mathbb{R}^p$ are the mean vectors, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ for $i = 1, \ldots, n$ and $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jp})'$ for $j = 1, \ldots, m$, $\boldsymbol{\xi}_x$ and $\boldsymbol{\xi}_y$ have zero means and control the variance and dependence structure that will be specified later for each setting. We will suppress the indices i and j in the following description if there is no confusion. Under the null hypothesis, the two groups have the same mean. Without loss of generality, let $\boldsymbol{\mu}_x = \boldsymbol{\mu}_y = \mathbf{0}$ under the null hypothesis. Under the alternative hypothesis, let $\boldsymbol{\mu}_x = \mathbf{0}$, and $0 < \beta \leq 1$ be the percentage of nonzero elements in $\boldsymbol{\mu}_y$, so that the first $p\beta$ elements are equal to a constant $\delta \neq 0$ and the rest elements are equal to 0. The proportion of signal β can control the strength of signal in the experiments. When $\beta = 0$, it reduces to the null hypothesis. The choice for signal varies with the setting, because too large signals will fail to demonstrate the comparison among the tests, and small signals cannot show the power converging to 1. The sample sizes and number of variables have two possible combinations: $\{n = 45, m = 60, p = 300\}$ and $\{n = 90, m = 120, p = 300\}$.

Because the two groups are always independent, the dependency only exists within the group. We consider four dependency structures, that is, independent, weakly dependent, strongly dependent, and long-range dependent (LR). $\boldsymbol{\xi}_x$ and $\boldsymbol{\xi}_y$ have the common structure in each case, so in the following settings we will use $\xi_{z,t}$, to denote the t^{th} element of $\boldsymbol{\xi}_x$ or $\boldsymbol{\xi}_y$, $\epsilon_{z,t}$ to denote the independent sample from the innovation distribution. The settings of innovation distributions will be introduced later.

- **Independent (IND)**: $\xi_{z,t} = \epsilon_{z,t}$, and $\epsilon_{z,t}$ is independently drawn from the innovation distribution.

- **Weakly dependent (WD)**: $\xi_{z,t}$ is generated according to ARMA(2, 2) using `arima.sim` function[1] from `stats` package in `R`. The relationship among elements satisfies $\xi_{z,t} = 0.4\xi_{z,t-1} - 0.1\xi_{z,t-2} + \epsilon_{z,t} + 0.2\epsilon_{z,t-1} + 0.3\epsilon_{z,t-2}$, where $t = 1, \ldots, p$, and $\epsilon_{z,t}$ is independently drawn from the innovation distribution. The first 10 autocorrelation values for this structure are 0.584, 0.325, 0.072, -3.86E-03, -8.71E-03, -3.10E-03, -3.68E-04, 1.62E-04, 1.02E-04, 2.45E-05. The correlation decays fast.

- **Strongly dependent (SD)**: $\xi_{z,t}$ is generated according to AR(1) satisfying $\xi_{z,t} = 0.9\xi_{z,t-1} + \epsilon_{z,t}$. It is also generated by the `arima.sim` function in `R`. $\epsilon_{z,t}$ is independently drawn from the innovation distribution. The first 10 autocorrelation values for this structure are 0.9, 0.81, 0.729, 0.656, 0.59, 0.531, 0.478, 0.43, 0.387, 0.349.

- **Long-range dependent (LR)**: The long-range dependent process is generated following the approach proposed by Hall et al. (1998). The self-similarity parameter is set to $H = 0.7$. Independently draw $\epsilon_{z,t}$ from the innovation distribution and stack them in a vector $\boldsymbol{\epsilon}_z = (\epsilon_{z,1}, \ldots, \epsilon_{z,p})$. Let $R = (r_{ij})$, where $r_{ij} = 0.5[(k+1)^{2H} + (k-1)^{2H} - 2k^{2H}]$, for $k = |i - j|$. Decompose R as $R = U'U$ by Cholesky factorization. Then $\boldsymbol{\xi}_z = U'\boldsymbol{\epsilon}_z$ for $z = x$ or y. The first 10 autocorrelation values for this structure are 0.32, 0.189, 0.146, 0.122, 0.107, 0.096, 0.087, 0.081, 0.075, 0.07. It is noting that the correlation still

[1] There is a "burn-in" period in the generating process, so that the beginning part of the generated sequence will be discarded. The details on how to determine the length of the burn-in period can be found in the help file: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/arima.sim.html.

tends= to 0 when the distance of the two components goes to infinity, but the decaying speed is slower than the ARMA model. In fact, the speed the correlation converging to 0 is $O(k^{-2(1-H)})$, specifically $O(k^{-0.6})$, when the distance k goes to infinity (Samorodnitsky, 2007). It is easy to check the sum of mixing coefficients, $\sum_{k=1}^{\infty} \alpha(k) < \infty$, may not hold with this decaying rate. The null distribution of our test statistic and the GCT test will not be valid in this case, but we will use this setting to check the robustness of the tests to the violation of the dependence regularity condition.

To check the robustness of the tests with respect to different distribution shapes, we consider four innovation distributions. The first one is the standard normal distribution served as a benchmark. The second is a skewed distribution, the shifted Gamma distribution with shape parameter 4 and scale parameter 2, which has mean 0, variance 16, and skewness 1. It is produced by shifting the ordinary Gamma distribution such that the center becomes 0. The third is the t distribution with degrees of freedom 3. The last one is the Cauchy(0, 0.1) distribution with density function $f(x) = [0.1\pi(1 + 100x^2)]^{-1}$ for $x \in \mathbb{R}$.

In the literature, there is no one test maintaining the best performance for all circumstances. Our test belongs to the sum-of-squares based test, so our scope of comparison is also confined to the same type of tests, including the GCT test (Gregory et al., 2015), the SKK test (Srivastava et al., 2013) and the CQ test (Chen and Qin, 2010). The GCT and SKK tests share the most similarities with our test. The CQ test is also a sum-of-squares test that improves the classical test proposed by Bai and Saranadasa (1996).

The window for our new test is chosen as $L = \lceil p^{3/8} \rceil$, where $\lceil r \rceil$ is the smallest integer not less than r. Under our settings, for example, $L = 9$ when $p = 300$. We choose the Parzen window with width 10 for the GCT test because this configuration was reported to have the best performance by Gregory et al. (2015). This window width is also close to our choice for L.

3.5.2 Numerical results

For each setting, we run tests under the Monte Carlo experiment with 2000 rounds. At a 5% significance level, the proportion of rejections out of the 2000 runs is recorded. If the data are generated under the null hypothesis, the proportion of rejections is the empirical type I error rate, whereas the proportion of rejections for the data generated under the alternative hypothesis is the empirical power. To account for the Monte Carlo error, we allow a margin of error $2\sqrt{0.05(1-0.05)/2000} \approx 0.01$. If the empirical type I error exceeds 0.06, then it fails to control the type I error and there is no further necessity to study its power.

The emphases of the simulation results are put on the following aspects:

- The effect of the high-order center corrections for the test statistic.
- The performance of our new test under different dependency structures.
- The robustness of our new test to different innovation distributions, including the violation of moment conditions.

Center correction

In Section 3.3, we discussed two versions of centering parameters. To investigate the effect of the center correction, we designed numerical experiments for our new test and GCT test, with and without the second-order center correction. When reporting the results, "ZWLm" and "ZWL" represent our new tests using J_{1n} and J_{2n} as test statistics, respectively, and "GCTm" and "GCT" are the moderate-p and large-p GCT tests, respectively.

The results for $n = 45$ and $m = 60$ under normal innovation distribution are displayed in Figure 3.1. As can be seen from the top-left panel, with the independent data, ZWL and ZWLm both successfully control the type I error. Both versions of the GCT tests have inflated type I error. ZWLm demonstrates more power than the ZWL test, and most of the power differences are more than the margin of error. For β value close to 1, i.e., when the signals are completely dense, all the tests show similar power close to 1. For our new

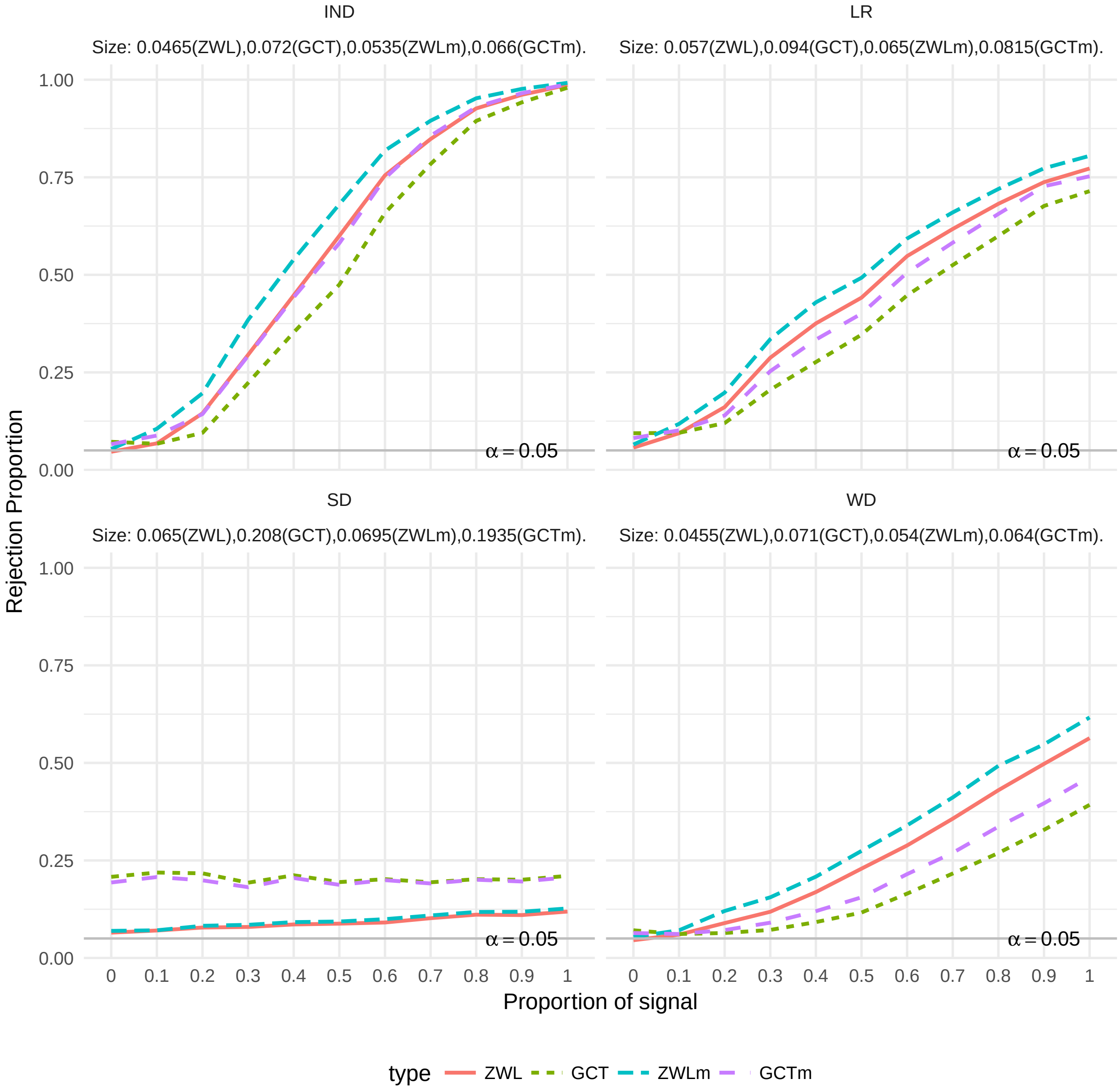


Figure 3.1: *The proportion of rejections of ZWLm, GCTm, CQ, and SKK tests based on 2000 runs when the innovation follows* **standard normal distribution** *and sample sizes are* $\boldsymbol{n = 45}$ *and* $\boldsymbol{m = 60}$*. The line "Size" on top of each graph shows the empirical type I error rates. "ZWLm" and "ZWL" are our new tests using* J_{1n} *and* J_{2n} *as test statistics, respectively. "GCTm" and "GCT" are the moderate-p and large-p GCT tests, respectively. The number of variables is* $p = 300$*. The signal magnitude is* $\delta = 0.125$*. "Proportion of signal" refers to* β*, which controls the sparsity of signal*

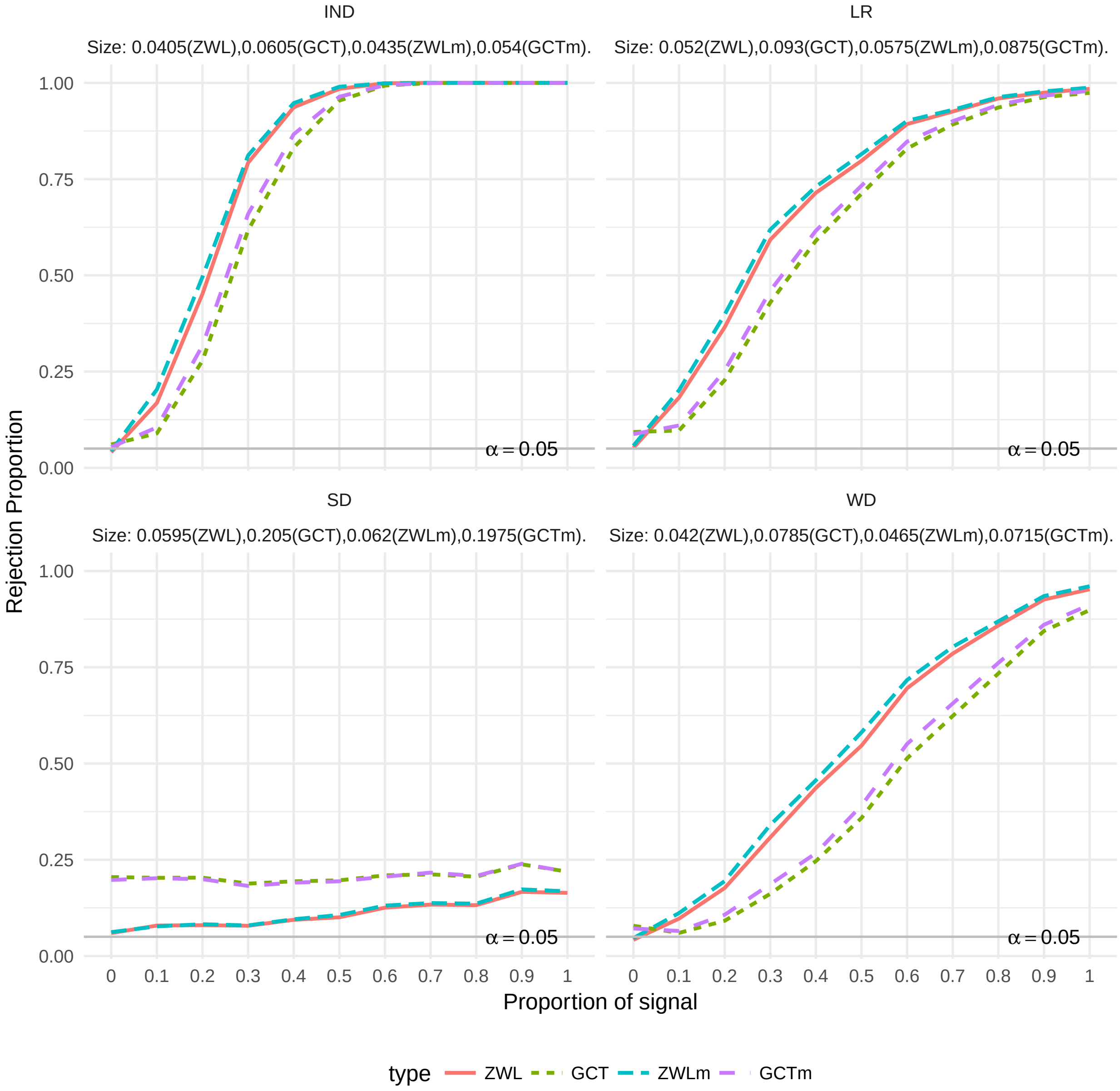


Figure 3.2: *The proportion of rejections of ZWLm, GCTm, CQ, and SKK tests based on 2000 runs when the innovation follows* **standard normal distribution** *and sample sizes are* $\boldsymbol{n = 90}$ *and* $\boldsymbol{m = 120}$*. The line "Size" on top of each graph shows the empirical type I error rates. "ZWLm" and "ZWL" are our new tests using* J_{1n} *and* J_{2n} *as test statistics, respectively. "GCTm" and "GCT" are the moderate-p and large-p GCT tests, respectively. The number of variables is* $p = 300$*. The signal magnitude is* $\delta = 0.125$*. "Proportion of signal" refers to* β*, which controls the sparsity of signal*

tests, the center correction helps to control the type I error but also leads to some power loss. In contrast, the correction brings higher empirical type I error and lower power for the GCT test. Compared with the independent setting, all the tests for weakly dependent data (bottom-right panel) suffer from inflated type I error and less power to some extent. However, the relative positions of the power curves do not change. Under strong dependency (bottom-left panel), GCT tests completely loss the control on type I error, but ZWL tests only have slight inflation. All the tests have very low power. With the long-range dependence, only ZWL can control the type I error.

Both our new test and the GCT test require that sample size and number of variables go to infinity in order to converge to normal distribution. If we increase the sample size to $n = 90$ and $m = 120$, as shown in Figure 3.2, the type I errors are mostly reduced, the powers are elevated and converge to 1 faster along with the signal proportion. The relative positions for the curves are unchanged compared with Figure 3.1. We can clearly observe the gap between the curves of our tests and the GCT tests. It is worth noting that the gap between the center-corrected version test and the uncorrected version has been reduced because the correction term $n^{-1}\hat{a}_n + n^{-2}\hat{b}_n$ diminishes as the sample size grows.

The center correction leads to some power loss for our new tests. For all the dependence structures, the effect of center correction is similar: it helps to control the type I error for the proposed tests, but it does not bring any benefit to the GCT tests.

The validity of center correction also relies on moment conditions. When the moment conditions are not satisfied, the non-corrected version exhibits more robustness. According to the numerical results, if we use the Cauchy innovation, with all other settings the same as the independent case in Figure 3.6, both the corrected version ZWL and GCT have type I error close to 1, losing the control completely.

The dependence structure

Figure 3.3 to Figure 3.6 summarize the results for four tests under four dependence structures and four innovation distributions in a factorial experiment. Within each figure, we can compare the performance under different dependence structures. From the discussion on center correction, the uncorrected version ZWL test has a similar type I error, higher power, and more robustness to heavy tail. The correction does not bring any benefit to the GCT test. Therefore, ZWLm and GCTm are used to compare with CQ and SKK tests in the following.

With the normal innovation (Figure 3.3), the GCTm test fails to control the type I error for all the cases. With independent structure, ZWLm, SKK, and CQ are able to control the type I error. SKK has the lowest type I error and highest power, whereas ZWLm has a similar performance. Under long-range dependence, SKK and CQ successfully control the type I error, but ZWLm fails to control. Both ZWLm and GCTm require the mixing coefficients to have a finite sum, which does not hold under long-range dependence. Under strong dependence, the CQ test provides exact control for type I error, SKK tends to be conservative, ZWLm is slightly liberal. Under weak dependence, the ZWLm test well controls the type I error and exhibits the highest power. The patterns with shifted Gamma and $t(3)$ distributions (see Figure 3.4 and 3.5) are similar to that with the Normal innovation.

When the innovation is changed to Cauchy distribution (Figure 3.6), the GCTm test still fails to control the type I error, especially with long-range and strong dependence. Under independence and weak dependence, SKK, CQ, and ZWLm tend to be conservative. Under LR structure, ZWLm is too liberal, whereas SKK is too conservative. Under SD structure, we make the signal stronger $\delta = 3$. CQ has low power, and ZWLm and SKK show conservative type I error, but ZWLm maintains the highest power.

In summary, when the data are independent or weakly dependent, the proposed test and the SKK test have the best performance. The dependency of data will impair powers for all the tests. When the dependence is very strong, the proposed test tends to become slightly liberal, while SKK tends to be conservative. In general, the GCT test cannot well control the type I error; CQ can always control the type I error but the power is usually inferior to the best performance. The type of innovation distribution also leads to complications of the effect of dependence structure. When the dependence is weak, the proposed test and the SKK tests are recommended. When the dependency is strong, it is suggested to use the CQ test. In practice, the dependence of the data can be determined in advance by checking the sample correlations.

Robustness to the innovation distribution

Among the tests being compared, SKK relies on the normality assumption of the data, and the proposed test, GCT test, and CQ test only assume some moment conditions. To examine the sensitivity to the choice of distributions, we consider the Normal, Gamma, t, and Cauchy distributions for the innovation distribution. It is noted that the Cauchy distribution will violate all the moment conditions because of the nonexistence of moments. For $t(3)$ distribution, all moments higher than the second moment do not exist, which also violates most of the moment assumptions stated in the theoretical results.

Figure 3.4 displays the results under skewed distribution, the shifted $Gamma(4, 2)$ distribution. This distribution only has skewness 1, so the results here only represent the scenario of slightly skewed populations. The pattern is similar to the standard normal distribution case in Figure 3.3.

The pattern of the t distribution case is also very similar to the Normal case. As can be seen from Figure 3.5, the type I error of our new test can be controlled better for the long-range and weak dependence settings, although the strong dependence setting still sees slightly inflated type I error rate.

Under the Cauchy innovation, as shown in Figure 3.6, the ZWLm and SKK all have conservative type I error for all the dependence structures except that ZWLm is liberal under long-range dependence. The ZWLm test is preferable to the others except for the

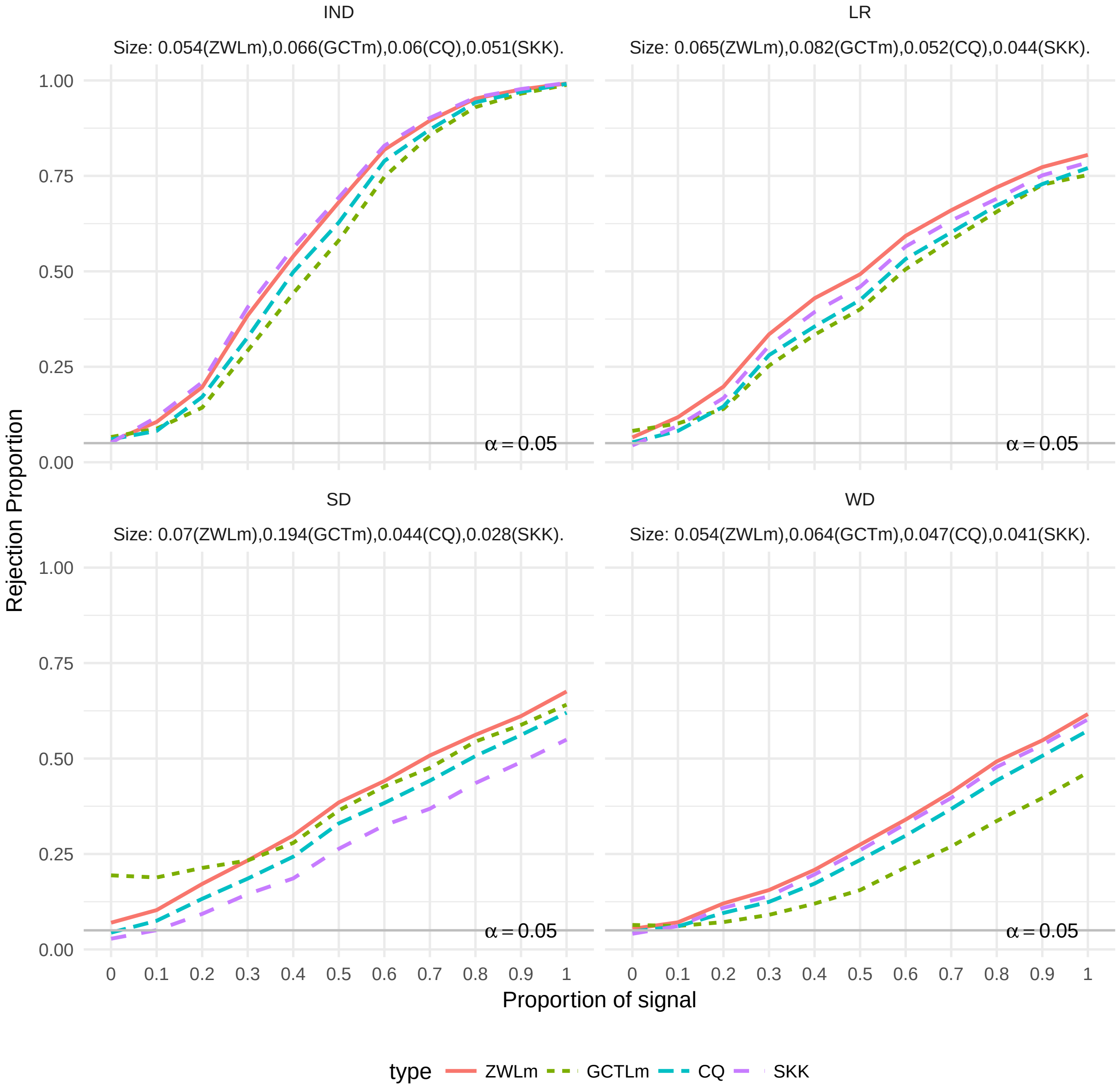


Figure 3.3: *The proportion of rejections of ZWLm, GCTm, CQ, and SKK tests based on 2000 runs when the innovation follows* ***standard normal distribution****. The line "Size" on top of each graph shows the empirical type I error rates. "ZWLm" is the test using J_{1n}; "GCTm" is the moderate-p GCT test; "CQ" is the test proposed by Chen and Qin (2010); "SKK" is the test suggested by Srivastava et al. (2013). The sample sizes are* ***n = 45, m = 60****. The number of variables is p = 300. The signal magnitude is $\delta = 0.125$ for IND, WD and LR, and 0.375 for SD. "Proportion of signal" refers to β, which controls the sparsity of signals.*

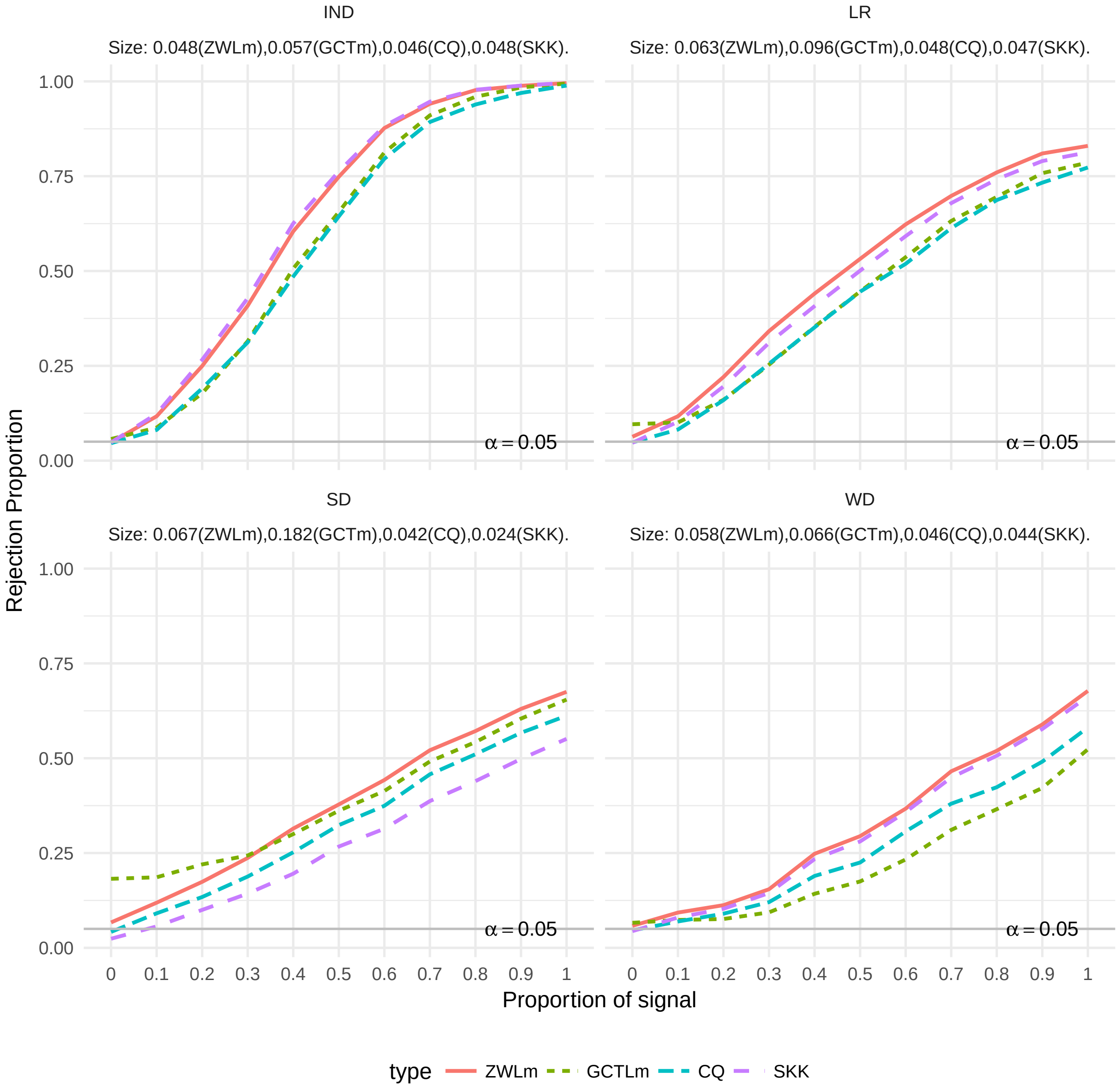


Figure 3.4: *The proportion of rejections of ZWLm, GCTm, CQ, and SKK tests based on 2000 runs when the innovation follows* ***shifted Gamma(4,2) distribution****. The line "Size" on top of each graph shows the empirical type I error rates. "ZWLm" is the test using J_{1n}; "GCTm" is the moderate-p GCT test; "CQ" is the test proposed by Chen and Qin (2010); "SKK" is the test suggested by Srivastava et al. (2013). The sample sizes are* $\boldsymbol{n = 45}$*,* $\boldsymbol{m = 60}$*. The number of variables is* $p = 300$*. The signal magnitude is* $\delta = 0.5$ *for IND, WD and LR, and 1.5 for SD. "Proportion of signal" refers to* β*, which controls the sparsity of signals.*

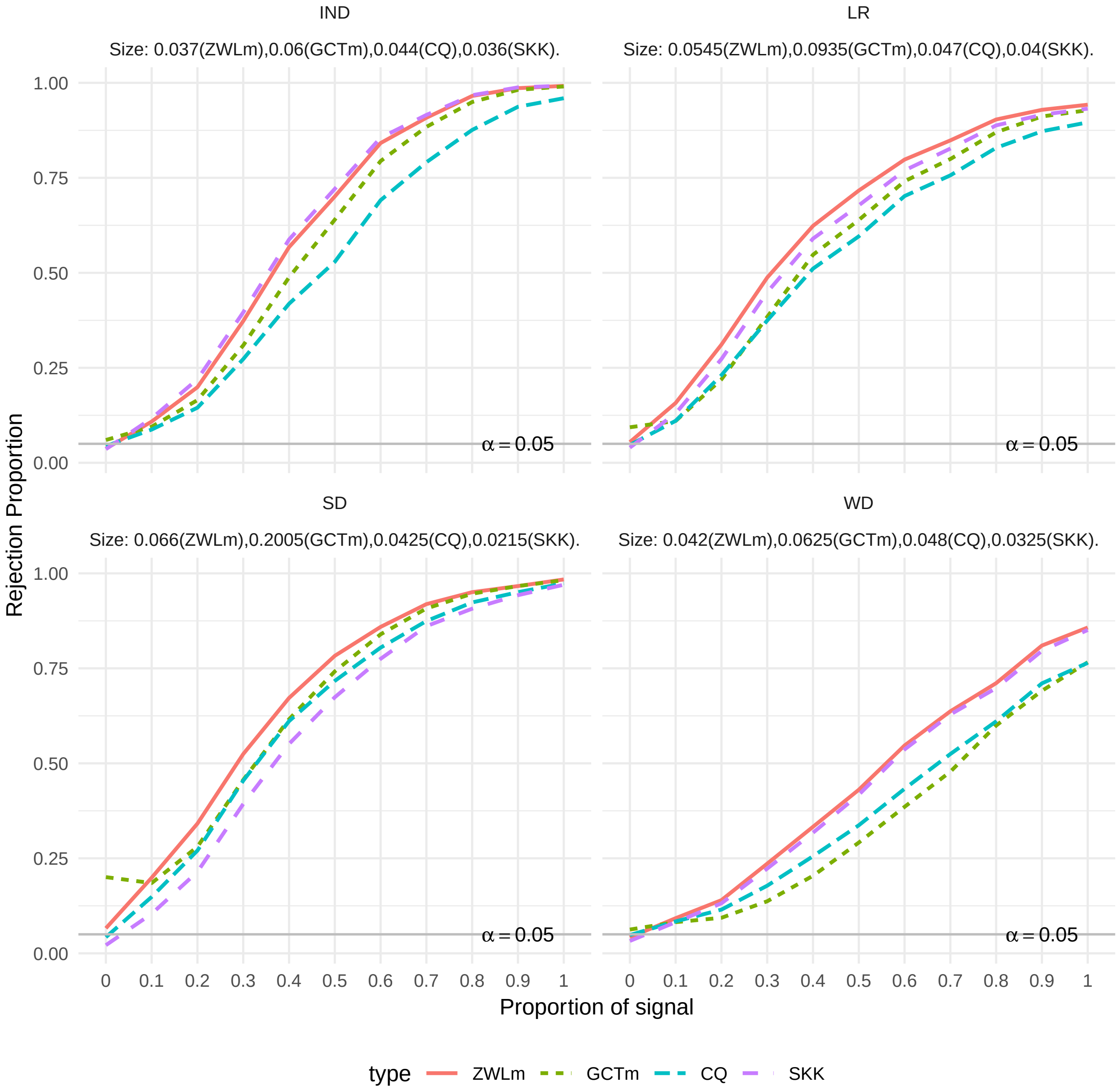


Figure 3.5: *The proportion of rejections of ZWLm, GCTm, CQ, and SKK tests based on 2000 runs when the innovation follows* **t(3) distribution**. *The line "Size" on top of each graph shows the empirical type I error rates. "ZWLm" is the test using* J_{1n}*; "GCTm" is the moderate-p GCT test; "CQ" is the test proposed by Chen and Qin (2010); "SKK" is the test suggested by Srivastava et al. (2013). The sample sizes are* $\boldsymbol{n = 45}$*,* $\boldsymbol{m = 60}$*. The number of variables is* $p = 300$*. The signal magnitude is* $\delta = 0.2$ *for IND, 0.25 for WD and LR, and 1 for SD. "Proportion of signal" refers to* β*, which controls the sparsity of signals.*

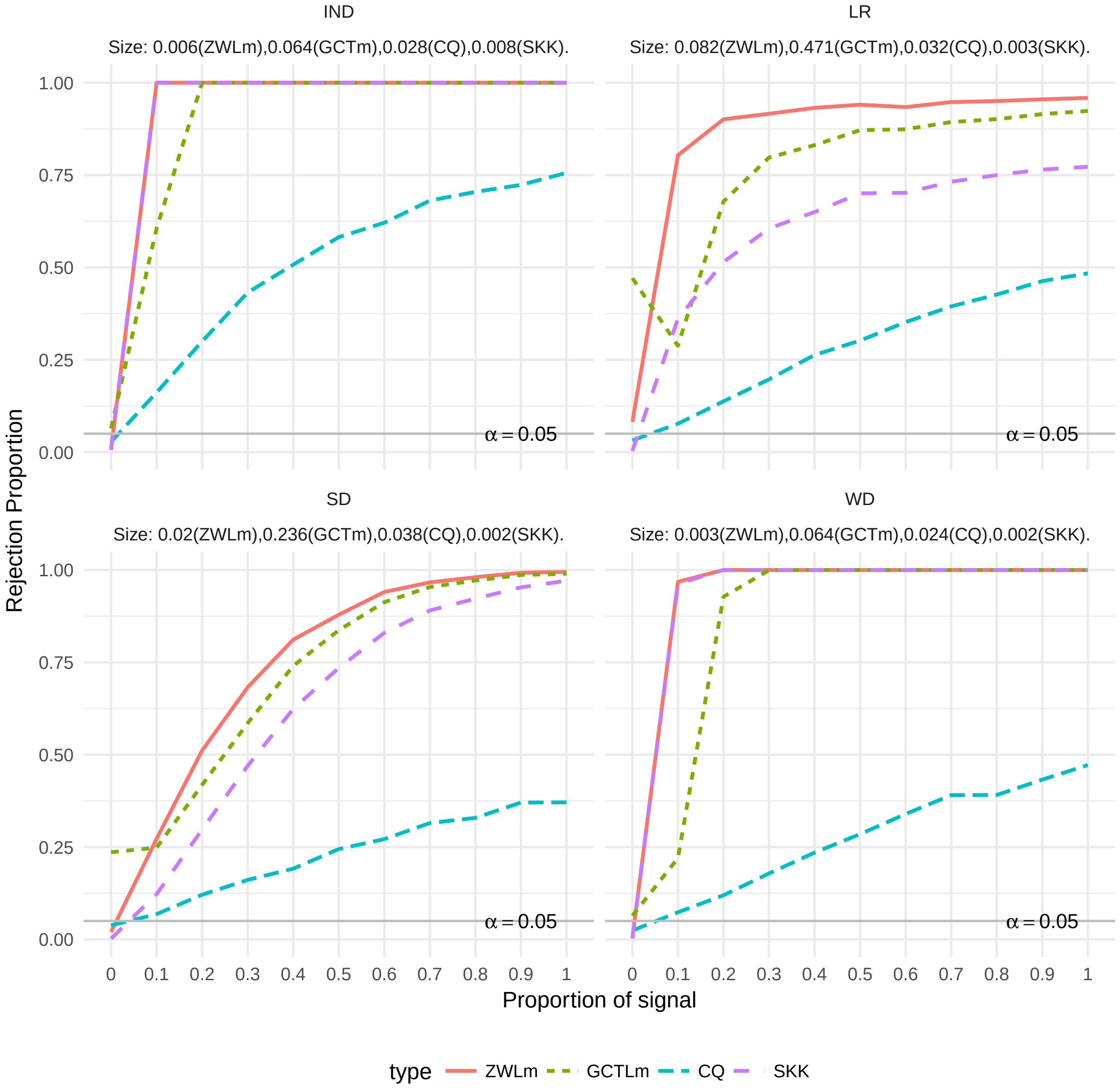


Figure 3.6: *The proportion of rejections of ZWLm, GCTm, CQ, and SKK tests based on 2000 runs when the innovation follows* **Cauchy(0, 0.1) distribution**. *The line "Size" on top of each graph shows the empirical type I error rates. "ZWLm" is the test using J_{1n}; "GCTm" is the moderate-p GCT test; "CQ" is the test proposed by Chen and Qin (2010); "SKK" is the test suggested by Srivastava et al. (2013). The sample sizes are* $\boldsymbol{n = 45}$, $\boldsymbol{m = 60}$. *The number of variables is $p = 300$. The signal magnitude is $\delta = 1$ for IND, WD and LR, and 3 for SD. "Proportion of signal" refers to β, which controls the sparsity of signals.*

long-range structure. Compared with the normal and gamma cases, the performances of ZWLm and SKK are still very close to each other at the independent or weakly dependent settings, and SKK shows lower power for strong or long-range dependencies. The CQ test suffers the most drop in power, whereas the GCTm test deviates the nominal level even seriously.

The result demonstrates that the proposed test is robust to slight skewness and heavy tails to some extent. Even the moment conditions do not hold, the new test still shows relatively good performance among the tests in comparison. In practice, the proposed test is recommended for testing the heavy-tailed data.

3.5.3 Comparison with Bootstrap tests

In simulation studies, we further consider the tests proposed by Blair et al. (1994). Their permutation test procedure was designed for paired samples. Formally, they first compute the difference vector $\mathbf{d}_i = \mathbf{X}_i - \mathbf{Y}_i$ for the i^{th} subject where $i = 1, \ldots, n$. Randomly assign $+$ or $-$ signs to each $\mathbf{d}_i$ to form a permutation. There are 2^n possible permutations of the difference vectors. They advised three versions of test statistics based on the univariate component-wise paired t-statistic, i.e., $\sum t_j$, $\max t_j$, and $\sum |t_j|$. These statistics are computed on each permutation of difference vectors to form a sampling distribution, then the p-value is obtained by calculating the percentage of statistics that exceed the value of the statistic computed on the original data set. As the sample size gets larger, 2^n becomes a huge number, so they suggested to take a large random sample, for example, 1,000 permutations, as an approximation. To accommodate to our independent two-sample setting, we propose a Bootstrapping version: mix all the $n + m$ observation vectors in the two samples, randomly select n observations to form the first sample, and the rest m observations are left as the second sample. Compute the univariate Welch's t-statistic for each axis of the p dimensions. The test statistics are $\sum t_j$, $\max t_j$, $\sum |t_j|$, following Blair et al. (1994), and $\sum t_j^2$, because our test is based on squared t-statistics. Repeat 1,000 times to get the

Bootstrapping distributions for the test statistics. The two-sided p-values are obtained by calculating the percentage of absolute values of statistics that exceed the absolute value of the statistic computed on the original data.

For a Monte Carlo study, we consider similar settings as the preceding section and report part of the results in Table 3.1 and 3.2. The tables report the proportion of rejections under the settings with Normal and Cauchy distribution, independent and strong dependent dependency, and our new test (moderate-p version) and the four Bootstrapping tests. Note that, $\beta = 0$ corresponds to the null hypothesis, and the proportion of rejections is the empirical type I errors.

In Table 3.1, the Bootstrap $\sum t_j$ shows a performance superior to other tests under the Normal distribution, then our new test follows, and the Bootstrap $\max t_j$ has low power. Under the Cauchy independent setting, all the tests show big power. This could be due to the signal $\delta = 1$ is too strong. For the Cauchy strong dependent case, the Bootstrap $\max t_j$ shows the best power, while Bootstrap $\sum t_j$ ranks the lowest. In a Cauchy distribution, it is easy to get an extremely large value. This value may be diluted in a summation by other smaller values, but the max can pick up this value in the final statistic. So it is not surprising that $\max t_j$ has the best performance under Cauchy distribution. It is also noted that Bootstrap $\sum |t_j|$ and $\sum t_j^2$ have similar results because comparing absolute values and comparing squared values are essentially the same.

It seems that the Bootstrap $\sum t_j$ is a good option for testing under the Normal distribution. If we further examine the construction of $\sum t_j$, a potential problem is that univariate t-statistics with different signs may cancel out each other. In the alternative hypothesis in Table 3.1, a positive quantity δ is added to some components of the second population, so most of the t-statistics corresponding to the non-zero components tend to have the same sign, and the effects will cumulate when adding them up. Hence, we further consider a setting that is less favorable to $\sum t_j$: for the alternative hypothesis, signals are set half positive and half negative. In this way, the t-statistics for non-zero components will also tend to be

Table 3.1: *The proportion of rejections for the new test and Bootstrap tests. The results are based on 2,000 runs.* H_a : $\boldsymbol{\mu}_x = \mathbf{0}$; $\boldsymbol{\mu}_y = \{\delta\mathbf{1}'_{p\beta}, \mathbf{0}'_{1-p\beta}\}$

		Normal $\delta = 0.125$					Cauchy $\delta = 1$				
	β	ZWLm	$\sum t_j$	$\sum t_j^2$	$\max t_j$	$\sum \lvert t_j \rvert$	ZWLm	$\sum t_j$	$\sum t_j^2$	$\max t_j$	$\sum \lvert t_j \rvert$
IND	0	0.048	0.055	0.000	0.060	0.000	0.004	0.050	0.000	0.026	0.011
	0.2	0.216	0.581	0.011	0.085	0.009	1.000	1.000	1.000	1.000	1.000
	0.4	0.522	0.990	0.101	0.126	0.099	1.000	1.000	1.000	1.000	1.000
	0.6	0.826	1.000	0.325	0.146	0.300	1.000	1.000	1.000	1.000	1.000
	0.8	0.957	1.000	0.627	0.185	0.623	1.000	1.000	1.000	1.000	1.000
	1	0.992	1.000	0.854	0.228	0.858	1.000	1.000	1.000	1.000	1.000
SD*	0	0.056	0.046	0.023	0.051	0.022	0.004	0.047	0.009	0.019	0.029
	0.2	0.076	0.069	0.032	0.059	0.033	0.292	0.163	0.335	0.625	0.267
	0.4	0.105	0.137	0.052	0.075	0.050	0.659	0.489	0.704	0.863	0.607
	0.6	0.138	0.264	0.073	0.082	0.070	0.853	0.782	0.878	0.940	0.818
	0.8	0.181	0.388	0.096	0.092	0.096	0.968	0.939	0.973	0.984	0.951
	1	0.231	0.619	0.134	0.102	0.135	0.986	0.985	0.991	0.994	0.979

* AR(1) coefficient = 0.8.

half positive and half negative. The t-statistics tend to be canceled when they are summed up. Table 3.2 displays the proportion of rejections in this setting. The Bootstrap $\sum t_j$ loses power in all the cases while other tests do not change much. Our new test performs the best under Normal distribution, and the Bootstrap $\max t_j$ has the best result under strong dependent Cauchy case.

3.6 Application to real data

In biological studies, identifying differentially expressed genes or gene-sets helps understand the complex mechanism at the genetic level. The metadata of genome annotations, for example, the Gene Ontology (GO), can boost the analysis of microarray data. The GO is a structured description of genes according to their biological functions. A GO term is a set of genes impact the same biological functionality. As a real-data example, we will apply our new high-dimensional test to detect differentially expressed GO terms associated with certain phenotypes. Through a series of experiments, our test shows good control in type I error, more statistical power, and consistency in detecting differentially expressed GO terms.

Table 3.2: *The proportion of rejections for the new test and Bootstrap tests. The results are based on 2,000 runs.* $H_a:\ \boldsymbol{\mu}_x = \mathbf{0};\ \boldsymbol{\mu}_y = \{\delta\mathbf{1}'_{0.5p\beta}, -\delta\mathbf{1}'_{0.5p\beta}, \mathbf{0}'_{1-p\beta}\}$

		Normal $\delta = 0.125$					Cauchy $\delta = 1$				
	β	ZWLm	$\sum t_j$	$\sum t_j^2$	$\max t_j$	$\sum \lvert t_j \rvert$	ZWLm	$\sum t_j$	$\sum t_j^2$	$\max t_j$	$\sum \lvert t_j \rvert$
IND	0	0.048	0.055	0.000	0.060	0.000	0.004	0.050	0.000	0.026	0.011
	0.2	0.222	0.045	0.009	0.081	0.012	1.000	0.322	1.000	1.000	1.000
	0.4	0.521	0.053	0.105	0.130	0.096	1.000	0.490	1.000	1.000	1.000
	0.6	0.832	0.042	0.338	0.160	0.319	1.000	0.570	1.000	1.000	1.000
	0.8	0.955	0.048	0.625	0.190	0.622	1.000	0.616	1.000	1.000	1.000
	1	0.992	0.048	0.862	0.219	0.873	1.000	0.643	1.000	1.000	1.000
SD*	0	0.056	0.046	0.023	0.051	0.022	0.004	0.047	0.009	0.019	0.029
	0.2	0.072	0.044	0.034	0.058	0.036	0.282	0.064	0.331	0.649	0.265
	0.4	0.114	0.051	0.059	0.076	0.060	0.652	0.071	0.704	0.864	0.606
	0.6	0.130	0.053	0.065	0.076	0.065	0.861	0.073	0.890	0.951	0.823
	0.8	0.183	0.051	0.105	0.091	0.105	0.960	0.107	0.972	0.982	0.942
	1	0.214	0.051	0.123	0.098	0.125	0.986	0.102	0.991	0.994	0.977

* AR(1) coefficient = 0.8.

The GO Consortium (http://geneontology.org/) provides annotations for the genes that map genes to GO terms based on three biological functionalities: biological process (BP), cellular components (CC), and molecular functions (MF). The GO terms are organized in a hierarchical structure and each term is a node in a directed acyclic graph (DAG) structure; low-level GO term is nested within a high-level one. A gene may appear in multiple GO terms, and a high-level GO term always contains the genes in its child node.

If one wanted to determine whether a GO term differentially expresses under two phenotypes, a multivariate test is demanded. The Hotelling's T^2 test is a feasible method if the data are in low dimension. For the GO terms with the number of genes greater than the sample sizes, however, a high-dimensional test is desired due to the restriction of Hotelling's T^2 in high dimension. We will demonstrate the use of our new test by applying it to acute lymphoblastic leukemia (ALL) dataset (Chiaretti et al., 2004). This dataset contains 128 cell observations with 12,625 microarray expression measures and 21 phenotypes. Chen and Qin (2010) applied their high-dimensional two-sample test to this dataset for the B-cell ALL with the BCR/ABL genetic translocation (sample size $n_1 = 37$) and cytogenetically normal

NEG B-cell ALL (sample size $n_2 = 42$). We will follow their data processing strategy. We obtained the data from the `ALL` (Li, 2009) package in `R` software, where the expression measures were preprocessed by a three-step robust multichip average method and subjected to base 2 logarithmic transformation. Following Dudoit et al. (2008), we further filtered the genes based on two criteria: first, retain the genes with expression measure greater than 100, in the absolute scale, in at least 25% of the 79 observations; second, retain the genes with expression measure having interquartile range greater than 0.5, in log base 2 scale. Finally, 2,391 genes entered our genes pool after filtering. The GO annotations are available in `R` package `hgu95av2.db`, and their maps to genes are provided by `topGO` package.

To evaluate and compare the high-dimensional two-sample tests on this real-data task, we designed an experimental scheme with two parts. Part 1 focuses on the statistical power of the tests: we compare the BCR/ABL and NEG groups by applying the tests on all the GO terms. It corresponds to the alternative hypothesis because we know the two groups have different phenotypes. Part 2 simulates a pseudo null hypothesis: we randomly split the NEG class into two groups and apply the two-sample tests on them. In this setting, the tests are supposed to not reject the null hypothesis.

Note that the Hotelling's T^2 is not well defined when $p > n + m - 2$ (Dempster, 1958). The NEG group contains 42 observations. When testing the two splits of NEG, n+m-2 =40. For number of dimension $p > 40$, we need high-dimensional tests. Due to this reason, we only keep the GO terms containing more than 40 genes. Finally, we obtained 1,256 GO terms and 2,310 genes in BP, 165 GO terms and 2,306 genes in MF, and 217 GO terms and 2,332 genes in CC. Summary statistics for the number of genes for our processed data are displayed in Table 3.3. Because the GO research community keeps adding new terms in the database, there are more GO terms in our analysis relative to Chen and Qin (2010).

In the experiments, we include our new test with test statistic J_{1n} (ZWLm), the GCT test (moderate-p version, denoted by GCTm), the Chen and Qin (2010) test (CQ), and the Srivastava et al. (2013) test (SKK). The parameter configurations are the same as those in

Table 3.3: *Summary for number of genes in the GO terms.*

Ontology Group	# of GO terms	Min	Max	Mean	Median	Std. Dev.
BP	1256	41	2310	207.2	95.5	286.6
MF	165	41	2306	218.3	104	331.5
CC	217	41	2332	334.0	128	512.0

Table 3.4: *Proportion of significant GO terms at* $\alpha = 0.05$ *for comparing BCR/ABL and NEG classes.*

Ontology	Bonferroni correction				BY procedure			
	ZWLm	GCTm	CQ	SKK	ZWLm	GCTm	CQ	SKK
BP	0.83	0.38	0.41	0.47	0.91	0.71	0.63	0.63
MF	0.86	0.57	0.29	0.32	0.92	0.83	0.48	0.52
CC	0.87	0.61	0.30	0.34	0.96	0.88	0.56	0.62

our simulation study if not specified otherwise (the window width $r = 9$ for the GCT test and $r = p^{3/8}$ for our new test). At significance level $\alpha = 0.05$, the proportion of significant GO terms are shown in Table 3.4. Within each ontology group there arises the multiple-testing issue, so we adjust the p-values using both the Bonferroni correction and the false discovery rate controlling procedure proposed by Benjamini and Yekutieli (2001), abbreviated as BY. The relative situation for the tests are similar under both adjustment approaches, so we mostly report the Bonferroni corrected version in later result displaying. Because the two samples belong to two phenotype classes, rejecting the null hypothesis indicates that the GO term expresses differentially. In most of the cases, our new test rejects more, the GCT ranks the second, and CQ and SKK reject relatively less often. It indicates that our new test has more statistical power in detecting differentially expressed GO terms than other tests.

Figure 3.7 shows the notch boxplot for the p-values. The notches give the 95% confidence interval for the median, which is median $\pm 1.57\text{IQR}/\sqrt{n}$, where IQR represents the interquartile range. The notches outside the plot represent that the bound of the confidence interval exceeds the range of 0 to 1. From this chart, the p-values of our test are more concentrating toward 0 than other tests. Differentially expressed GO terms are more likely being detected by using our new test.

We do not have *a-priori* biological knowledge about which GO terms have a major

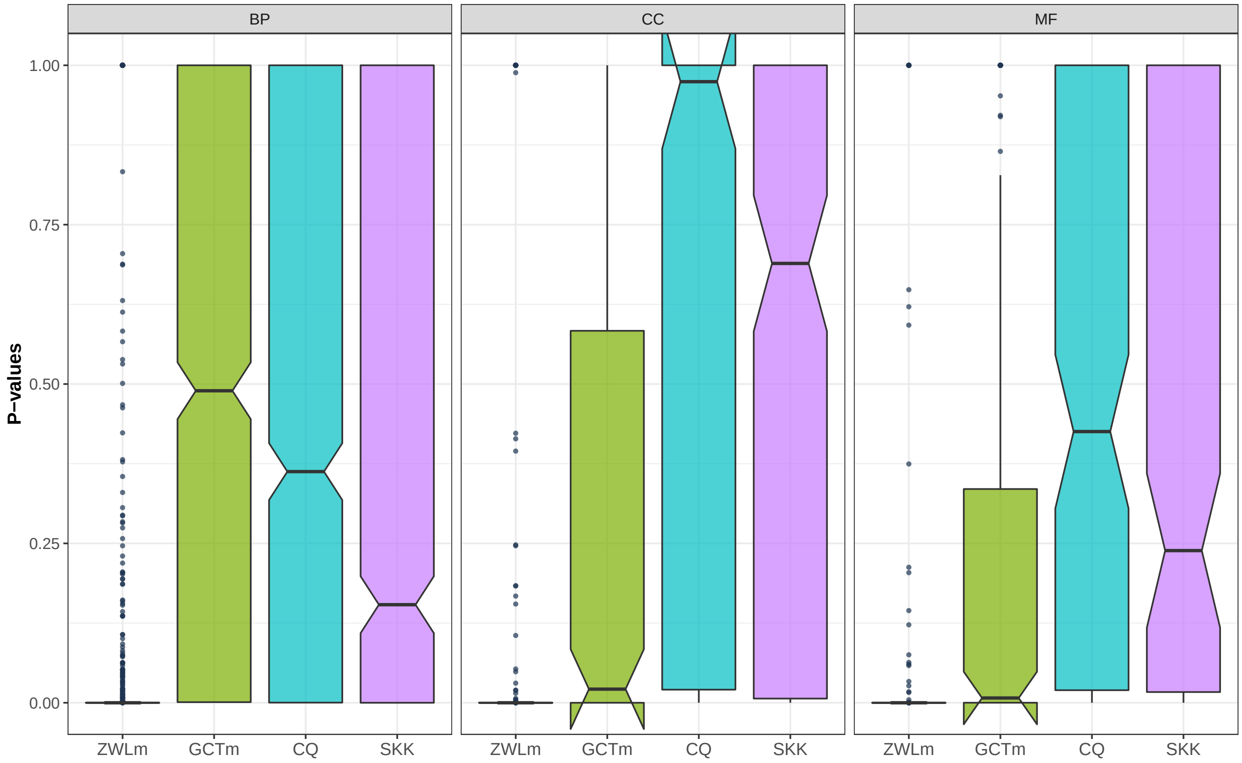


Figure 3.7: *Boxplot of adjusted p-values (Bonferroni correction) for tests comparing NEG and BCR/ABL.*

contribution to the difference of the phenotypes. Based on the two-class data, we further designed an experiment to evaluate the robustness of the tests in finding the differentially expressed GO terms. **A good procedure is expected to consistently identify the important GO terms on different datasets.** Since we do not have multiple datasets, we artificially create some by random sampling: randomly split the 37 BCR/ABL observations into two subsets BCR/ABL_1 with sample size 18 and BCR/ABL_2 with size 19; similarly, randomly split the 42 NEG observations into NEG_1 and NEG_2, both with sample size 21. For each type of hypothesis tests, run the test on BCR/ABL_1 vs. NEG_1 and BCR/ABL_2 vs. NEG_2 and for each GO term, then collect the names of significant GO terms and store them in sets $\mathcal{S}_1$ and $\mathcal{S}_2$. The p-values are adjusted by the Bonferroni approach, and the tests are conducted under significance level 0.05. For each type of the tests, compute the following robustness score (RS):

$$RS = \frac{|\mathcal{S}_1 \cap \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|},$$

where $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S}$. The idea of our robust score is borrowed from the Intersection over Union (IoU) metric in computer vision literature, which is the most commonly used metric for comparing the similarity between two arbitrary shapes, for example, in object detection (Everingham et al., 2010) and image segmentation (Rahman and Wang, 2016). If a GO term plays an important role, the hypothesis test should be able to identify it based on different data subsets. High RS indicates that the test can produce a robust and consistent result in detecting important GO terms. There is randomness in the data splitting, so we run the experiment 100 rounds with different random seeds. The average RS out of the 100 rounds, categorized by ontology groups, are reported in Table 3.5.

To illustrate the result more intuitively, Figure 3.8 shows the Venn diagrams of rejected GO terms. This chart is based on one split of the data and the GO terms in the BP group only. For example, the leftmost Venn diagram shows that the new test identified (674+49=) 723 important GO terms on the first portion of the data and (674+178=) 852 GO terms on the second portion. In this case, the robust score is 674/(49+674+178)=0.748.

Both the table and figure indicate that our new test is overwhelmingly more robust than other tests. Although there is no ground truth available in this experiment, our test shows an advantage over other methods in terms of consistently finding important GO terms.

Table 3.5: *The average Robust Score (RS) based on 100 random splits.*

Ontology	number of GO terms	ZWLm	GCTm	CQ	SKK
BP	1256	0.658	0.283	0.030	0.395
MF	165	0.506	0.255	0.027	0.261
CC	217	0.575	0.399	0.000	0.332

Another aspect we want to examine is the rejections under the null hypothesis. Within the NEG class, we randomly split the 42 observations into two subgroups with equal sample sizes. For each GO term, we apply the tests to determine the equality of means. The p-values

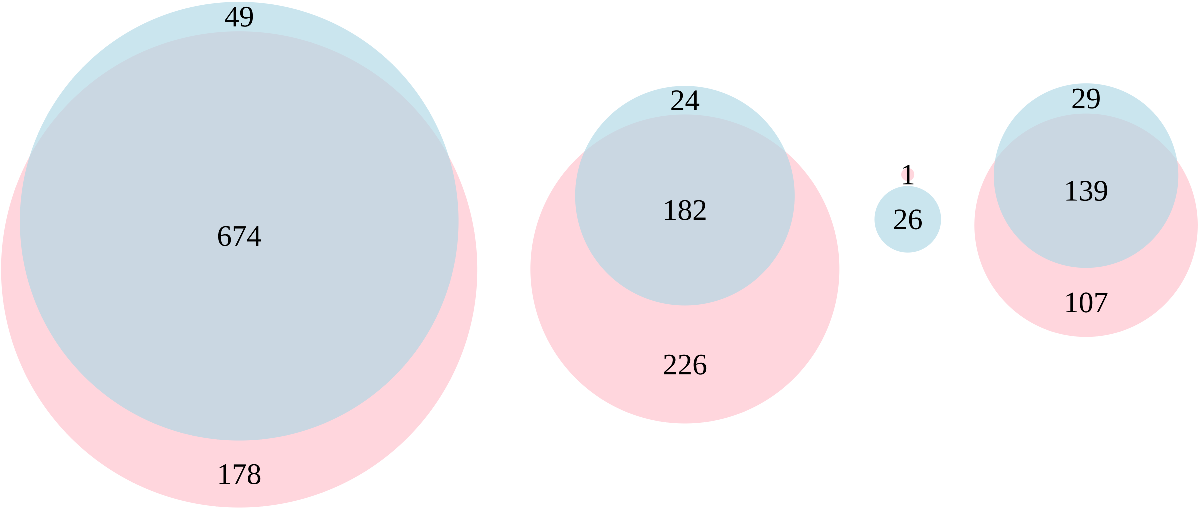


Figure 3.8: *The Venn diagram for GO term detection. From left to right they are the results for ZWLm, GCTm, CQ, SKK. The Robust Score from left to right are 0.748 0.421 0.000, and 0.505. This diagram is based on one of the random splits of the dataset for the BP ontology group. The blue and pink circles are the GO terms detected in the two subsets of the split data, and the numbers mark the counts. Larger overlap area indicates the test is more robust in detections.*

in this setting are expected to be larger than the two-class comparison above. The boxplot of the adjusted p-values is shown in Figure 3.9. As can be seen from the chart, the p-values of the new test, the CQ test and the SKK test cluster at 1, but the p-values for the GCT have more variability. In this setting, the GCT test is more likely to make type I error than others.

Combining Figure 3.7, Figure 3.9, and other experiment results, in this task of detecting differentially expressed GO terms, we can conclude that the new test can successfully control the type I error, maintain large statistical power, and produce robust detections. The GCT test is more liberal than others, and the CQ and SKK are relatively conservative. All other tests in comparison are less robust than our test.

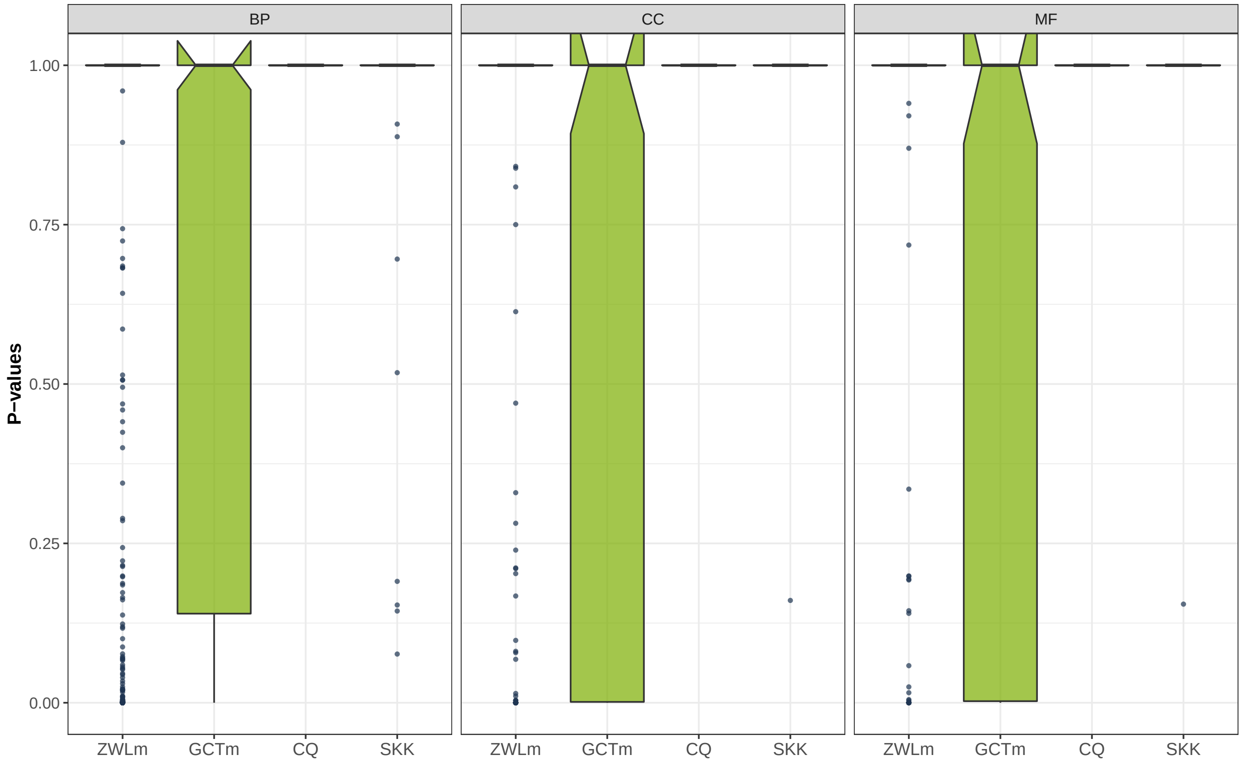


Figure 3.9: *Boxplot for adjusted p-values (Bonferroni correction) for comparing two random "NEG" sub-groups. The panels from left to right are for BP, CC and MF ontologies.*

3.7 Summary

This chapter proposed a new test for comparing the means for two high-dimensional populations. The new test statistic is based on the average of component-wise squared t-statistics, which is also used by the GCT test (Gregory et al., 2015). We suggested a new scaling parameter to overcome the drawbacks of that used by the GCT test. The new scaling parameter directly uses the sample information, does not require stationarity assumption, and achieves faster convergence to the true parameter. We established the asymptotic normality of the test statistic under H_0 and derived the power function. The simulation study showed that our test has better performance than the GCT test under all the settings. The numerical results support that using 1 as the center parameter is better than higher-order center correction. The new test is robust to moderate dependency, but very strong dependency will lead to loss of power. Moreover, the new test is robust to slightly skewed data and heavy-tailed data. In a real-data example, our new test and other existing tests were used to detect differentially expressed GO terms. The new test shows good control in type I error and more statistical power. Notably, on different datasets, our test can provide much better consistency in identifying the important GO terms.

Chapter 4

Adjusted power analysis

According to our asymptotic results, all the tests discussed in the previous two chapters should have empirical type I error rates converging to the nominal level $\alpha = 0.05$ when the sample size goes to infinity. But in some cases of our Monte Carlo simulations, especially for small sample sizes, we saw the type I error rates depart from the nominal level. This is because the converging speed is not fast enough. It brings a difficulty to the power comparison because the type I error rates for competing tests are different. To handle this issue, some existing studies suggested adjusting the power according to their actual type I error rate, for example, Zhang and Boos (1994), Lloyd (2005), and Cavus et al. (2019). We will give a brief description of their methods below.

Zhang and Boos (1994) suggested using the empirical percentile of the Monte Carlo test statistics under the null hypothesis as the critical value. Denote the test statistic as T. Suppose the Monte Carlo simulation draws B_0 independent samples from the population under the null hypothesis, and B_1 samples under the alternative hypothesis. With these samples, compute the test statistics $T_{01}, \ldots, T_{0B_0}, T_{11}, \ldots, T_{1B_1}$. Let C_α be the 100α-th upper percentile of $T_{01}, \ldots, T_{0B_0}$, then they proposed to adjust the power by

$$\text{power}_{\text{ZB}} = \frac{1}{B_1} \sum_{j=1}^{B_1} \mathcal{I}(T_{1j} > C_\alpha).$$

This method substantially changes the original test because it changes the critical value to the empirical percentile. In this way, the sampling distribution of the test statistic is generated by the Monte Carlo experiments, instead of from theoretical derivation. Besides that, this method does not fit in our case for another reason: if two tests have the same test statistic, their adjusted power will be the same. Recall that the TCFU tests have the same statistic as the classical t-test. Our high-order approximation for the sampling distribution will be ignored in their proposed procedure.

Lloyd (2005) proposed two methods for power adjustment. The first one is based on the normal approximation of the ROC curve, which describes the relationship between the power and the size. At the nominal level of α^*, the adjusted power is given by

$$\text{power}_{\text{Lloyd}} = \Phi(\Phi^{-1}(\hat{\beta}) - \Phi^{-1}(\hat{\alpha}) + \Phi^{-1}(\alpha^*)), \tag{4.1}$$

where Φ is the standard normal CDF function, $\hat{\alpha}$ is the empirical type I error rate and $\hat{\beta}$ is the empirical power. Recall that the power and type I error rate differences of TCFU, TT and classical t-test exist at $O(n^{-1/2})$ or higher order. The coarse normal approximation of the ROC curve in their method omits the $O(n^{-1/2})$ term, which will hide the high-order differences. Besides, this method tends to give unfair advantage to conservative tests. It can be explained by taking Taylor expansion at $\hat{\beta}$. Let $\Delta_n = \Phi^{-1}(\alpha^*) - \Phi^{-1}(\hat{\alpha})$. Because $\hat{\alpha}$ converges to α^*, Δ_n tends to 0 as the sample size increases. Therefore,

$$\text{power}_{\text{Lloyd}} = \Phi(\Phi^{-1}(\hat{\beta}) + \Delta_n) = \hat{\beta} + \phi(\Phi^{-1}(\hat{\beta}))\Delta_n + o_p(\Delta_n).$$

When a test is liberal, $\hat{\alpha} > \alpha^*$ and $\Delta_n < 0$, the power will be adjusted downward. When a test is conservative, $\hat{\alpha} < \alpha^*$ and $\Delta_n > 0$, the power will be boosted. Figure 4.1 gives an illustration of this adjustment method for three different $\hat{\beta}$. For any empirical power value, the power can be boosted to a high value close to 1 if the type I error rate is sufficiently close to 0. For small $\hat{\beta}$, this boosting effect is stronger. Following this method, an extremely

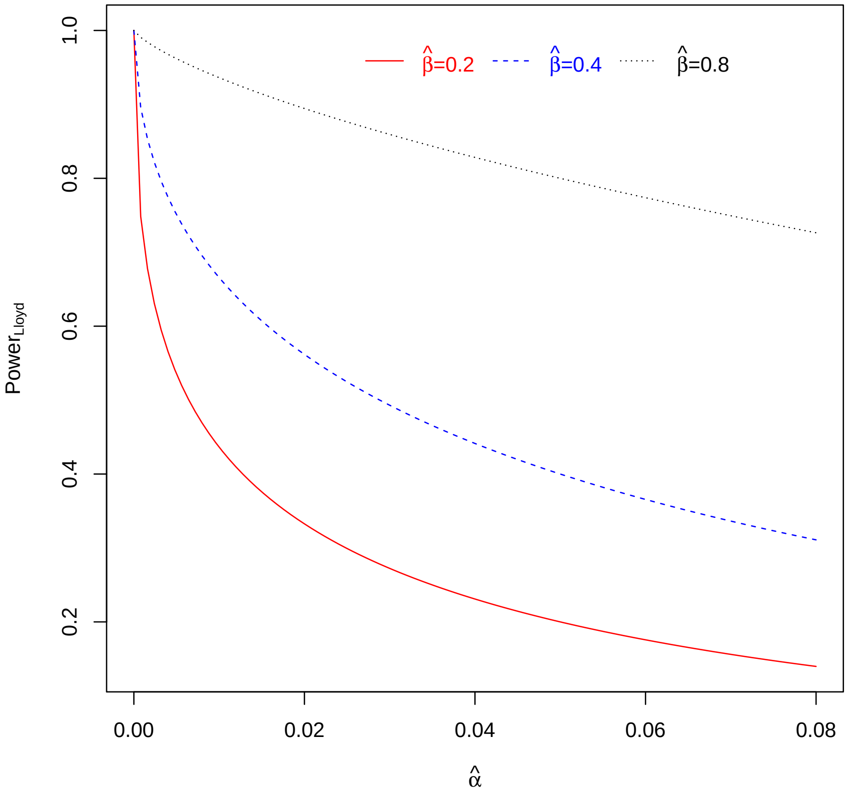


Figure 4.1: *Adjusted power at* $\alpha^* = 0.05$ *using Lloyd (2005) method 1*

conservative test is more likely to have larger adjusted power. For example, suppose we want to compare two competing tests at $\alpha^* = 0.05$. One test has $\hat{\alpha} = 0.0001, \hat{\beta} = 0.2$, where the adjusted power is 0.89; the other test has $\hat{\alpha} = 0.05, \hat{\beta} = 0.8$, where the adjusted power is only 0.8. But in reality, the first test is not useful because it does not provide any power. For hypothesis testing, the nominal level α^* prescribes the upper limit of the type I error rate. As long as a test can control the type I error rate within α^*, it is a reasonable test, and there is no need for any adjustment.

Lloyd (2005) also proposed another method to compare competing tests using the partial area under the ROC curve (PAUC) instead of power. It involves computation of Mann-Whitney statistic for the simulated test statistics $T_{01}, \ldots, T_{0B_0}$ under H_0 and $T_{11}, \ldots, T_{1B_1}$ under H_a. If the relative ranks of those test statistics do not change, the PAUC will also be the same. The TT tests have transformed t-statistic $G(T)$ as their test statistics. Those transformation functions are all monotone and the ranks for $G(T_{01}), \ldots, G(T_{0B_0}), G(T_{11}), \ldots, G(T_{1B_1})$ will be the same as the ranks of the original $T_{01}, \ldots, T_{0B_0}, T_{11}, \ldots, T_{1B_1}$. Using this method, the TT tests and the classical t-test will have the same PAUC. But from the theoretical analysis in Section 2.3, we know their power functions are different. Additionally, the TCFU tests and the classical t-test have the same test statistics but different critical values. Using Lloyd PAUC power adjustment, they will have identical power. This is not right because their power functions are calculated from different distributions.

Cavus et al. (2019) proposed a penalization approach and suggested to adjust the power by

$$\text{power}_{\text{CYS}} = \frac{\hat{\beta}}{\sqrt{1 + |1 - \frac{\hat{\alpha}}{\alpha^*}|}}. \tag{4.2}$$

It penalizes any departure of the type I error $\hat{\alpha}$ from the nominal level α^* from both sides. In fact, equation (4.1) also can be seen as a penalization function. The main difference between Lloyd (2005) and Cavus et al. (2019) is whether to penalize or boost the power when $\hat{\alpha} < \alpha^*$. As we have pointed out in the discussion about Lloyd (2005) method 1, the

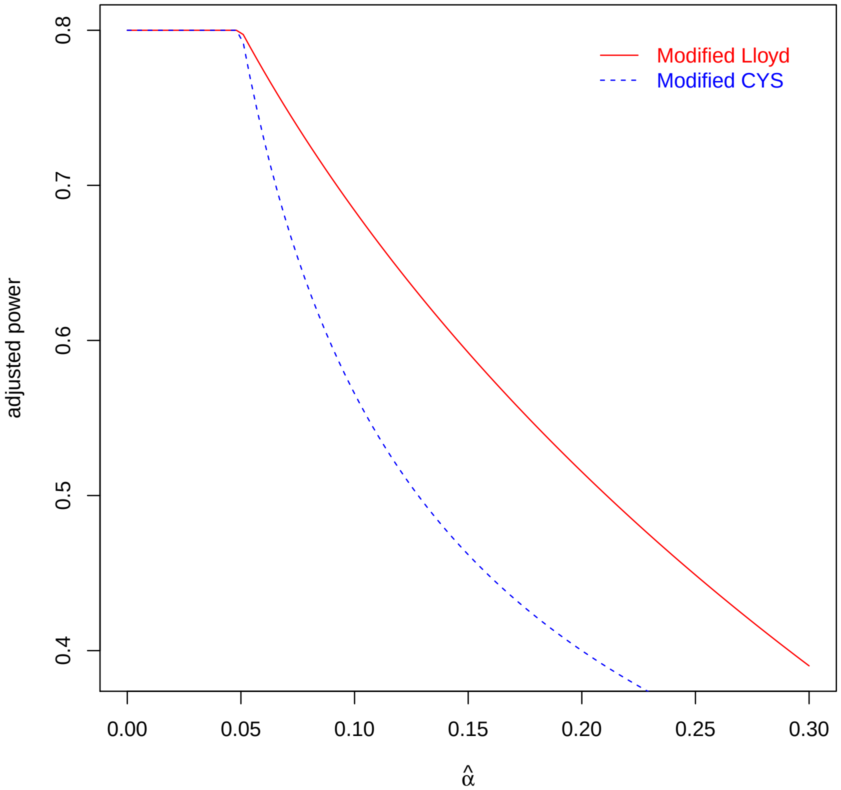


Figure 4.2: *Comparison of different penalty functions for power adjustment. "Modified Lloyd" is for the adjustment using (4.3), and "Modified CYS" is for (4.4). The empirical power is* $\hat{\beta} = 0.8$*. The nominal level is* $\alpha^* = 0.05$.

type I error rates within α^* should be deemed acceptable, so the power of conservative tests should not be further penalized.

Here we propose an asymmetric power penalization method by slightly adjusting (4.1) or (4.2) as follows:

$$\text{power}_{\text{Lloyd\_new}} = \Phi(\Phi^{-1}(\hat{\beta}) - \Phi^{-1}(\max(\hat{\alpha}, \alpha^*)) + \Phi^{-1}(\alpha^*)), \tag{4.3}$$

and

$$\text{power}_{\text{CYS\_new}} = \frac{\hat{\beta}}{\sqrt{1 + |1 - \frac{\max(\hat{\alpha},\alpha^*)}{\alpha^*}|}}. \tag{4.4}$$

They only impose penalty on power when $\hat{\alpha} > \alpha^*$, and keep the original power when $\hat{\alpha} < \alpha^*$.

These different functions penalize the power at different degrees when the empirical type I error rate approaches the nominal level. Figure 4.2 gives an example for $\hat{\beta} = 0.8$. The function (4.3) adds smaller penalty than (4.4) at $\hat{\alpha}$ close to α. This makes more sense because the empirical type I error is an estimate which has variation and naturally fluctuates around the nominal level. It is not reasonable to impose a stronger penalty for smaller departure.

4.1 Power adjustment for Chapter 2

Using (4.3), the adjusted empirical power for simulation setting 1 is displayed in Table 4.1-4.3.

Table 4.1: *The adjusted power for **upper-tailed** alternative in setting 1 at level* $\alpha^* = 0.05$.

n_1	TCFU_1	TCFU_2	TT_1	TT_2	TT_3	TT_4	t
15	0.421	0.345	0.460	0.359	0.476	0.468	0.249
25	0.595	0.545	0.634	0.524	0.650	0.638	0.424
40	0.792	0.772	0.823	0.750	0.828	0.824	0.674
50	0.832	0.841	0.865	0.815	0.872	0.864	0.770
80	0.961	0.961	0.967	0.956	0.968	0.967	0.942
120	0.992	0.992	0.993	0.991	0.994	0.993	0.989
160	0.999	1.000	0.999	1.000	1.000	0.999	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Based on Table 4.1, in the upper-tailed scenario, the TT_1, TT_3 and TT_4 tests have the highest adjusted power for most of the sample sizes, while t-test has the lowest adjusted power. Recall from Table 2.3 that t-test is very conservative in this simulation setting, while TT_1, TT_3 and TT_4 successfully control the type I error rates. The result shows that TT_4, TT_1 and TT_3 are more powerful than TT_2, which is consistent with the theoretical result in Table 2.1.

For the lower-tailed and two-sided H_a in setting 1, we only reported the power when sample sizes are relatively large (Table 2.4, Table 2.5, Figure 2.2, and Figure 2.3) and did not show the empirical power for small sample sizes due to inflated type I error rates. Now, we can evaluate them with the power adjustment technique (see Table 4.2 and 4.3). For the lower-tailed alternative, classical t and TT_3 have high power. For the two-sided alternative, TT_1, TT_3 and TT_4 have high power, but t-test shows very low power compared to TCFU and TT tests.

Table 4.2: *The adjusted power for* ***lower-tailed*** *alternative in setting 1 at level* $\alpha^* = 0.05$.

n_1	TCFU_1	TCFU_2	TT_1	TT_2	TT_3	TT_4	t
15	0.225	0.216	0.209	0.236	0.076	0.214	0.239
25	0.311	0.299	0.288	0.330	0.352	0.295	0.341
40	0.433	0.417	0.394	0.440	0.465	0.408	0.451
50	0.472	0.451	0.432	0.484	0.497	0.447	0.490
80	0.628	0.609	0.587	0.658	0.665	0.599	0.666
120	0.754	0.737	0.718	0.773	0.777	0.730	0.784
160	0.863	0.846	0.833	0.884	0.889	0.843	0.886
250	0.936	0.930	0.919	0.946	0.951	0.926	0.952

Table 4.3: *The adjusted power for* ***two-sided*** *alternative in setting 1 at level* $\alpha^* = 0.05$.

n_1	TCFU_1	TCFU_2	TT_1	TT_2	TT_3	TT_4	t
15	0.218	0.121	0.207	0.101	0.396	0.205	0.049
25	0.373	0.289	0.387	0.210	0.535	0.391	0.124
40	0.630	0.578	0.648	0.480	0.715	0.647	0.367
50	0.690	0.659	0.725	0.569	0.778	0.719	0.484
80	0.910	0.913	0.927	0.857	0.936	0.925	0.800
120	0.974	0.977	0.980	0.960	0.984	0.981	0.955
160	0.999	0.998	0.999	0.998	0.999	0.999	0.995
250	1.000	1.000	1.000	1.000	1.000	1.000	1.000

4.2 Power adjustment for Chapter 3

Figure 3.1 and 3.2 are sufficient for the discussion on the different versions of center corrections, so we did not further adjust the power. The empirical type I error rates for Figure 3.3-3.6 are very different for competing tests, so we adjust the power using formula (4.3) and report them in Figure 4.3-4.6.

In the standard Normal, Gamma, and $t(3)$ cases, the new test and SKK test are almost always more powerful than the other two tests except the strong dependence setting, where the new test and CQ test have more power. In the Cauchy innovation setting, our test has

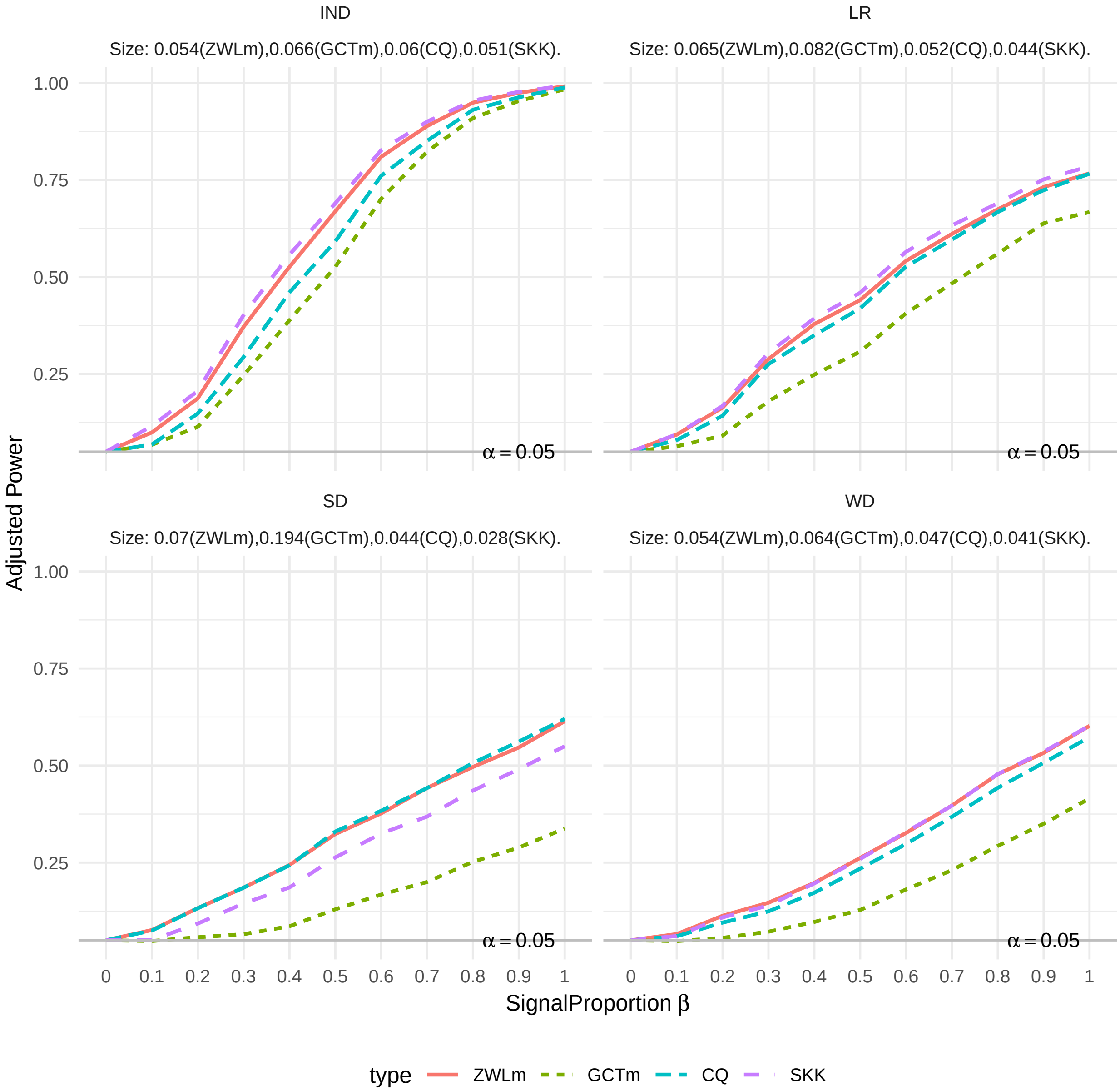


Figure 4.3: *The adjusted power of ZWLm, GCTm, CQ, and SKK tests when the innovation follows the* ***standard normal distribution****. Other configurations are the same as Figure 3.3. The "Size" on top of each graph shows the original empirical type I error rates. The signal magnitude is $\delta = 0.125$ for IND, WD and LR, and 0.375 for SD.*

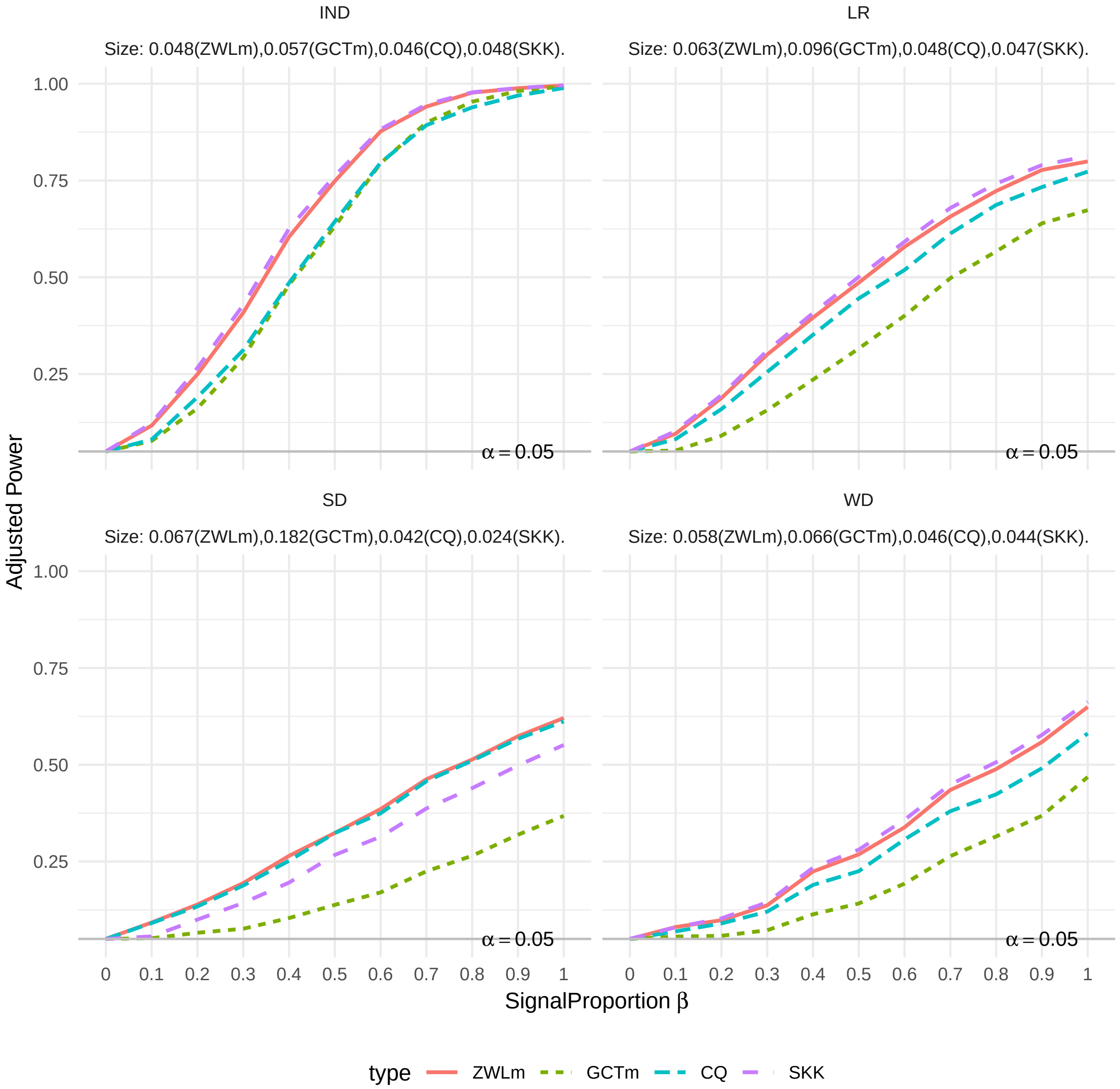


Figure 4.4: *The adjusted power of ZWLm, GCTm, CQ, and SKK tests when the innovation follows the* ***shifted Gamma(4,2) distribution****. Other configurations are the same as Figure 3.4. The "Size" on top of each graph shows the original empirical type I error rates. The signal magnitude is $\delta = 0.5$ for IND, WD and LR, and 1.5 for SD.*

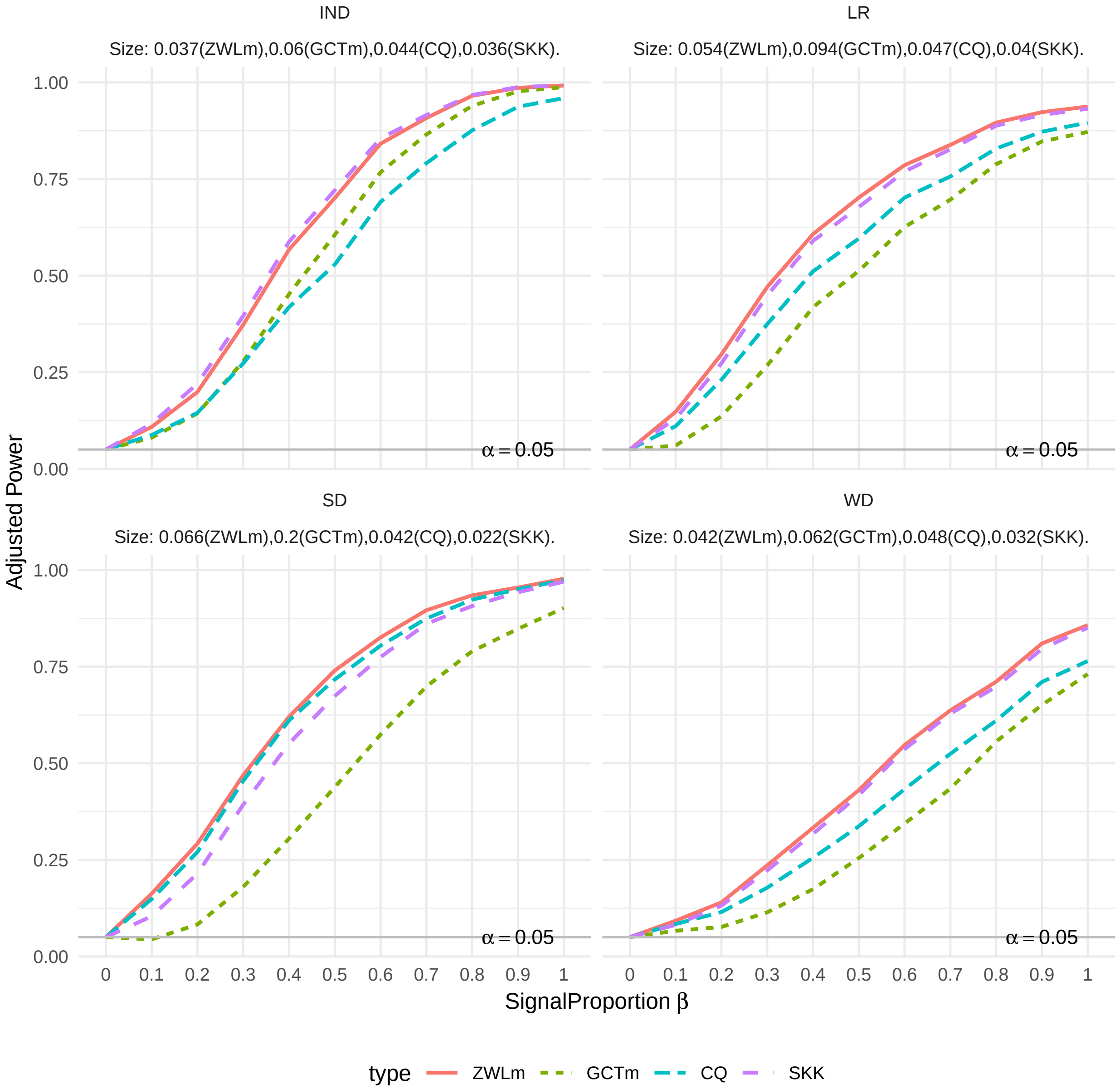


Figure 4.5: *The adjusted power of ZWLm, GCTm, CQ, and SKK tests when the innovation follows the* ***t(3) distribution****. Other configurations are the same as Figure 3.5. The "Size" on top of each graph shows the original empirical type I error rates. The signal magnitude is* $\delta = 0.2$ *for IND, 0.25 for WD and LR, and 1 for SD.*

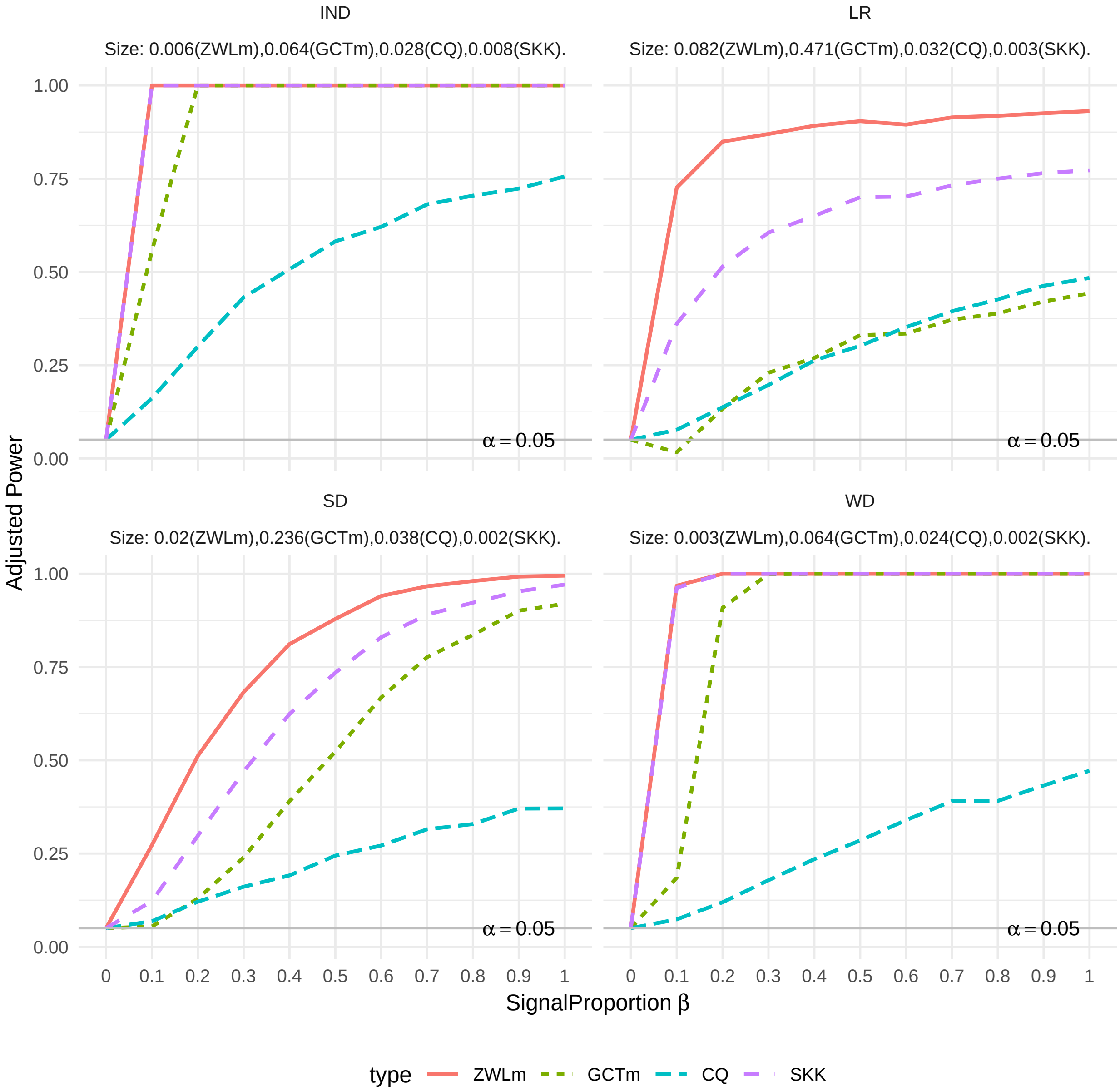


Figure 4.6: *The adjusted power of ZWLm, GCTm, CQ, and SKK tests when the innovation follows the* ***Cauchy(0,0.1) distribution****. Other configurations are the same as Figure 3.6. The "Size" on top of each graph shows the original empirical type I error rates. The signal magnitude is $\delta = 1$ for IND, WD and LR, and 3 for SD.*

an obvious advantage over other tests for long-range and strong dependence settings and maintains the same high power as SKK in independent and weak dependent settings.

4.3 Summary

The patterns in the adjusted power are mostly consistent with our previous findings. Although the power adjustment technique provides us an alternative viewpoint, we still recommend examining the type I error rates carefully when reading and interpreting the adjusted power.

Summary

This dissertation concerns about two-sample mean tests for univariate and high-dimensional populations. In Chapter 2, we proposed the TCFU and the TT tests for testing the equality of means of two independent univariate populations. The TCFU tests extended the TCF test (Wang et al., 2017) to the unequal-variance scenario. The TT tests considered four transformations including one correcting a transformation proposed by Hall (1992b). Through the theoretical power functions, we compared the new tests in terms of power and type I error and derived the analytical conditions leading to high power. These conditions depend on the relative skewness, the pooled kurtosis, and the adjusted effect size. For one-sided alternatives, the TCFU$_2$, TT$_1$ and TT$_4$ tests have type I error rate converging to the nominal level faster than that of the TT$_2$ and TT$_3$ tests. The power ranking $TCFU_1 \succ TT_4 \succ TT_1$ always holds for all types of alternatives. We also presented the coverage probabilities accurate to $O(n^{-1})$ for two-sided transformation-based confidence intervals. Using these theoretical results, we can provide a more rigorous explanation of the simulation results obtained by Zhou and Dinh (2005). Monte Carlo simulation studies showed that no test can achieve the best performance for all scenarios. An example from genetic studies demonstrated that the TCFU and TT tests can identify more significantly different genes than the t-test and successfully control the type I error.

Although Bootstrap-t procedure amounts to the infinite-order Edgeworth expansion, where moments are replaced by plug-in estimators, it can only achieve the same order of accuracy as the TCFU procedure does. The asymptotic properties of the TCFU tests can provide an appealing alternative to estimate the power of the Bootstrap-t test. The power of the bootstrapping test could not be computed with simple bootstrap because it is always conducted under H_0. Moreover, the TCFU tests are much more efficient with computation,

which is especially important when testing a large number of hypotheses.

In Chapter 3, we proposed a new test for comparing means for two high-dimensional populations. The new test statistic is based on the average of component-wise squared t-statistics, which is also used by the GCT test (Gregory et al., 2015). We suggested a new scaling parameter to overcome the drawbacks of that used by the GCT test. The new scaling parameter directly uses the sample information, does not require stationarity assumption, and achieves faster convergence to the true parameter. We established the asymptotic normality of the test statistic under H_0 and derived the power function. The simulation study showed that our test has better performance than the GCT test under all the settings. The numerical results support that using 1 as the center parameter is better than higher-order center correction. The new test is robust to moderate dependency, but very strong dependency will lead to loss of power. Moreover, the new test is robust to slightly skewed data and heavy-tailed data. With acute lymphoblastic leukemia gene expression data, we demonstrated the new tests can be used to give more consistent results in detecting differently expressed Gene Ontology terms than competing tests.

In the last part of the dissertation, we considered power adjustments to address a question of how to fairly compare the power of competing methods in simulation studies when they have different empirical type I error rates. After discussing some existing methods and their drawbacks, we proposed to modify two existing methods and give an asymmetric penalty for the departures of type I error rates from the nominal level. The new power adjustment method was used to compare the simulation results in the previous two chapters of the dissertation. This asymmetric power adjustment method can have broader applications in power comparison for future studies.

Bibliography

Lavy Abramovitch and Kesar Singh. Edgeworth corrected pivotal statistics and the bootstrap. *The Annals of Statistics*, 13(1):116–132, 1985.

Lihua An and Ejaz S. Ahmed. Improving the performance of kurtosis estimator. *Computational Statistics & Data Analysis*, 52(5):2669–2681, 2008.

Krishna B Athreya and Soumendra N Lahiri. *Measure theory and probability theory*. Springer Science & Business Media, 2006.

Zhidong Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329, 1996.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 08 2001.

Rabi N Bhattacharya, Jayanta K Ghosh, et al. On the validity of the formal edgeworth expansion. *The Annals of Statistics*, 6(2):434–451, 1978.

Patrick Billingsley. *Probability and measure*. Wiley series in probability and mathematical statistics. Wiley, New York, 1995.

Yvonne Bishop, Stephen Fienberg, and Paul Holland. *Discrete multivariate analysis: Theory and practice*. MIT Press, Cambridge, MA, 2007.

Clifford R. Blair, James J. Higgins, Walt Karniski, and Jeffrey D. Kromrey. A study of multivariate permutation tests which may replace hotelling's t2 test in prescribed circumstances. *Multivariate Behavioral Research*, 29(2):141–163, 1994.

Peter Brockwell and Richard Davis. *Time Series: Theory and Methods*. Springer-Verlag New York, 1991.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Tony Cai, Weidong Liu, and Yin Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372, 2014.

Mustafa Cavus, Berna Yazici, and Ahmet Sezer. Penalized power approach to compare the power of the tests when type i error probabilities are different. *Communications in Statistics - Simulation and Computation*, 0(0):1–15, 2019. doi: 10.1080/03610918.2019.1588310.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835, 2010.

Sabina Chiaretti, Xiaochun Li, Robert Gentleman, Antonella Vitale, Marco Vignetti, Franco Mandelli, Jerome Ritz, and Robin Foa. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778, 2004. ISSN 0006-4971.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

Yu A Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory of Probability & Its Applications*, 13(4):691–696, 1968.

A. P. Dempster. A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 29(4):995–1010, 1958.

D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, May 1995. ISSN 0018-9448. doi: 10.1109/18.382009.

Sandrine Dudoit, Sündüz Keleş, and Mark J. van der Laan. *Multiple tests of association with biological annotation metadata*, pages 153–218. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

Weibo Gong, Yong Liu, Vishal Misra, and Don Towsley. On the tails of web file size distributions. *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, 39(1):192–201, 2001.

Karl B. Gregory, Raymond J. Carroll, Veerabhadran Baladandayuthapani, and Soumendra N. Lahiri. A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, 110(510):837–849, 2015.

Peter Hall. *The bootstrap and Edgeworth expansion.* Springer Series in Statistics, New York, 1992a.

Peter Hall. On the removal of skewness by transformation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):221–228, 1992b.

Peter Hall, Bing-Yi Jing, and Soumendra Nath Lahiri. On the sampling window method for long-range dependent data. *Statistica Sinica*, 8(4):1189–1204, 1998.

Jiang Hu and Zhidong Bai. A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices. *Science China Mathematics*, 59(12): 2281–2300, 2016.

Xiaochun Li. *ALL: A data package*, 2009. R package version 1.22.0.

Chris J. Lloyd. Estimating test power adjusted for size. *Journal of Statistical Computation and Simulation*, 75(11):921–933, 2005. doi: 10.1080/00949650412331321160.

Miles E. Lopes, Laurent Jacob, and Martin J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 1206–1214, USA, 2011. Curran Associates Inc.

Dimitris N. Politis and Joseph P. Romano. Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis*, 16(1):67–103, 1995.

Konstantinos Psounis, Pablo Molinero-Fernández, Balaji Prabhakar, and Fragkiskos Papadopoulos. Systems with multiple servers under heavy-tailed workloads. *Performance Evaluation*, 62(1-4):456–474, 2005.

Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016.

Gennady Samorodnitsky. Long range dependence. *Foundations and Trends® in Stochastic Systems*, 1(3):163–257, 2007.

Muni S. Srivastava and Meng Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386 – 402, 2008.

Muni S. Srivastava, Shota Katayama, and Yutaka Kano. A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358, 2013.

Radhendushka Srivastava, Ping Li, and David Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970, 2016.

Måns Thulin. A high-dimensional two-sample test for the mean using random subspaces. *Computational Statistics & Data Analysis*, 74:26–38, 2014.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA*, 99:6567 – 6572, 2002.

Bo Tong. *More accurate two sample comparisons for skewed populations*. PhD thesis, Kansas State University, 2016.

Dimitri Van De Ville, Thierry Blu, and Michael Unser. Integrated wavelet processing and spatial statistical testing of fmri data. *NeuroImage*, 23(4):1472–1485, 2004.

Haiyan Wang. *Testing in multifactor heteroscedastic ANOVA and repeated measures designs with large number of levels*. PhD thesis, Pennsulvania State University, 2004.

Haiyan Wang and Michael G Akritas. Inference from heteroscedastic functional data. *Journal of Nonparametric Statistics*, 22(2):149–168, 2010.

Haiyan Wang, Bo Tong, Huaiyu Zhang, and Xukun Li. New two-sample tests for skewed populations and their connection to theoretical power of bootstrap-t test. *TEST*, 26(3): 661–683, 2017.

Yujun Wu, Marc G. Genton, and Leonard A. Stefanski. A multivariate two-sample mean test for small sample size and missing data. *Biometrics*, 62(3):877–885, 2006.

Gongjun Xu, Lifeng Lin, Peng Wei, and Wei Pan. An adaptive two-sample test for high-dimensional means. *Biometrika*, 103(3):609–624, 2016.

Jin Xu. Asymptotic expansion of the non null distribution of the two-sample t-statistic under non normality with application in power comparison. *Communications in Statistics—Theory and Methods*, 39(11):1915–1921, 2010.

Jin Xu, Xinping Cui, and Arjun K. Gupta. Improved statistics for contrasting means of two samples under non-normality. *British Journal of Mathematical and Statistical Psychology*, 62(1):21–40, 2009.

Ji Zhang and Dennis D Boos. Adjusted power estimates in monte carlo experiments. *Communications in Statistics - Simulation and Computation*, 23(1):165–173, 1994. doi: 10.1080/03610919408813162.

Xiao Hua Zhou and Phillip Dinh. Nonparametric confidence intervals for the one-and two-sample problems. *Biostatistics*, 6(2):187–200, 2005.

Roger S. Zoh, Abhra Sarkar, Raymond J. Carroll, and Bani K. Mallick. A powerful bayesian test for equality of means in high dimensions. *Journal of the American Statistical Association*, 113(524):1733–1741, 2018. doi: 10.1080/01621459.2017.1371024.

Appendix A

Appendices for Chapter 2

A.1 The expressions of $p_1(x)$, $p_2(x)$, $p_{10}(x)$, and $p_{20}(x)$

Based on the result provided by Xu (2010), $p_1(x) = -\sum_{j=1}^{3} a_j H_{j-1}(x)$, and $p_2(x) = -\sum_{j=1}^{6} b_j H_{j-1}(x)$ where $H_1(x) = x$, $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, $H_5(x) = x^5 - 10x^3 + 15x$ are Hermite polynomials. Let

$$A = \left(\frac{\sigma_1^2}{\lambda_1} + \frac{\sigma_2^2}{\lambda_2}\right)^{-3/2} \left(\frac{\sigma_1^3\gamma_1}{\lambda_1^2} - \frac{\sigma_2^3\gamma_2}{\lambda_2^2}\right),$$

$$B_1 = \left(\frac{\sigma_1^2}{\lambda_1} + \frac{\sigma_2^2}{\lambda_2}\right)^{-2} \left(\frac{\sigma_1^4\tau_1}{\lambda_1^3} + \frac{\sigma_2^4\tau_2}{\lambda_2^3}\right),\ B_2 = \left(\frac{\sigma_1^2}{\lambda_1} + \frac{\sigma_2^2}{\lambda_2}\right)^{-2} \left(\frac{\sigma_1^4}{\lambda_1^3} + \frac{\sigma_2^4}{\lambda_2^3}\right).$$

The coefficients are

$$a_1 = -A/2,\ a_2 = -Aw/2,\ a_3 = -A/3,$$

$$b_1 = 3w(B_1 + 2B_2)/8,\ b_2 = A^2 + B_2 + w^2(B_1 + 2B_2)/8,\ b_3 = 7A^2w/8 + B_2w/2,$$

$$b_4 = -B_1/12 + 2A^2/3 + B_2/2 + A^2w^2/8,\ b_5 = A^2w/6,\ b_6 = A^2/18.$$

Under H_0, $w = 0$, $p_{10}(x) = -\sum_{j=1}^{3} a_{j0} H_{j-1}(x) = (2x^2+1)A/6$, and $p_{20}(x) = -\sum_{j=1}^{6} b_{j0} H_{j-1}(x)$ where $a_{10} = -A/2,\ a_{20} = 0,\ a_{30} = -A/3,\ b_{10} = b_{30} = b_{50} = 0,\ b_{20} = A^2 + B_2,\ b_{40} =$

$-B_1/12 + 2A^2/3 + B_2/2$, and $b_{60} = A^2/18$.

A.2 Regularity conditions

Suppose the true mean of X_{ij} is μ_i and let $Y^*_{ij} = \frac{X_{ij}-\mu_i}{\sigma_i}$, $\bar{Y}^*_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y^*_{ij}$ for $i = 1, 2$ and $j = 1, ..., n_i$. The existence of the Edgeworth expansion requires the following moment and Cramér's conditions. They were referred as regularity conditions in the theorem.

(C.1) (Moment condition) $E(Y^*_{ij})^8 < \infty$, $i = 1, 2$;

(C.2) (Cramér's condition) $\limsup_{||(t_1,t_2)||\to\infty} |\chi_i(t_1, t_2)| < 1$ for $i = 1, 2$, where $\chi_i(t_1, t_2)$ is the characteristic function of $(Y^*_{i1}, Y^{*2}_{i1})'$.

(C.3) Assume the characteristic function χ^X_i of the sample X_{i1} satisfies $\int_{-\infty}^{\infty} |\chi^X_i|^c < \infty$ for some $c \geq 1$, for $i = 1, 2$.

The moment condition (C.1) is needed because computation of the fourth cumulant of the test statistic involves term $E[n_i^{-1}\sum_{j=1}^{n_i} Y^{*2}_{ij} - 1]^4$. The Cramér's condition implies that any atoms of the distribution of $(\overline{Y}^*_i, n_i^{-1}\sum_{j=1}^{n_i} Y^{*2}_{ij})'$ have exponentially small mass so that the distribution of $\sqrt{n_i}A(\overline{Y}^*_i, n_i^{-1}\sum_{j=1}^{n_i} Y^{*2}_{ij})$ are virtually continuous (see page 57 of Hall (1992a)), where A is a function with fourth continuous derivatives in a neighborhood of $(E(Y^*_{i1}), E(Y^{*2}_{i1}))' = (0, 1)'$. The condition (C.3) ensures that $\overline{X}_{i.}$ has a bounded density (see page 78 of Hall (1992a)).

A.3 Proofs

A.3.1 Proof for Theorem 1

Proof. First we consider the TCFU$_2$ test, i.e., the critical value is $\hat{\xi}_{2,\alpha}$. Note that under H$_a$

$$\begin{aligned} P(T_n \leq \hat{\xi}_{2,\alpha}) &= P(T_n - [\hat{\xi}_{2,\alpha} - \xi_{2,\alpha}] \leq \xi_{2,\alpha}) \\ &= P(T_n - [\hat{\xi}_{1,\alpha} - \xi_{1,\alpha} + n^{-1}(\hat{q}_{20}(z_\alpha) - q_{20}(z_\alpha))] \leq \xi_{2,\alpha}). \end{aligned} \tag{A.1}$$

Since $\hat{\gamma}_i - \gamma_i = O_p(n_i^{-1/2})$ and $\hat{\tau}_i - \tau_i = O_p(n_i^{-1/2})$, we know that $\hat{\xi}_{1,\alpha} - \xi_{1,\alpha} = O_p(n^{-1})$ and $\hat{q}_{20}(z_\alpha) - q_{20}(z_\alpha) = O_p(n^{-1/2})$.

Applying the Delta method in Section 2.7 of Hall (1992a) to (A.1), we have

$$P(T_n \leq \hat{\xi}_{2,\alpha}) = P(T_n - [\hat{\xi}_{1,\alpha} - \xi_{1,\alpha}] \leq \xi_{2,\alpha}) + O(n^{-3/2}),$$

Denote

$$R_n = T_n - [\hat{\xi}_{1,\alpha} - \xi_{1,\alpha}] = T_n + n^{-1}\Delta_n, \tag{A.2}$$

where

$$\Delta_n = -n[\hat{\xi}_{1,\alpha} - \xi_{1,\alpha}] = -n^{1/2}[\hat{q}_{10}(z_\alpha) - q_{10}(z_\alpha)] = n^{1/2}(2z_\alpha^2 + 1)(\hat{A} - A)/6$$

Notice that both R_n and T_n are smooth functions of the sample means of the standardized random variables $(X_{ij} - \mu_i)/\sigma_i$ and $(X_{ij} - \mu_i)^2/\sigma_i^2$ with all the derivatives. By Theorem 2 of Bhattacharya et al. (1978), the Edgeworth expansion of R_n and T_n can be obtained by formally inverting the characteristic function under the assumptions of the theorem.

The cumulants of R_n have the following relationship with the cumulants of T and

$E(T_0^k \Delta_n)$, where $k = 1, 2, 3$, and T_0 is T_n under the null hypothesis:

$$\begin{aligned}
\kappa_1(R_n) &= \kappa_1(T_n) + O(n^{-3/2}), \\
\kappa_2(R_n) &= \kappa_2(T_n) + 2n^{-1}E(T_0\Delta_n) + O(n^{-2}), \\
\kappa_3(R_n) &= \kappa_3(T_n) + 3n^{-1}E(T_0^2\Delta_n) + O(n^{-3/2}), \\
\kappa_4(R_n) &= \kappa_4(T_n) + 4n^{-1}[E(T_0^3\Delta_n) - 3E(T_0^2)E(T_0\Delta_n)] + O(n^{-2})
\end{aligned}$$

Further calculation gives $E(T_0\Delta_n) = c_\alpha + O(n^{-1})$, where $c_\alpha = (B/6 - A^2/4)(2z_\alpha^2 + 1)$, $E(T_0^2\Delta_n) = O(n^{-1/2})$, and $E(T_0^3\Delta_n) - 3E(T_0^2)E(T_0\Delta_n) = O(n^{-1})$. Only $\kappa_2(R_n)$ differs from $\kappa_2(T)$ by an order less than $O(n^{-3/2})$. Accounting for the change of κ_2 in the derivation of Edgeworth expansion, we have

$$P(R_n \leq x) = P(T_n \leq x) - n^{-1}c_\alpha(x - w)\phi(x - w) + O(n^{-3/2}),$$

Therefore,

$$P(T_n \leq \hat{\xi}_{2,\alpha}) = P(R_n \leq \xi_{2,\alpha}) + O(n^{-3/2}) = F_{T,w}(\xi_{2,\alpha}) - n^{-1}c_\alpha(\xi_{2,\alpha} - w)\phi(\xi_{2,\alpha} - w) + O(n^{-3/2}).$$

Because $\xi_{2,\alpha} = z_\alpha + O(n^{-1/2})$, applying the Taylor expansion, $\phi(\xi_{2,\alpha} - w) = \phi(z_\alpha - w) + O(n^{-1/2})$, we have

$$n^{-1}c_\alpha(\xi_{2,\alpha} - w)\phi(\xi_{2,\alpha} - w) = n^{-1}c_\alpha(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2}).$$

Above results can be applied directly to $P(T \leq \hat{\xi}_{1,\alpha})$ to get

$$P(T_n \leq \hat{\xi}_{1,\alpha}) = F_{T,w}(\xi_{1,\alpha}) - n^{-1}c_\alpha(\xi_{1,\alpha} - w)\phi(\xi_{1,\alpha} - w) + O(n^{-3/2}).$$

□

A.3.2 Proof for Corollary 1

Proof. Using the result of Theorem 1, we have

$$
\begin{aligned}
& P(T \leq \hat{\xi}_{2,\alpha}) - P(T \leq \hat{\xi}_{1,\alpha}) \\
= \; & F_{T,w}(\xi_{2,\alpha} - w) - F_{T,w}(\xi_{1,\alpha} - w) + O(n^{-3/2}) \\
= \; & \Phi(\xi_{2,\alpha} - w) + n^{-1/2} p_1(\xi_{2,\alpha} - w)\phi(\xi_{2,\alpha} - w) + n^{-1} p_2(\xi_{2,\alpha} - w)\phi(\xi_{2,\alpha} - w) \\
& -[\Phi(\xi_{1,\alpha} - w) + n^{-1/2} p_1(\xi_{1,\alpha} - w)\phi(\xi_{1,\alpha} - w) + n^{-1} p_2(\xi_{1,\alpha} - w)\phi(\xi_{1,\alpha} - w)] + O(n^{-3/2})
\end{aligned}
$$

Note that

$$p_1(\xi_{2,\alpha} - w) = p_1(\xi_{1,\alpha} - w) + p_1'(\xi_{1,\alpha} - w)(\xi_{2,\alpha} - \xi_{1,\alpha}) + o(\xi_{2,\alpha} - \xi_{1,\alpha}) = p_1(\xi_{1,\alpha} - w) + O(n^{-1}),$$

and

$$\phi(\xi_{2,\alpha} - w) = \phi(\xi_{1,\alpha} - w) + \phi'(\xi_{1,\alpha} - w)(\xi_{2,\alpha} - \xi_{1,\alpha}) + o((\xi_{2,\alpha} - \xi_{1,\alpha})) = \phi(\xi_{1,\alpha} - w) + O(n^{-1}).$$

Hence using the Taylor expansion, we have

$$
\begin{aligned}
P(T \leq \hat{\xi}_{2,\alpha}) - P(T \leq \hat{\xi}_{1,\alpha}) &= \Phi(\xi_{2,\alpha} - w) - \Phi(\xi_{1,\alpha} - w) + O(n^{-3/2}) \\
&= \phi(\xi_{1,\alpha} - w) n^{-1} q_{20}(z_\alpha) + O(n^{-3/2}) \\
&= \phi(z_\alpha - w) n^{-1} q_{20}(z_\alpha) + O(n^{-3/2})
\end{aligned}
$$

□

A.3.3 Proof of Theorem 2

Proof. The inverse of function $G_1(U) = U + \hat{A}U^2/3 + \hat{A}^2U^3/27 + \hat{A}/(6n)$ is $G_1^{-1}(t) = (\hat{A}/3)^{-1}[1 + \hat{A}(t - \hat{A}/(6n))]^{1/3} - (\hat{A}/3)^{-1}$. Then the CDF of the test statistic of TT_1 test is

$$\begin{aligned}
& P\left(\sqrt{n}G_1\left(\frac{T_n}{\sqrt{n}}\right) < x\right) \\
= \; & P\left(T_n < \sqrt{n}\left(\frac{\hat{A}}{3}\right)^{-1}\left[1 + \hat{A}\left(\frac{x}{\sqrt{n}} - \frac{\hat{A}}{6n}\right)\right]^{1/3} - \sqrt{n}\left(\frac{\hat{A}}{3}\right)^{-1}\right) \\
= \; & P\left(T_n < \sqrt{n}\left(\frac{\hat{A}}{3}\right)^{-1}\left[1 + \frac{\hat{A}}{3}\left(\frac{x}{\sqrt{n}} - \frac{\hat{A}}{6n}\right) - \frac{\hat{A}^2}{9}\left(\frac{x}{\sqrt{n}} - \frac{\hat{A}}{6n}\right)^2 + \frac{5\hat{A}^3}{81}\left(\frac{x}{\sqrt{n}} - \frac{\hat{A}}{6n}\right)^3\right.\right. \\
& \left.\left. + O_p(n^{-2})\right] - \sqrt{n}\left(\frac{\hat{A}}{3}\right)^{-1}\right) \\
= \; & P\left(T_n < \sqrt{n}\left(\frac{\hat{A}}{3}\right)^{-1}\left[\frac{\hat{A}x}{3\sqrt{n}} - \frac{\hat{A}^2}{18n} - \frac{\hat{A}^2}{9}\left(\frac{x^2}{n} - \frac{\hat{A}x}{3n^{3/2}}\right) + \frac{5\hat{A}^3x^3}{81n^{3/2}} + O_p(n^{-2})\right]\right) \\
= \; & P\left(T_n < x - \frac{\hat{A}}{6\sqrt{n}} - \frac{\hat{A}x^2}{3\sqrt{n}} + \frac{\hat{A}^2x}{9n} + \frac{5\hat{A}^2x^3}{27n}\right) + O(n^{-3/2}) \\
= \; & P\left(T_n + \frac{(2x^2+1)(\hat{A} - A)}{6\sqrt{n}} < x - \frac{(2x^2+1)A}{6\sqrt{n}} + \frac{(5x^3+3x)A^2}{27n}\right) + O(n^{-3/2}),
\end{aligned}$$

where the second equal sign is due to Taylor expansion and the fourth equal sign is due to the Delta method of Edgeworth expansion. Then the power of the lower-tailed TT_1 test is

$$\begin{aligned}
& P\left(T_n + \frac{(2z_\alpha^2+1)(\hat{A} - A)}{6\sqrt{n}} < z_\alpha - \frac{(2z_\alpha^2+1)A}{6\sqrt{n}} + \frac{(5z_\alpha^3+3z_\alpha)A^2}{27n}\right) + O(n^{-3/2}) \\
= \; & P\left(R_{1n} < z_\alpha - \frac{(2z_\alpha^2+1)A}{6\sqrt{n}} + \frac{(5z_\alpha^3+3z_\alpha)A^2}{27n}\right) + O(n^{-3/2}) \\
= \; & F_{T,w}(\eta_{1,\alpha}) - n^{-1}c_\alpha(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2})
\end{aligned}$$

where $R_n = T_n + (2z_\alpha^2+1)(\hat{A} - A)/(6\sqrt{n})$ was defined in (A.2) in A.3.1,

$$\eta_{1,\alpha} = z_\alpha - (2z_\alpha^2+1)A/(6\sqrt{n}) + (5z_\alpha^3+3z_\alpha)A^2/(27n). \tag{A.3}$$

The last equality follows from Theorem 1 and $c_\alpha = (B/6 - A^2/4)(2z_\alpha^2 + 1)$.

The inverse function of $G_2(U) = (2/3n^{-1/2}\hat{A})^{-1}[\exp(2/3n^{-1/2}\hat{A}U) - 1] + \hat{A}/(6n)$ is $G_2^{-1}(t) = (2\hat{A}/(3\sqrt{n}))^{-1}\log[1 + 2\hat{A}/(3\sqrt{n})(t - \hat{A}/(6n))]$. Then the CDF of the test statistic of TT_2 test is

$$\begin{aligned}
& P\left(\sqrt{n}G_2\left(\frac{T_n}{\sqrt{n}}\right) < x\right) \\
= \;& P\left(T_n < \sqrt{n}\left(\frac{2\hat{A}}{3\sqrt{n}}\right)^{-1}\log\left[1 + \frac{2\hat{A}}{3\sqrt{n}}\left(\frac{x}{\sqrt{n}} - \frac{\hat{A}}{6n}\right)\right]\right) \\
= \;& P\left(T_n < n\left(\frac{2\hat{A}}{3}\right)^{-1}\left[\frac{2\hat{A}}{3\sqrt{n}}\left(\frac{x}{\sqrt{n}} - \frac{\hat{A}}{6n}\right) - \frac{1}{2}\left(\frac{2\hat{A}}{3\sqrt{n}}\left(\frac{x}{\sqrt{n}} - \frac{\hat{A}}{6n}\right)\right)^2 + O_p(n^{-3})\right]\right) \\
= \;& P\left(T_n < x - \frac{\hat{A}}{6\sqrt{n}} - \frac{\hat{A}x^2}{3n}\right) + O(n^{-3/2}) \\
= \;& P\left(T_n + \frac{\hat{A} - A}{6\sqrt{n}} < x - \frac{A}{6\sqrt{n}} - \frac{Ax^2}{3n}\right) + O(n^{-3/2}),
\end{aligned}$$

where the second equal sign is due to Taylor expansion and the third is due to the Delta method of Edgeworth expansion. Then the power of the lower-tailed TT_2 test is

$$\begin{aligned}
& P\left(T_n + \frac{\hat{A} - A}{6\sqrt{n}} < z_\alpha - \frac{A}{6\sqrt{n}} - \frac{Az_\alpha^2}{3n}\right) + O(n^{-3/2}) \\
= \;& P\left(R_{2n} < z_\alpha - \frac{A}{6\sqrt{n}} - \frac{Az_\alpha^2}{3n}\right) + O(n^{-3/2}) \\
= \;& F_{T,w}(\eta_{2,\alpha}) - n^{-1}c_{2,\alpha}(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2})
\end{aligned}$$

where $R_{2n} = T_n + (\hat{A} - A)/(6\sqrt{n})$, $\eta_{2,\alpha} = z_\alpha - A/(6\sqrt{n}) - Az_\alpha^2/3n$. Using the argument analogous to A.3.1, we can show $c_{2,\alpha} = B/6 - A^2/4$.

The inverse function of $G_3(U) = U + U^2 + U^3/3 + \hat{A}/(6n)$ is $G_3^{-1}(t) = [1 + 3(t - \hat{A}/(6n))]^{1/3} -$

1. Then the CDF of the test statistic of TT_3 test is

$$\begin{aligned}
& P\left(\sqrt{n}G_3\left(\frac{T_n}{\sqrt{n}}\right) < x\right) \\
= \;& P\left(T_n < \sqrt{n}[1+3(x/\sqrt{n}-\hat{A}/(6n))]^{1/3}-1\right) \\
= \;& P\left(T_n < \sqrt{n}[(x/\sqrt{n}-\hat{A}/(6n)) - (x/\sqrt{n}-\hat{A}/(6n))^2 + 5(x/\sqrt{n}-\hat{A}/(6n))^3/3 + O_p(n^{-2})]\right) \\
= \;& P\left(T_n + \frac{\hat{A}-A}{6\sqrt{n}} < x - \frac{A+6x^2}{6\sqrt{n}} + \frac{5x^3+Ax}{3n}\right) + O(n^{-3/2}),
\end{aligned}$$

where the second equal sign is due to Taylor expansion and the third is due to the Delta method of Edgeworth expansion. Then the power of the lower-tailed TT_3 test is

$$\begin{aligned}
& P\left(T_n + \frac{\hat{A}-A}{6\sqrt{n}} < z_\alpha - \frac{A+6z_\alpha^2}{6\sqrt{n}} + \frac{5z_\alpha^3+Az_\alpha}{3n}\right) + O(n^{-3/2}) \\
= \;& P\left(R_{2n} < z_\alpha - \frac{A+6z_\alpha^2}{6\sqrt{n}} + \frac{5z_\alpha^3+Az_\alpha}{3n}\right) + O(n^{-3/2}) \\
= \;& F_{T,w}(\eta_{3,\alpha}) - n^{-1}c_{3,\alpha}(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2})
\end{aligned}$$

where $\eta_{3,\alpha} = z_\alpha - (A+6z_\alpha^2)/(6\sqrt{n}) + (5z_\alpha^3 + Az_\alpha)/(3n)$, $c_{3,\alpha} = c_{2,\alpha} = B/6 - A^2/4$.

The inverse function of $G_4(U) = (2/3\hat{A})^{-1}[\exp(2/3\hat{A}U) - 1] + \hat{A}/(6n)$ is $G_4^{-1}(t) =$

$(2\hat{A}/3)^{-1}\log[1+2\hat{A}/3(t-\hat{A}/(6n))]$. Then the CDF of the test statistic of TT$_4$ test is

$$\begin{aligned}
& P\left(\sqrt{n}G_4\left(\frac{T_n}{\sqrt{n}}\right)<x\right)\\
= \; & P\left(T_n<\sqrt{n}\left(\frac{2\hat{A}}{3}\right)^{-1}\log\left[1+\frac{2\hat{A}}{3}\left(\frac{x}{\sqrt{n}}-\frac{\hat{A}}{6n}\right)\right]\right)\\
= \; & P\left(T_n<\sqrt{n}\left(\frac{2\hat{A}}{3}\right)^{-1}\left[\frac{2\hat{A}}{3}\left(\frac{x}{\sqrt{n}}-\frac{\hat{A}}{6n}\right)-\frac{1}{2}\left(\frac{2\hat{A}}{3}\left(\frac{x}{\sqrt{n}}-\frac{\hat{A}}{6n}\right)\right)^2\right.\right.\\
& \left.\left.+\frac{1}{3}\left(\frac{2\hat{A}}{3}\left(\frac{x}{\sqrt{n}}-\frac{\hat{A}}{6n}\right)\right)^3+O_p(n^{-2})\right]\right)\\
= \; & P\left(T_n<x-\frac{\hat{A}}{6\sqrt{n}}-\frac{\hat{A}x^2}{3\sqrt{n}}+\frac{\hat{A}^2x}{9n}+\frac{4\hat{A}^2x^3}{27n}\right)+O(n^{-3/2})\\
= \; & P\left(T_n+\frac{(2x^2+1)(\hat{A}-A)}{6\sqrt{n}}<x-\frac{(2x^2+1)A}{6\sqrt{n}}+\frac{3A^2x+4A^2x^3}{27n}\right)+O(n^{-3/2}),
\end{aligned}$$

where the second equal sign is due to Taylor expansion and the third is due to the Delta method of Edgeworth expansion. Then the power of the lower-tailed TT$_4$ test is

$$\begin{aligned}
& P\left(T_n+\frac{(2z_\alpha^2+1)(\hat{A}-A)}{6\sqrt{n}}<z_\alpha-\frac{(2z_\alpha^2+1)A}{6\sqrt{n}}+\frac{3A^2z_\alpha+4A^2z_\alpha^3}{27n}\right)+O(n^{-3/2})\\
= \; & P\left(R_n<z_\alpha-\frac{(2z_\alpha^2+1)A}{6\sqrt{n}}+\frac{A^2(4z_\alpha^3+3z_\alpha)}{27n}\right)+O(n^{-3/2})\\
= \; & F_{T,w}(\eta_{4,\alpha})-n^{-1}c_\alpha(z_\alpha-w)\phi(z_\alpha-w)+O(n^{-3/2})
\end{aligned}$$

where R_n was defined in (A.2) in Appendix A.3.1, $\eta_{4,\alpha}=z_\alpha-(2z_\alpha^2+1)A/(6\sqrt{n})+(4z_\alpha^3+3z_\alpha)A^2/(27n)$. □

A.3.4 Proof for Corollary 2

Proof. Due to different $\eta_{i,\alpha}$ being used in the power functions, the comparison cannot be directly conducted. We apply Taylor expansion to the power functions at $\xi_{2,\alpha}$ defined in equation (2.7) to unify the leading terms.

To facilitate the derivation, the first derivative of $F_{T,w}(x)$ is

$$f_{T,w}(x) = \phi(x-w) + n^{-1/2}[p_1'(x-w)\phi(x-w) + p_1(x-w)\phi'(x-w)] + O(n^{-1}),$$

and second derivative is

$$f_{T,w}'(x) = \phi'(x-w) + O(n^{-1/2}).$$

Because they only appear at higher-order terms, we suppressed the small terms.

The power function of the lower-tailed TT_1 test can be rewritten as

$$\begin{aligned}
& P(\sqrt{n}G_1(T/\sqrt{n}) \leq z_\alpha) \\
= \; & F_{T,w}(\eta_{1,\alpha}) - n^{-1}c_{1,\alpha}(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2}) \\
= \; & F_{T,w}(\xi_{2,\alpha}) + f_{T,w}(\xi_{2,\alpha})(\eta_{1,\alpha} - \xi_{2,\alpha}) + \frac{1}{2}f_{T,w}'(\xi_{2,\alpha})(\eta_{1,\alpha} - \xi_{2,\alpha})^2 \\
& -n^{-1}c_{1,\alpha}(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2}) \\
= \; & F_{T,w}(\xi_{2,\alpha}) + f_{T,w}(\xi_{2,\alpha})(\eta_{1,\alpha} - \xi_{2,\alpha}) - n^{-1}c_{1,\alpha}(z_\alpha - w)\phi(z_\alpha - w) + O(n^{-3/2}) \\
= \; & F_{T,w}(\xi_{2,\alpha}) + n^{-1}Q_{1,w,\alpha} + O(n^{-3/2}),
\end{aligned}$$

where $\eta_{1,\alpha}$ was defined in (A.3) and $Q_{1,w,\alpha} = \phi(z_\alpha - w)[A^2(5z_\alpha^3 + 3z_\alpha)/27 - q_{20}(z_\alpha)] - c_{1,\alpha}(z_\alpha - w)\phi(z_\alpha - w)$. The third equality is due to the fact that $\eta_{1,\alpha} - \xi_{2,\alpha} = O(n^{-1})$. Similar technique can show

$$\begin{aligned}
P(\sqrt{n}G_2(T/\sqrt{n}) \leq z_\alpha) =& F_{T,w}(\xi_{2,\alpha}) + n^{-1/2}f_{T,w}(\xi_{2,\alpha})z_\alpha^2 A/3 + n^{-1}Q_{2,w,\alpha} + O(n^{-3/2}), \\
P(\sqrt{n}G_3(T/\sqrt{n}) \leq z_\alpha) =& F_{T,w}(\xi_{2,\alpha}) + n^{-1/2}f_{T,w}(\xi_{2,\alpha})z_\alpha^2(A-3)/3 + n^{-1}Q_{3,w,\alpha} + O(n^{-3/2}), \\
P(\sqrt{n}G_4(T/\sqrt{n}) \leq z_\alpha) =& F_{T,w}(\xi_{2,\alpha}) + n^{-1}Q_{4,w,\alpha} + O(n^{-3/2}),
\end{aligned}$$

where

$$
\begin{aligned}
Q_{2,w,\alpha} &= \phi(z_\alpha - w)[-Az_\alpha^2/3 - q_{20}(z_\alpha)] + A^2 z_\alpha^4 \phi'(z_\alpha - w)/18 - c_{2,\alpha}(z_\alpha - w)\phi(z_\alpha - w) \\
Q_{3,w,\alpha} &= \phi(z_\alpha - w)[(5z_\alpha^3 + Az_\alpha)/3 - q_{20}(z_\alpha)] + (A-3)^2 z_\alpha^4 \phi'(z_\alpha - w)/18 - c_{3,\alpha}(z_\alpha - w)\phi(z_\alpha - w) \\
Q_{4,w,\alpha} &= \phi(z_\alpha - w)[A^2(4z_\alpha^3 + 3z_\alpha)/27 - q_{20}(z_\alpha)] - c_{4,\alpha}(z_\alpha - w)\phi(z_\alpha - w).
\end{aligned}
$$

Now the power functions all have $F_{T,w}(\xi_{2,\alpha})$ as the leading term, and the power differences can be computed by subtraction. □

A.3.5 Proof for Corollary 3

Proof. For $i = 1$ or 4, using the results in A.3.4, we have

$$
\begin{aligned}
& P_{H_0}(\sqrt{n}G_i(T_n/\sqrt{n}) < z_{\alpha/2}) + P_{H_0}(\sqrt{n}G_i(T_n/\sqrt{n}) > z_{1-\alpha/2}) \\
= \; & F_{T,0}(\xi_{2,\alpha/2}) + n^{-1}Q_{i,0,\alpha/2} + 1 - [F_{T,0}(\xi_{2,1-\alpha/2}) + n^{-1}Q_{i,0,1-\alpha/2}] + O(n^{-3/2}) \\
= \; & \alpha/2 + n^{-1}Q_{i,0,\alpha/2} + \alpha/2 - n^{-1}Q_{i,0,1-\alpha/2} + O(n^{-3/2}) \\
= \; & \alpha + 2Q_{i,0,\alpha/2} + O(n^{-3/2}),
\end{aligned}
$$

where the second equality holds because $Q_{i,0,\alpha} = -Q_{i,0,1-\alpha}$. We can show $2Q_{i,0,\alpha/2} = \Lambda_i$ for $i = 1$ or 4 in the corollary using some basic algebra.

For the $i = 2$,

$$
\begin{aligned}
& P_{H_0}(\sqrt{n}G_2(T_n/\sqrt{n}) < z_{\alpha/2}) + P_{H_0}(\sqrt{n}G_2(T_n/\sqrt{n}) > z_{1-\alpha/2}) \\
= \; & F_{T,0}(\xi_{2,\alpha/2}) + n^{-1/2}f_{T,0}(\xi_{2,\alpha/2})z_{\alpha/2}^2 A/3 + n^{-1}Q_{2,0,\alpha/2} \\
& +1 - [F_{T,0}(\xi_{2,1-\alpha/2}) + n^{-1/2}f_{T,0}(\xi_{2,1-\alpha/2})z_{1-\alpha/2}^2 A/3 + n^{-1}Q_{2,0,1-\alpha/2}] + O(n^{-3/2}) \\
= \; & \alpha + n^{-1/2}z_{\alpha/2}^2 A/3[f_{T,0}(\xi_{2,\alpha/2}) - f_{T,0}(\xi_{2,1-\alpha/2})] + n^{-1}[Q_{2,0,\alpha/2} - Q_{2,0,1-\alpha/2}] + O(n^{-3/2}),
\end{aligned}
$$

where $f_{T,0}(x)$ is the first derivative of $F_{T,0}(x)$, and it can be expanded as

$$f_{T,0}(x) = \phi(x) + n^{-1/2}[p'_{10}(x)\phi(x) + p_{10}(x)\phi'(x)] + O(n^{-1}). \qquad \text{(A.4)}$$

The second derivative of $F_{T,0}(x)$ is

$$f'_{T,0}(x) = \phi'(x) + O(n^{-1/2}). \qquad \text{(A.5)}$$

Because they only appear at high-order terms, we suppressed the small terms. Then by (A.4), we have

$$\begin{aligned} n^{-1/2}f_{T,0}(\xi_{2,\alpha/2}) &= n^{-1/2}\phi(\xi_{2,\alpha/2}) + n^{-1}[p'_{10}(z_{\alpha/2})\phi(z_{\alpha/2}) + p_{10}(z_{\alpha/2})\phi'(z_{\alpha/2})] + O(n^{-3/2}) \\ &= n^{-1/2}[\phi(z_{\alpha/2}) + \phi'(z_{\alpha/2})(\xi_{2,\alpha/2} - z_{\alpha/2})] \\ &\quad + n^{-1}[p'_{10}(z_{\alpha/2})\phi(z_{\alpha/2}) + p_{10}(z_{\alpha/2})\phi'(z_{\alpha/2})] + O(n^{-3/2}) \\ &= n^{-1/2}\phi(z_{\alpha/2}) + n^{-1}\phi'(z_{\alpha/2})q_{10}(z_{\alpha/2}) \\ &\quad + n^{-1}[p'_{10}(z_{\alpha/2})\phi(z_{\alpha/2}) + p_{10}(z_{\alpha/2})\phi'(z_{\alpha/2})] + O(n^{-3/2}) \\ &= n^{-1/2}\phi(z_{\alpha/2}) + n^{-1}p'_{10}(z_{\alpha/2})\phi(z_{\alpha/2}) + O(n^{-3/2}), \end{aligned}$$

and

$$n^{-1/2}f_{T,0}(\xi_{2,1-\alpha/2}) = n^{-1/2}\phi(z_{1-\alpha/2}) + n^{-1}p'_{10}(z_{1-\alpha/2})\phi(z_{1-\alpha/2}) + O(n^{-3/2}).$$

From A.1, we know that $p_{10}(x) = (2x^2+1)A/6$, and $p'_{10}(x) = 2Ax/3$. Also note the fact that $\phi'(x) = -x\phi(x)$. Therefore

$$\begin{aligned} & P_{H_0}(\sqrt{n}G_2(T_n/\sqrt{n}) < z_{\alpha/2}) + P_{H_0}(\sqrt{n}G_2(T_n/\sqrt{n}) > z_{1-\alpha/2}) \\ = \;& \alpha + n^{-1}\left[2z_{\alpha/2}^2 A/3p'_{10}(z_{\alpha/2})\phi(z_{\alpha/2}) + Q_{2,0,\alpha/2} - Q_{2,0,1-\alpha/2}\right] + O(n^{-3/2}) \\ = \;& \alpha + n^{-1}[4z_{\alpha/2}^3 A^2/9\phi(z_{\alpha/2}) + Q_{2,0,\alpha/2} - Q_{2,0,1-\alpha/2}] + O(n^{-3/2}) \\ = \;& \alpha + n^{-1}\Lambda_2 + O(n^{-3/2}). \end{aligned}$$

Analogously, because $Q_{3,0,\alpha} = -Q_{3,0,1-\alpha}$, we have

$$\begin{aligned} & P_{H_0}(\sqrt{n}G_3(T_n/\sqrt{n}) < z_{\alpha/2}) + P_{H_0}(\sqrt{n}G_3(T_n/\sqrt{n}) > z_{1-\alpha/2}) \\ = \;& \alpha + n^{-1}[2z_{\alpha/2}^2(A-3)/3p_{10}'(z_{\alpha/2})\phi(z_{\alpha/2}) + 2Q_{3,0,\alpha/2}] + O(n^{-3/2}) \\ = \;& \alpha + n^{-1}[4z_{\alpha/2}^3 A(A-3)/9\phi(z_{\alpha/2}) + 2Q_{3,0,\alpha/2}] + O(n^{-3/2}) \\ = \;& \alpha{+}n^{-1}\Lambda_3 + O(n^{-3/2}). \end{aligned}$$

□

Appendix B

Appendices for Chapter 3

B.1 Proofs

B.1.1 Proof of Lemma 1

Proof. Recall that $t_j^2 = (\overline{X}_j - \overline{Y}_j)^2/(S_{1j}^2/n + S_{2j}^2/m)$. Because the samples are independently identically distributed, basic central limit theorem can prove that the sample variances all have asymptotic normal distributions centered at their corresponding population variances. We can hence write $S_{1j}^2 = \sigma_{1j}^2 + O_p(N^{-1/2})$, $S_{2j}^2 = \sigma_{2j}^2 + O_p(N^{-1/2})$. Then rewrite t_j^2 as

$$\begin{aligned}
t_j^2 &= \frac{(\overline{X}_j - \overline{Y}_j)^2}{\dfrac{\sigma_{1j}^2}{n} + \dfrac{\sigma_{2j}^2}{m}} \frac{\dfrac{\sigma_{1j}^2}{n} + \dfrac{\sigma_{2j}^2}{m}}{\dfrac{S_{1j}^2}{n} + \dfrac{S_{2j}^2}{m}} \\
&= \frac{N(\overline{X}_j - \overline{Y}_j)^2}{\dfrac{\sigma_{1j}^2}{\lambda_1} + \dfrac{\sigma_{2j}^2}{\lambda_2}} \frac{\dfrac{\sigma_{1j}^2}{\lambda_1} + \dfrac{\sigma_{2j}^2}{\lambda_2}}{\dfrac{\sigma_{1j}^2}{\lambda_1} + \dfrac{\sigma_{2j}^2}{\lambda_2} + O_p(N^{-1/2})} \\
&= \frac{N(\overline{X}_j - \overline{Y}_j)^2}{\dfrac{\sigma_{1j}^2}{\lambda_1} + \dfrac{\sigma_{2j}^2}{\lambda_2}} [1 + O_p(N^{-1/2})]^{-1} \\
&= \frac{N(\overline{X}_j - \overline{Y}_j)^2}{\dfrac{\sigma_{1j}^2}{\lambda_1} + \dfrac{\sigma_{2j}^2}{\lambda_2}} + O_p(N^{-1/2}),
\end{aligned}$$

where the last equality follows from the Taylor series $(1+x)^{-1} = 1 - x + o(x)$ at $x = 0$. Next, consider that

$$\begin{aligned}
&\mathrm{cov}\left[(\overline{X}_j - \overline{Y}_j)^2, (\overline{X}_{j'} - \overline{Y}_{j'})^2\right] \\
=&\mathrm{cov}\left(\overline{X}_j^2 + \overline{Y}_j^2 - 2\overline{X}_j\overline{Y}_j,\ \overline{X}_{j'}^2 + \overline{Y}_{j'}^2 - 2\overline{X}_{j'}\overline{Y}_{j'}\right) \\
=&\mathrm{cov}\left(\overline{X}_j^2, \overline{X}_{j'}^2\right) + \mathrm{cov}\left(\overline{X}_j^2, \overline{Y}_{j'}^2\right) - 2\mathrm{cov}\left(\overline{X}_j^2, \overline{X}_{j'}\overline{Y}_{j'}\right) + \mathrm{cov}\left(\overline{Y}_j^2, \overline{X}_{j'}^2\right) + \mathrm{cov}\left(\overline{Y}_j^2, \overline{Y}_{j'}^2\right) \\
&- 2\mathrm{cov}\left(\overline{Y}_j^2, \overline{X}_{j'}\overline{Y}_{j'}\right) - 2\mathrm{cov}\left(\overline{X}_j\overline{Y}_j, \overline{X}_{j'}^2\right) - 2\mathrm{cov}\left(\overline{X}_j\overline{Y}_j, \overline{Y}_{j'}^2\right) \\
&+ 4\mathrm{cov}\left(\overline{X}_j\overline{Y}_j, \overline{X}_{j'}\overline{Y}_{j'}\right)
\end{aligned} \tag{B.1}$$

Noting the independence between $\mathbf{X}_j$ and $\mathbf{Y}_j$, $\mathrm{cov}\left(\overline{X}_j^2, \overline{Y}_{j'}^2\right) = \mathrm{cov}\left(\overline{Y}_j^2, \overline{X}_{j'}^2\right) = 0$. Under the null hypothesis, $E(X_{1j}) = E(Y_{1j})$, so without loss of generality we can assume that $E(X_{1j}) = E(Y_{1j}) = 0$ for $j = 1, \ldots, p$. Then

$$\begin{aligned}
&\mathrm{cov}\left(\overline{X}_j^2, \overline{X}_{j'}\overline{Y}_{j'}\right) \\
=&E\left(\overline{X}_j^2\overline{X}_{j'}\overline{Y}_{j'}\right) - E\left(\overline{X}_j^2\right)E\left(\overline{X}_{j'}\overline{Y}_{j'}\right) \\
=&E\left(\overline{X}_j^2\overline{X}_{j'}\right)E(\overline{Y}_{j'}) - E\left(\overline{X}_j^2\right)E\left(\overline{X}_{j'}\right)E(\overline{Y}_{j'}) \\
=&0
\end{aligned}$$

Similarly, $\mathrm{cov}\left(\overline{Y}_j^2, \overline{X}_{j'}\overline{Y}_{j'}\right) = \mathrm{cov}\left(\overline{X}_j\overline{Y}_j, \overline{X}_{j'}^2\right) = \mathrm{cov}\left(\overline{X}_j\overline{Y}_j, \overline{Y}_{j'}^2\right) = 0$. Also note that,

$$\begin{aligned}
&\mathrm{cov}\left(\overline{X}_j^2, \overline{X}_{j'}^2\right) \\
=&n^{-4}\mathrm{cov}\left(\sum_{i_1=1}^n\sum_{i_2=1}^n X_{ji_1}X_{ji_2}, \sum_{i_3=1}^n\sum_{i_4=1}^n X_{j'i_3}X_{j'i_4}\right) \\
=&n^{-4}E\left(\sum_{i_1=1}^n\sum_{i_2=1}^n\sum_{i_3=1}^n\sum_{i_4=1}^n X_{ji_1}X_{ji_2}X_{j'i_3}X_{j'i_4}\right) - \\
&n^{-4}E\left(\sum_{i_1=1}^n\sum_{i_2=1}^n X_{ji_1}X_{ji_2}\right)E\left(\sum_{i_3=1}^n\sum_{i_4=1}^n X_{j'i_3}X_{j'i_4}\right)
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&n^{-4}E\left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{i_3=1}^{n}\sum_{i_4=1}^{n}X_{ji_1}X_{ji_2}X_{j'i_3}X_{j'i_4}\right)\\
=&n^{-4}E\left(\sum_{i_1=i_2}\sum_{i_3=i_4\neq i_1}X_{ji_1}X_{ji_2}X_{j'i_3}X_{j'i_4}\right)+n^{-4}E\left(\sum_{i_1=i_3}\sum_{i_2=i_4\neq i_1}X_{ji_1}X_{ji_2}X_{j'i_3}X_{j'i_4}\right)\\
&+n^{-4}E\left(\sum_{i_1=i_4}\sum_{i_2=i_3\neq i_1}X_{ji_1}X_{ji_2}X_{j'i_3}X_{j'i_4}\right)+n^{-4}E\left(\sum_{i_1=i_2=i_3=i_4}X_{ji_1}X_{ji_2}X_{j'i_3}X_{j'i_4}\right)\\
=&n^{-4}\sum_{i_1}EX_{ji_1}^2\sum_{i_3\neq i_1}EX_{j'i_3}^2+2n^{-4}\sum_{i_1}E(X_{ji_1}X_{j'i_1})\sum_{i_2\neq i_1}E(X_{ji_2}X_{j'i_2})+O(n^{-3})\\
=&n^{-4}n(n-1)\sigma_{1j}^2\sigma_{1j'}^2+2n^{-4}n(n-1)\sigma_{1jj'}^2+O(n^{-3})\\
=&n^{-2}\sigma_{1j}^2\sigma_{1j'}^2+2n^{-2}\sigma_{1jj'}^2+O(n^{-3}).
\end{aligned}$$

The second term is

$$n^{-4}E\left(\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}X_{ji_1}X_{ji_2}\right)E\left(\sum_{i_3=1}^{n}\sum_{i_4=1}^{n}X_{j'i_3}X_{j'i_4}\right)=n^{-4}(n\sigma_{1j}^2)(n\sigma_{1j'}^2)=n^{-2}\sigma_{1j}^2\sigma_{1j'}^2.$$

Then $\operatorname{cov}\left(\overline{X}_j^2,\overline{X}_{j'}^2\right)=2n^{-2}\sigma_{1jj'}^2+O(n^{-3})$. In a similar way, we can show $\operatorname{cov}\left(\overline{Y}_j^2,\overline{Y}_{j'}^2\right)=2n^{-2}\sigma_{2jj'}^2+O(m^{-3})$, $\operatorname{cov}\left(\overline{X}_j\overline{Y}_j,\overline{X}_{j'}\overline{Y}_{j'}\right)=(mn)^{-1}\sigma_{1jj'}\sigma_{2jj'}$. Collect all the terms in equation (B.1), we have $\operatorname{cov}\left[(\overline{X}_j-\overline{Y}_j)^2,(\overline{X}_{j'}-\overline{Y}_{j'})^2\right]=2n^{-2}\sigma_{1jj'}^2+2m^{-2}\sigma_{2jj'}^2+4(mn)^{-1}\sigma_{1jj'}\sigma_{2jj'}=2(n^{-1}\sigma_{1jj'}+m^{-1}\sigma_{2jj'})^2$. Using Lemma 7 and the conditions $\sup_j E(X_{1j}^4)<\infty$, $\sup_j E(Y_{1j}^4)<\infty$, we have

$$\operatorname{cov}(t_j^2,t_{j'}^2)=\frac{2N^2(n^{-1}\sigma_{1jj'}+m^{-1}\sigma_{2jj'})^2}{(\sigma_{1j}^2/\lambda_1+\sigma_{2j}^2/\lambda_2)(\sigma_{1j'}^2/\lambda_1+\sigma_{2j'}^2/\lambda_2)}+O(N^{-1/2}).$$

□

B.1.2 Proof for Lemma 3

Proof. Note that $t_j^2 = g(\mathbf{Z}_j)$ for a Borel-measurable function $g : \mathcal{R}^{m+n} \to \mathcal{R}$. For an arbitrary Borel set B, $\{t_j^2 \in B\} = \{\omega : g(\mathbf{Z}_j(\omega)) \in B\} = \{\omega : \mathbf{Z}_j(\omega) \in B'\}$ for Borel set $B' \in \mathcal{B}^{m+n}$, where $\mathcal{B}^{m+n}$ is the Borel σ-algebra of subsets of $\mathcal{R}^{m+n}$. Hence, the σ-algebra generated by t_j^2 is a sub-σ-algebra of the σ-algebra generated by $\mathbf{Z}_j$. Then the result follows from the definition of strong mixing coefficient in (3.11). □

B.1.3 Proof of Lemma 5

Proof. First note that

$$\text{var}(\sqrt{p}T_n) = p^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}\text{cov}(t_j^2, t_{j'}^2) = p^{-1}\sum_{j=1}^{p}\text{var}(t_j^2) + 2p^{-1}\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}\text{cov}(t_j^2, t_{j'}^2).$$

We have $\lim_{p\to\infty} p^{-1}\sum_{j=1}^{p}\text{var}(t_j^2) < \infty$ because $\sup_j \text{var}(t_j^2) \le \sup_j E(t_j^4) < \infty$. Applying inequality (3.12), we have

$$\begin{aligned} p^{-1}\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p}\text{cov}(t_j^2, t_{j'}^2) &\le p^{-1}\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} 12[\alpha(j-j')]^{\nu/(2+\nu)}|E(t_j^2)^{2+\nu}|^{1/(2+\nu)}|E(t_{j'}^2)^{2+\nu}|^{1/(2+\nu)} \\ &= 12\sup_j |E(t_j^2)^{2+\nu}|^{2/(2+\nu)}p^{-1}\sum_{j=1}^{p-1}\sum_{r=1}^{p-j}[\alpha(r)]^{\nu/(2+\nu)} \\ &\le 12\sup_j |E(t_j^2)^{2+\nu}|^{2/(2+\nu)}\sum_{r=1}^{p}[\alpha(r)]^{\nu/(2+\nu)} \\ &< \infty, \text{ as } p \to \infty. \end{aligned}$$

Recall from Lemma 1 that $\lim_{N\to\infty}\text{cov}(t_j^2, t_{j'}^2) = \gamma_{j,j'} \ge 0$. Because $\min\{\sigma_{1j}^2, \sigma_{2j}^2\} > 0$, $\gamma_{j,j} > 0$ from the definition in (3.5). Then

$\lim_{N,p\to\infty}\text{var}(\sqrt{p}T_n) = \lim_{p\to\infty} p^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}\gamma_{j,j'} \ge \lim_{p\to\infty} p^{-1}\sum_{j=1}^{p}\gamma_{j,j} > 0.$ □

B.1.4 Proof of Theorem 3

Proof. Through Lemma 3 and Lemma 5, we have established that the sequence of $\{t_1^2, t_2^2, \ldots\}$ is α-mixing, and that $0 < \lim_{p\to\infty} \operatorname{var}(\sqrt{p}T_n) < \infty$. To show the asymptotic normality of T_n, we can follow the proof of Lemma 2.1 in Wang and Akritas (2010) to finish the rest of our proof.

Their lemma requires $\limsup_j E(t_j^{32}) < \infty$, but we will show a way that can relax the moment conditions. Denote $z_j = t_j^2 - E(t_j^2)$ for $j = 1, 2, \ldots$. In their proof, the only place that requires $\limsup_j E(t_j^{32}) < \infty$ is the inequality, using (3.13),

$$|E(z_{j_1}z_{j_2}z_{j_3}z_{j_4})| \leq 8[1 + E(z_{j_1}^4) + E(z_{j_2}^4 z_{j_3}^4 z_{j_4}^4)][\alpha(j_2 - j_1)]^{1/2},\ j_1 < j_2 < j_3 < j_4.$$

To show the finiteness of $E(z_{j_2}^4 z_{j_3}^4 z_{j_4}^4)$, they invoked the Cauchy-Schwarz inequality twice that demands the finiteness of 16^{th} moment of z_j, hence the 32^{nd} moment of t_j.

Alternatively, using the Davydov inequality (3.12) setting $s = q = 2 + \nu$ with $\nu > 0$, we have

$$|E(z_{j_1}z_{j_2}z_{j_3}z_{j_4})| \leq 12[\alpha(j_2 - j_1)]^{\nu/(2+\nu)} |Ez_{j_1}^{2+\nu}|^{1/(2+\nu)} |Ez_{j_2}^{2+\nu} z_{j_3}^{2+\nu} z_{j_4}^{2+\nu}|^{1/(2+\nu)}.$$

The assumption $\limsup_j E(t_{1j}^{12+6\nu}) < \infty$ is equivalent to $\limsup_j E(z_j^{6+3\nu}) < \infty$, so

$$|Ez_{j_2}^{2+\nu} z_{j_3}^{2+\nu} z_{j_4}^{2+\nu}| \leq \left[E|z_{j_2}^{3(2+\nu)}|\right]^{1/3} \left[E|z_{j_3}^{3(2+\nu)}|\right]^{1/3} \left[E|z_{j_4}^{3(2+\nu)}|\right]^{1/3} < \infty$$

from Hölder's inequality. It follows that $|E(z_{j_1}z_{j_2}z_{j_3}z_{j_4})| \leq K_1[\alpha(j_2 - j_1)]^{\nu/(2+\nu)}$ for some constant K_1. Similarly, $|E(z_{j_1}z_{j_2}z_{j_3}z_{j_4})| \leq K_2[\alpha(j_4 - j_3)]^{\nu/(2+\nu)}$ for some constant K_2. This shows that $|E(z_{j_1}z_{j_2}z_{j_3}z_{j_4})| \leq K \min\{[\alpha(j_2 - j_1)]^{\nu/(2+\nu)}, [\alpha(j_4 - j_3)]^{\nu/(2+\nu)}\}$ for some constant

K. Then

$$\begin{aligned}
\sum_{j_1<j_2<j_3<j_4}^{p} |E(z_{j_1}z_{j_2}z_{j_3}z_{j_4})| &\leq \sum_{j_1<j_2<j_3<j_4}^{p} K\min\{[\alpha(j_2-j_1)]^{\nu/(2+\nu)}, [\alpha(j_4-j_3)]^{\nu/(2+\nu)}\} \\
&= K\sum_{j_1=1}^{p-3}\sum_{j_2=j_1+1}^{p-2}\sum_{j_3=j_2+1}^{p-1}\sum_{r=1}^{p-j_3}\min\{[\alpha(j_2-j_1)]^{\nu/(2+\nu)}, [\alpha(r)]^{\nu/(2+\nu)}\} \\
&= K\sum_{j_1=1}^{p-3}\sum_{s=1}^{p-2-j_1}\sum_{j_3=j_1+s+1}^{p-1}\sum_{r=1}^{p-j_3}\min\{[\alpha(s)]^{\nu/(2+\nu)}, [\alpha(r)]^{\nu/(2+\nu)}\} \\
&\leq Kp^2\sum_{s=1}^{p}\sum_{r=1}^{p}\min\{[\alpha(s)]^{\nu/(2+\nu)}, [\alpha(r)]^{\nu/(2+\nu)}\} \\
&\leq Kp^2\left[\sum_{s=1}^{p}[\alpha(s)]^{\nu/(2+\nu)} + 2\sum_{s=1}^{p-1}\sum_{r=s+1}^{p}[\alpha(r)]^{\nu/(2+\nu)}\right] \\
&\leq 2Kp^2\sum_{r=1}^{p} r[\alpha(r)]^{\nu/(2+\nu)} \\
&= \tilde{K}p^2
\end{aligned}$$

for some $\tilde{K} < \infty$ as $p \to \infty$. Note that, the condition $\alpha(r) = O(r^{-5})$, as stated in Lemma 4, is used to control the summation $\sum_{r=1}^{p} r[\alpha(r)]^{1/2}$ in their proof. In our theorem, we avoid this condition by assuming (C.1): $\sum_{r=1}^{p} r[\alpha(r)]^{\nu/(2+\nu)} < \infty$. The rest of the proof is analogous to the proof of Lemma 2.1 in Wang and Akritas (2010). □

B.1.5 Proof for the asymptotic normality of $\hat{\gamma}(k)$

Proof. To facilitate the derivation, we assume $k \geq 0$ in this proof. The $k < 0$ situation can be easily shown by using the symmetry property $\gamma(k) = \gamma(-k)$, and $\hat{\gamma}(k) = \hat{\gamma}(-k)$. Suppose $0 \leq k \leq L < \infty$, and $p \to \infty$ we have

$$\hat{\gamma}(k) - \gamma(k) = \frac{1}{p-k}\sum_{j=1}^{p-k}[(t_j^2 - T_n)(t_{j+k}^2 - T_n) - \mathrm{cov}(t_j^2, t_{j+k}^2)], \tag{B.2}$$

and

$$
\begin{aligned}
&\frac{1}{p-k}\sum_{j=1}^{p-k}(t_j^2-T_n)(t_{j+k}^2-T_n)\\
=&\frac{1}{p-k}\sum_{j=1}^{p-k}(t_j^2-1+1-T_n)(t_{j+k}^2-1+1-T_n)\\
=&\frac{1}{p-k}\sum_{j=1}^{p-k}[(t_j^2-1)(t_{j+k}^2-1)+(t_j^2-1)(1-T_n)+(t_{j+k}^2-1)(1-T_n)]+(1-T_n)^2 \qquad \text{(B.3)}\\
=&\frac{1}{p-k}\sum_{j=1}^{p-k}[(t_j^2-1)(t_{j+k}^2-1)]-(1-T_n)^2+O_p(p^{-1})\\
=&\frac{1}{p-k}\sum_{j=1}^{p-k}[(t_j^2-1)(t_{j+k}^2-1)]+O_p(p^{-1}).
\end{aligned}
$$

The last equality rests on the fact that $\sqrt{p}(T_n-1)$ converges to normal distribution. Let $z_{j,k}=(t_j^2-1)(t_{j+k}^2-1)-\text{cov}(t_j^2,t_{j+k}^2)$, we have

$$
\begin{aligned}
E(z_{j,k}) &= E[(t_j^2-1)(t_{j+k}^2-1)]-\text{cov}(t_j^2,t_{j+k}^2)\\
&= E[(t_j^2-Et_j^2+Et_j^2-1)(t_{j+k}^2-Et_{j+k}^2+Et_{j+k}^2-1)]-\text{cov}(t_j^2,t_{j+k}^2)\\
&= \text{cov}(t_j^2,t_{j+k}^2)+(Et_j^2-1)(Et_{j+k}^2-1)-\text{cov}(t_j^2,t_{j+k}^2)\\
&= O(n^{-2})
\end{aligned}
\qquad \text{(B.4)}
$$

Then the sequence $\{z_{1,k},z_{2,k},...\}$ is an α-mixing sequence with approximate mean 0. Adopting the CLT for α-mixing process Wang and Akritas (2010) will prove that $\sqrt{p}[\hat{\gamma}(k)-\gamma(k)]$ weakly converges to a normal distribution centered at 0 provided $p=o(n^4)$. □

B.1.6 Proof of Proposition 2

Proof. Consider

$$\begin{aligned} t_j^{2k} &= (\overline{X}_j - \overline{Y}_j)^{2k}(S_{1j}^2/n + S_{2j}^2/m)^{-k} \\ &= (\sigma_{1j}^2/n + \sigma_{2j}^2/m)^{-k}(\overline{X}_j - \overline{Y}_j)^{2k}(1+B_n)^k \\ &= (\sigma_{1j}^2 N/n + \sigma_{2j}^2 N/m)^{-k}(\sqrt{N}\overline{X}_j - \sqrt{N}\overline{Y}_j)^{2k}(1+B_n)^k, \end{aligned}$$

where

$$B_n = -\frac{(S_{1j}^2 - \sigma_{1j}^2) + \frac{n}{m}(S_{2j}^2 - \sigma_{2j}^2)}{S_{1j}^2 + \frac{n}{m}S_{2j}^2}.$$

Applying Hölder's inequality on the expectation of t_j^{2k}, we have

$$E\left[\left|(\sqrt{N}\overline{X}_j - \sqrt{N}\overline{Y}_j)^{2k}(1+B_n)^k\right|\right] \le \left[E|\sqrt{N}\overline{X}_j - \sqrt{N}\overline{Y}_j|^{2k+2}\right]^{\frac{2k}{2k+2}}\left[E|1+B_n|^{k(2k+2)}\right]^{\frac{1}{2k+2}}.$$

We first show that $E|1+B_n|^{k(2k+2)} < \infty$. Based on the strong law of large number, $B_n \to 0$ almost surely. Applying Taylor expansion, we have $(1+B_n)^{k(2k+2)} = 1+k(2k+2)B_n+O_p(B_n^2)$. By Cauchy-Schwarz inequality,

$$\begin{aligned} E|B_n| &= E[|B_n|(S_{1j}^2 + \tfrac{n}{m}S_{2j}^2)(S_{1j}^2 + \tfrac{n}{m}S_{2j}^2)^{-1}] \\ &\le \{E[|B_n|(S_{1j}^2 + \tfrac{n}{m}S_{2j}^2)]^2\}^{\frac{1}{2}}[E(S_{1j}^2 + \tfrac{n}{m}S_{2j}^2)^{-2}]^{\frac{1}{2}}. \end{aligned} \tag{B.5}$$

For the first term,

$$E[B_n(S_{1j}^2 + \tfrac{n}{m}S_{2j}^2)]^2 = E[(S_{1j}^2 - \sigma_{1j}^2) + \tfrac{n}{m}(S_{2j}^2 - \sigma_{2j}^2)]^2 = \mathrm{Var}(S_{1j}^2) + \tfrac{n^2}{m^2}\mathrm{Var}(S_{2j}^2) \to 0, \tag{B.6}$$

when n and m go to infinity based on the central limit theorem for the sample variances. For the second term in (B.5), using Taylor approximation again, we have $(S_{1j}^2 + S_{2j}^2 n/m)^{-2} = (\sigma_{1j}^2 + \sigma_{2j}^2 n/m)^{-2}(1-2D_n)+O_p(D_n^2)$, where $D_n = (S_{1j}^2 + S_{2j}^2 n/m)/(\sigma_{1j}^2 + \sigma_{2j}^2 n/m) - 1$ converges to 0 almost surely. Because $E(D_n) = 0$, then $E(S_{1j}^2 + S_{2j}^2 n/m)^{-2} = (\sigma_{1j}^2 + \sigma_{2j}^2 n/m)^{-2} +$

$E(O_p(D_n^2))$. When $k > 1$, the conditions $\sup_j E(X_j^{2k+2}) < \infty$ and $\sup_j E(Y_j^{2k+2}) < \infty$ imply $\sup_j E(X_j^{4+\epsilon}) < \infty$, $\sup_j E(Y_j^{4+\epsilon}) < \infty$ for some $\epsilon > 0$. Then the uniform integrability of D_n^2 can be guaranteed, which leads to $E(D_n^2) \to 0$. Thus $E(S_{1j}^2 + S_{2j}^2 n/m)^{-2} = (\sigma_{1j}^2 + \sigma_{2j}^2 n/m)^{-2} + o(1) < \infty$. (B.5) and (B.6) together verify that $E(B_n) \to 0$.

Next, we will show that $E|\sqrt{N}\overline{X}_j - \sqrt{N}\overline{Y}_j|^{2k+2} \leq \infty$. Due to c_r-inequality ,

$$E|\sqrt{N}\overline{X}_j - \sqrt{N}\overline{Y}_j|^{2k+2} \leq 2^{2k+1}\left[E\left(|\sqrt{N}\overline{X}_j|^{2k+2}\right) + E\left(|\sqrt{N}\overline{Y}_j|^{2k+2}\right)\right]. \tag{B.7}$$

For brevity, we drop the subscript j for the j^{th} component X_{ij}. Because n, m and N are all of the same order, we will consider $\sqrt{n}\overline{X}$ and $\sqrt{m}\overline{Y}$ in below. Note that if $EX_1^{2k+2} < \infty$,

$$\begin{aligned}
E(\sqrt{n}|\overline{X}|)^{2k+2} &= n^{(2k+2)/2} n^{-(2k+2)} E \sum_{i_1}^{n} X_{i_1} \sum_{i_2}^{n} X_{i_2} \ldots \sum_{i_{2k+2}}^{n} X_{i_{2k+2}} \\
&= n^{-k-1}\left[\sum_{i_1} EX_{i_1}^{2k+2} + \binom{2k+2}{2} \sum_{i_1 \neq i_2} EX_{i_1}^{2k} EX_{i_2}^{2} + \ldots \right. \\
&\quad \left. + \frac{(2k+2)!}{2^{k+1}} \sum_{i_1 \neq \ldots \neq i_{k+1}} EX_{i_1}^2 \ldots EX_{i_{k+1}}^2 \right] \\
&= n^{-k} C_1 EX_1^{2k+2} + n^{-k+1} C_2 EX_1^{2k-1} EX_1^2 + \ldots + C_{k+1}(EX_1^2)^{k+1} < \infty,
\end{aligned}$$

for some constants C_1, C_2, ..., $C_{k+1} < \infty$. It is analogous to show $E(\sqrt{m}|\overline{Y}|)^{2k+2} < \infty$. □

B.1.7 Proof of Lemma 6

Proof. With the moment conditions assumed, $\sup_j E|X_{1j}Y_{1j}| < \infty$ by Cauchy-Schwarz inequality, and $\gamma_{j,j'} < \infty$. Because the samples are independently identically distributed, using central limit theorem, those sample covariances and sample variances multiplied by the square root of their sample sizes have asymptotic normal distributions centered at their population counterparts. We can write the sample estimators as $S_{1jj'} = \sigma_{1jj'} + O_p(N^{-1/2})$, $S_{2jj'} = \sigma_{2jj'} + O_p(N^{-1/2})$, $S_{1j} = \sigma_{1j}^2 + O_p(N^{-1/2})$, $S_{2j} = \sigma_{2j}^2 + O_p(N^{-1/2})$, $S_{1j'} = \sigma_{1j'}^2 + O_p(N^{-1/2})$, $S_{2j'} = \sigma_{2j'}^2 + O_p(N^{-1/2})$.

When $\gamma_{j,j'} > 0$, the difference can be written as $\tilde{\gamma}_{j,j'} - \gamma_{j,j'} = \gamma_{j,j'}(\tilde{\gamma}_{j,j'}/\gamma_{j,j'} - 1)$. The ratio can be further expanded as

$$\tilde{\gamma}_{j,j'}/\gamma_{j,j'} = \left[\frac{S_{1jj'}/\lambda_1 + S_{2jj'}/\lambda_2}{\sigma_{1jj'}/\lambda_1 + \sigma_{2jj'}/\lambda_2}\right]^2 \frac{\sigma^2_{1j}/\lambda_1 + \sigma^2_{2j}/\lambda_2}{S^2_{1j}/\lambda_1 + S^2_{2j}/\lambda_2} \ \frac{\sigma^2_{1j'}/\lambda_1 + \sigma^2_{2j'}/\lambda_2}{S^2_{1j'}/\lambda_1 + S^2_{2j'}/\lambda_2}.$$

For the first ratio term, note that

$$\frac{S_{1jj'}/\lambda_1 + S_{2jj'}/\lambda_2}{\sigma_{1jj'}/\lambda_1 + \sigma_{2jj'}/\lambda_2} = \frac{[\sigma_{1jj'} + O_p(N^{-1/2})]/\lambda_1 + [\sigma_{2jj'} + O_p(N^{-1/2})]/\lambda_2}{\sigma_{1jj'}/\lambda_1 + \sigma_{2jj'}/\lambda_2} = 1 + O_p(N^{-1/2}).$$

Similarly, the rest two ratio terms can be shown as $1 + O_p(N^{-1/2})$ by applying Taylor expansion $[1 + O_p(N^{-1/2})]^{-1} = 1 + O_p(N^{-1/2})$. Then $\hat{\gamma}_{j,j'}/\gamma_{j,j'} = 1 + O_p(N^{-1/2})$. With $0 < |\gamma_{j,j'}| < \infty$, the result is proved.

Then we consider $\gamma_{j,j'} = 0$ which occurs only when $\sigma_{1jj'} = \sigma_{2jj'} = 0$. By similar claim as $\gamma_{j,j'} > 0$ case, we have

$$\begin{aligned}
\tilde{\gamma}_{j,j'} &= \frac{2(S_{1jj'}/\lambda_1 + S_{2jj'}/\lambda_2)^2}{(S^2_{1j}/\lambda_1 + S^2_{2j}/\lambda_2)(S^2_{1j'}/\lambda_1 + S^2_{2j'}/\lambda_2)} \\
&= \frac{[O_p(N^{-1/2})]^2}{(\sigma^2_{1j}/\lambda_1 + \sigma^2_{2j}/\lambda_2)(1 + O_p(N^{-1/2}))(\sigma^2_{1j'}/\lambda_1 + \sigma^2_{2j'}/\lambda_2)(1 + O_p(N^{-1/2}))} \\
&= O_p(N^{-1}).
\end{aligned}$$

□

B.1.8 Proof of Theorem 4

Proof. First we reorganize the difference between the estimator and the true variance as

$$\begin{aligned}
\tilde{\zeta}_n^2 - \mathrm{var}(\sqrt{p}T_n) &= p^{-1} \sum_{|j-j'|\le L} \tilde{\gamma}_{j,j'} - \mathrm{var}\left(p^{-\frac{1}{2}} \sum_{j=1}^{p} t_j^2\right) \\
&= p^{-1} \sum_{|j-j'|\le L} \tilde{\gamma}_{j,j'} - p^{-1} \sum_{j=1}^{p} \sum_{j'=1}^{p} \mathrm{cov}(t_j^2, t_{j'}^2) \\
&= p^{-1} \sum_{|j-j'|\le L} \left(\tilde{\gamma}_{j,j'} - \mathrm{cov}(t_j^2, t_{j'}^2)\right) - p^{-1} \sum_{|j-j'|> L} \mathrm{cov}(t_j^2, t_{j'}^2).
\end{aligned}$$

For the first term, using Lemma 5 and 6 we have

$$\begin{aligned}
\left| p^{-1} \sum_{|j-j'|\le L} \left(\tilde{\gamma}_{j,j'} - \mathrm{cov}(t_j^2, t_{j'}^2)\right) \right| &\le p^{-1} \sum_{|j-j'|\le L} \left|\tilde{\gamma}_{j,j'} - \mathrm{cov}(t_j^2, t_{j'}^2)\right| \\
&= p^{-1} \sum_{|j-j'|\le L} \left|\mathrm{cov}(t_j^2, t_{j'}^2)\right| \left| \frac{\tilde{\gamma}_{j,j'}}{\mathrm{cov}(t_j^2, t_{j'}^2)} - 1 \right| \\
&= p^{-1} \sum_{|j-j'|\le L} \left|\mathrm{cov}(t_j^2, t_{j'}^2)\right| O_p(N^{-1/2}) \\
&\le var(\sqrt{p}T_n) O_p(N^{-1/2}) \\
&= O_p(N^{-1/2})
\end{aligned}$$

The second term has the following property due to inequality (3.12):

$$\left| p^{-1} \sum_{|j-j'|> L} \mathrm{cov}(t_j^2, t_{j'}^2) \right| = \left| 2p^{-1} \sum_{j=1}^{p-L} \sum_{r=L}^{p-j} \mathrm{cov}(t_j^2, t_{j+r}^2) \right| \le C \sum_{r=L}^{p} [\alpha(r)]^{\nu/(2+\nu)}$$

for some constant C.

If $\alpha(r)$ satisfies $\sum_{r=1}^{\infty} r[\alpha(r)]^{\nu/(2+\nu)} < \infty$, then

$$\sum_{r=L}^{p} [\alpha(r)]^{\nu/(2+\nu)} \le CL^{-1} \sum_{r=L}^{p} r[\alpha(r)]^{\nu/(2+\nu)} = O(L^{-1}).$$

If $\alpha(r) = O(r^{-h})$ for some $a > 1$, then $\sum_{r=L}^{p}[\alpha(r)]^{\nu/(2+\nu)} = K\sum_{r=L}^{p} r^{-h\nu/(2+\nu)}$ for some constant K. If $\nu > 2/(h-1)$, then $h\nu/(2+\nu) > 1$ and we have

$$\sum_{r=L}^{p} r^{-h\nu/(2+\nu)} \leq \int_{L}^{p} r^{-h\nu/(2+\nu)}\, \mathrm{d}r = \frac{r^{1-h\nu/(2+\nu)}}{1-h\nu/(2+\nu)}\Big|_{L}^{p} = O(p^{1-h\nu/(2+\nu)}).$$

□

B.1.9 Proof of Corollary 4

Proof. The consistency of $\tilde{\zeta}_n^2$ is shown in Theorem 4. We only need to examine the estimator for the center $E(T_n) = 1+n^{-1}a_n+n^{-2}b_n+O(n^{-3})$. J_{n1} uses 1 as the center estimator. When $p = o(N^2)$, $\sqrt{p}[E(T_n)-1] = o(N)[n^{-1}a_n+n^{-2}b_n+O(n^{-3})] = o(1)$. J_{n2} uses $1+n^{-1}\hat{a}_n+n^{-2}\hat{b}_n$ as the center estimator. When $p = o(N^6)$, $\sqrt{p}[E(T_n) - (1+n^{-1}\hat{a}_n+n^{-2}\hat{b}_n)] = n^{-1}\sqrt{p}(a_n - \hat{a}_n)+n^{-2}\sqrt{p}(b_n-\hat{b}_n)+\sqrt{p}O(n^{-3}) = o_p(1)$. To see this, let $A_n = n^{-1}\sqrt{p}(a_n-\hat{a}_n)$. Due to the α-mixing condition, we have $E(A_n^2) = n^{-2}p^{-1}\sum_{j=1}^{p}\sum_{j'=1}^{p}(c_{nj}-\hat{c}_{nj})(c_{nj'}-\hat{c}_{nj'}) = O(n^{-2})$, so $A_n \xrightarrow{p} 0$ when $n \to \infty$. Similarly $n^{-2}\sqrt{p}(b_n-\hat{b}_n) \xrightarrow{p} 0$. □

B.1.10 Proof of Corollary 7

Proof.

$$\begin{aligned} E[|X_n|^k I_{|X_n|^k \geq M}] =& E[|X_n|^k I_{|X_n|\geq M^{1/k}}] \leq E[|X_n|^{b-k} M^{-(b-k)/k}|X_n|^k I_{|X_n|\geq M^{1/k}}] \\ =& M^{1-b/k}E[|X_n|^b I_{|X_n|\geq M^{1/k}}] \leq M^{1-b/k}E|X_n|^b. \end{aligned}$$

Since $1-b/k < 0$, we conclude that $\lim_{M\to\infty}\limsup_{N\to\infty} E[|X_n|^k I_{|X_n|^k\geq M}] = 0$, or that X_n^k is uniformly integrable, and $E(X_n^k) \to E(X^k)$. □

B.2 Expressions for c_{nj}, d_{nj}, $\hat{c}_{nj}$ and $\hat{d}_{nj}$

The expressions are directly obtained from Gregory et al. (2015). Firstly, define c_{nj} and d_{nj} in the following way.

$$c_{nj} = \tau_{nj}^{-2}[\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2] + 2\tau_{nj}^{-6}[\mu'_{3j} + (n/m)^2\eta'_{3j}]$$

$$\begin{aligned}
d_{nj} =& \tau_{nj}^{-4}\{[\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2] - [\mu'_{4j} - 3\sigma_{1j}^4 + (n/m)^4(\eta'_{4j} - 3\sigma_{2j}^4)]\} \\
&+ \tau_{nj}^{-6}[\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2][\mu'_{4j} - \sigma_{1j}^4 + (n/m)^3(\eta'_{4j} - \sigma_{2j}^4)]\} \\
&- 4\tau_{nj}^{-6}[\mu'_{3j} + (n/m)^2\eta'_{3j}][\mu'_{3j} + (n/m)^3\eta'_{3j}] \\
&- 2\tau_{nj}^{-6}[(\mu'_{3j})^2 + (n/m)^5(\eta'_{3j})^2] \\
&- 6\tau_{nj}^{-8}[\mu'_{3j} + (n/m)^2\eta'_{3j}][\mu'_{5j} - 2\mu'_{3j}\sigma_{1j}^2 + (n/m)^4(\eta'_{5j} - 2\eta'_{3j}\sigma_{2j}^2)] \\
&- 3\tau_{nj}^{-8}[(\mu'_{4j} - \sigma_{1j}^4) + (n/m)^3(\eta'_{4j} - \sigma_{2j}^4)]^2 \\
&+ 6\tau_{nj}^{-8}[\sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2][\mu'_{3j} + (n/m)^2\eta'_{3j}]^2 \\
&+ 3\tau_{nj}^{-10}[\sigma_{1j}^2 + (n/m)\sigma_{2j}^2][\mu'_{4j} - \sigma_{1j}^4 + (n/m)^3(\eta'_{4j} - \sigma_{2j}^4)]^2 \\
&+ 12\tau_{nj}^{-10}[\mu'_{3j} + (n/m)^2\eta'_{3j}]^2[\mu'_{4j} - \sigma_{1j}^4 + (n/m)^3(\eta'_{4j} - \sigma_{2j}^4)]
\end{aligned}$$

where $\tau_{nj}^2 = \sigma_{1j}^2 + (n/m)^2\sigma_{2j}^2$, μ'_{kj} and η'_{kj} are the k^{th} central moments of X_{1j} and Y_{1j}, respectively. Replacing the moment parameters in c_{nj} and d_{nj} with their sample estimates, we can get $\hat{c}_{nj}$ and $\hat{d}_{nj}$, respectively.

B.3 Calculation of var($\sqrt{p}T_n^{(1)}$)

Much of the calculation is the same as the proof of Lemma 1. We only need to compute

$$\begin{aligned}
&\operatorname{cov}[(\overline{X}_j - \overline{Y}_j)^2, (\overline{X}_{j'} - \overline{Y}_{j'})^2] \\
=&\operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)} + \delta_j)^2, (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)} + \delta_{j'})^2] \\
=&\operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})^2, (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)})^2] + 2\delta_j \operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)}), (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)})^2] \\
&+ 2\delta_{j'} \operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})^2, (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)})] + 4\delta_j\delta_{j'} \operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)}), (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)})]
\end{aligned} \tag{B.8}$$

The first term is simply the covariance under H_0 shown in Lemma 1. For the rest terms, noticing the independence between X and Y and the fact that $\overline{X}_j^{(0)}$ and $\overline{Y}_j^{(0)}$ are of zero means, we have

$$\begin{aligned}
&\operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)}), (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)})^2] \\
=&\operatorname{cov}([\overline{X}_j^{(0)}, (\overline{X}_{j'}^{(0)})^2] + \operatorname{cov}[\overline{Y}_j^{(0)}, (\overline{Y}_{j'}^{(0)})^2] \\
=&n^{-3}\operatorname{cov}(\sum_{i_1=1}^{n} X_{ji_1}^{(0)}, \sum_{i_2=1}^{n}\sum_{i_3=1}^{n} X_{j'i_2}^{(0)} X_{j'i_3}^{(0)}) + m^{-3}\operatorname{cov}(\sum_{i_1=1}^{m} Y_{ji_1}^{(0)}, \sum_{i_2=1}^{m}\sum_{i_3=1}^{m} Y_{j'i_2}^{(0)} Y_{j'i_3}^{(0)}) \\
=&n^{-3}E(\sum_{i_1=i_2=i_3}^{n} X_{ji_1}^{(0)} X_{j'i_2}^{(0)} X_{j'i_3}^{(0)}) + m^{-3}E(\sum_{i_1=i_2=i_3}^{m} Y_{ji_1}^{(0)} Y_{j'i_2}^{(0)} Y_{j'i_3}^{(0)}) \\
=&n^{-2}E(X_{j1}^{(0)}(X_{j'1}^{(0)})^2) + m^{-2}E(Y_{j1}^{(0)}(Y_{j'1}^{(0)})^2).
\end{aligned} \tag{B.9}$$

Similarly it follows that $\operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)})^2, (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)})] = n^{-2}E(X_{j'1}^{(0)}(X_{j1}^{(0)})^2) + m^{-2}E(Y_{j'1}^{(0)}(Y_{j1}^{(0)})^2)$. The last term is $\operatorname{cov}[(\overline{X}_j^{(0)} - \overline{Y}_j^{(0)}), (\overline{X}_{j'}^{(0)} - \overline{Y}_{j'}^{(0)})] = \operatorname{cov}[\overline{X}_j^{(0)}, \overline{X}_{j'}^{(0)}] + \operatorname{cov}[\overline{Y}_j^{(0)}, \overline{Y}_{j'}^{(0)}] = n^{-1}\sigma_{1jj'} + m^{-1}\sigma_{2jj'}$. Then the result is as equation (3.19) with some basic algebra.