

Article

High-Dimensional U-Statistics Type Hypothesis Testing via Jackknife Pseudo-Values with Multiplier Bootstrap

Mingjuan Zhang and Libin Jin *

School of Statistics and Mathematics, Interdisciplinary Research Institute of Data Science, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China; zhang_mingjuan@lixin.edu.cn

* Correspondence: jinlb1987@hotmail.com

Abstract: High-dimensional parameter testing is commonly used in bioinformatics to analyze complex relationships in gene expression and brain connectivity studies, involving parameters like means, covariances, and correlations. In this paper, we present a novel approach for testing U -statistics-type parameters by leveraging jackknife pseudo-values. Inspired by Tukey's conjecture, we establish the asymptotic independence of these pseudo-values, allowing us to reformulate U -statistics-type parameter testing as a sample mean testing problem. This reformulation enables the use of established sample mean testing frameworks, simplifying the testing procedure. We apply a multiplier bootstrap method to obtain critical values and provide a rigorous theoretical analysis to validate the approach. Simulation studies demonstrate the robustness of our method across a variety of scenarios. Additionally, we apply our approach to investigate differences in the dependency structures of a subset of genes within the Wnt signaling pathway, which is associated with lung cancer.

Keywords: high-dimensional; hypothesis testing; U -statistic; jackknife; multiplier bootstrap

MSC: 62F03



Citation: Zhang, M.; Jin, L. High-Dimensional U-Statistics Type Hypothesis Testing via Jackknife Pseudo-Values with Multiplier Bootstrap. *Mathematics* **2024**, *12*, 3837. <https://doi.org/10.3390/math12233837>

Academic Editor: Gaorong Li

Received: 30 October 2024

Revised: 2 December 2024

Accepted: 3 December 2024

Published: 4 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the development of technology, high-dimensional parameter testing is widely used in various bioinformatics analyses. Numerous studies employ mean tests to investigate changes in genes of interest, such as [1,2]. Additionally, some studies use covariance matrix or precision matrix tests to illustrate brain connectivity analysis and apply correlation matrix tests for gene co-expression network analysis, such as [3–5]. In this paper, we concentrate on U -statistics type parameter testing, encompassing but not limited to mean, covariance, and correlation tests.

High-dimensional mean tests have been well studied, as reviewed by [6]. We categorize the proposed test statistics into five broad groups. Firstly, ℓ_2 -norm-based test statistics, such as [7–11], are known for their effectiveness under dense alternatives. Another group is the ℓ_∞ -norm-based tests, as evidenced by [12–14], which are more suitable for sparse alternatives. Since the alternative hypothesis is usually unknown in advance, a third category of test statistics aims to accommodate diverse alternatives by combining the p -values of tests based on various norms, such as [15–17]. Additionally, some tests simplify the problem by projecting the high-dimensional mean vector onto lower dimensions. For instance, some studies explore random projections, such as [18,19], while ref. [20] seeks the optimal projection directions. Most of the tests mentioned earlier impose sparsity conditions on covariance matrices, but dense patterns are common in reality. To address this, studies such as [21–24] enhance the signal strength and test performance by incorporating common factors.

High-dimensional covariance tests have also achieved significant advancements in recent years. The one-sample covariance test ($H_0: \Sigma = I$) mainly used methods including spectral norm-based tests, such as [25–27], and Frobenius norm-based tests, such as [28–30].

For the two-sample covariance test, Frobenius norm-based tests, such as [30–33], perform well for dense alternatives, while tests based on maximum entry-wise norm, such as [34], show a strong performance for sparse alternatives.

High-dimensional correlation tests have received significant attention. For the one-sample Pearson correlation test, ref. [35] proposes a test suitable for sparse alternatives, while ref. [36] provides a test that is powerful for dense alternatives. For the two-sample correlation test, ref. [37] introduces a test suitable for sparse alternatives, and ref. [38] develops a general framework for testing correlation structures across one, two, and multiple sample scenarios. In addition to Pearson correlation matrix tests, rank-based correlation matrix tests, including Kendall's tau and Spearman's rank correlation, have been well studied, as demonstrated by [39–41]. Furthermore, ref. [42] proposes a framework for the equality test of U -statistic-based correlation matrices.

For all the aforementioned parameter tests, there exist U -statistic-based test statistics, such as [10,30,33,34,42]. There are also some adaptive and unified tests. Ref. [16] proposes a unified framework for testing high-dimensional parameters that can be estimated by U -statistic-based vectors. Ref. [43] constructs a family of U -statistics as unbiased estimators for ℓ_p -norms of test parameters, further combining the p -values of different orders. Ref. [44] proposes a two-step Gaussian approximation for high-dimensional non-degenerate U -statistics and a bootstrap method for computing their probabilities within hyper-rectangles.

In this paper, we propose a novel approach for U -statistics-type parameters. Inspired by Tukey's conjecture [45], we establish the asymptotic independence of jackknife pseudo-values for the U -statistic estimator. By constructing test statistics based on the sample means of these pseudo-values, we effectively transfer U -statistic-type testing into the sample mean testing framework. This reformulation allows us to apply established methods from sample mean testing, simplifying the testing procedure. We derive the critical values for our test statistics by applying a multiplier bootstrap to the pseudo-values. In addition, we conduct a comprehensive theoretical analysis of our proposed test, including validation of the multiplier bootstrap procedure for accurate critical value estimation, as well as an assessment of its asymptotic properties, such as size control and power performance.

The rest of this paper is organized as follows: In Section 2, we present the detailed testing procedures. Section 3 verifies the effectiveness of the multiplier bootstrap used in Section 2 and analyzes the theoretical performance of our proposed tests. Section 4 presents simulation results to justify the empirical performance of our methods. In Section 5, we apply our methodology to analyze the dependency differences in the Wnt signaling pathway between lung cancer patients and control patients. Finally, some conclusions and discussions are provided in Section 6.

2. Methodology

Let $\mathbf{X} = (X_1, \dots, X_d)^\top$ and $\mathbf{Y} = (Y_1, \dots, Y_d)^\top$ be two d -dimensional random vectors independent of each other. $\mathbf{X}_1, \dots, \mathbf{X}_{n_1}$ are independent and identically distributed (i.i.d.) random samples from $\mathbf{X}$ with $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{id})^\top$. Similarly, $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2}$ are i.i.d. random samples from $\mathbf{Y}$ with $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{id})^\top$. We set $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_{n_1}\}$, $\mathcal{Y} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2}\}$, and

$$\begin{aligned}\tilde{u}_{1,s} &= \binom{n_1}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n_1} \Phi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}), \\ \tilde{u}_{2,s} &= \binom{n_2}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n_2} \Phi_s(\mathbf{Y}_{i_1}, \dots, \mathbf{Y}_{i_m}),\end{aligned}\quad (1)$$

where $s = 1, \dots, q$, with q being the dimensionality of the parameter we are interested in, and Φ_s is an m -order symmetric kernel function. We assume that Φ_s is symmetric and that

each kernel function is of the same order m only for notational simplicity. We then define two U -statistic based vectors as (1),

$$\tilde{\mathbf{u}}_1 := (\tilde{u}_{1,1}, \tilde{u}_{1,2}, \dots, \tilde{u}_{1,q})^\top, \quad \text{and} \quad \tilde{\mathbf{u}}_2 := (\tilde{u}_{2,1}, \tilde{u}_{2,2}, \dots, \tilde{u}_{2,q})^\top.$$

We use $\mathbf{u}_\gamma$ to denote the expectation of $\tilde{\mathbf{u}}_\gamma$, i.e., $\mathbf{u}_\gamma = (u_{\gamma,1}, u_{\gamma,2}, \dots, u_{\gamma,q})^\top$ with $u_{\gamma,s} = \mathbb{E}[\tilde{u}_{\gamma,s}]$ for $\gamma = 1, 2$ and $s = 1, \dots, q$. We are interested in testing the following hypotheses:

(i) (One-sample problem) For a given $\mathbf{u}_0 \in \mathbb{R}^q$

$$\mathbf{H}_0 : \mathbf{u}_1 = \mathbf{u}_0 \quad \text{v.s.} \quad \mathbf{H}_1 : \mathbf{u}_1 \neq \mathbf{u}_0. \quad (2)$$

(ii) (Two-sample problem)

$$\mathbf{H}_0 : \mathbf{u}_1 = \mathbf{u}_2 \quad \text{v.s.} \quad \mathbf{H}_1 : \mathbf{u}_1 \neq \mathbf{u}_2. \quad (3)$$

Intuitively, with the estimations for $\mathbf{u}_\gamma$ ($\gamma = 1, 2$) in (1), the two problems above in (2) and (3) can be treated as the one sample mean and two sample mean tests in high dimensions. However, it is difficult to obtain or calculate the asymptotic distribution for the test statistics based on the U -statistics estimation in (1). Hence, to overcome this obstacle, we propose the test based on the jackknife pseudo-values, which are defined by

$$\hat{\mathbf{u}}_{\gamma,i} = n_\gamma \tilde{\mathbf{u}}_\gamma - (n_\gamma - 1) \tilde{\mathbf{u}}_\gamma^{-i}, \quad (4)$$

where $\tilde{\mathbf{u}}_1^{-i}$ is the U -statistic vector based on $\mathcal{X} \setminus \{\mathbf{X}_i\}$, and $\tilde{\mathbf{u}}_2^{-i}$ is the U -statistic vector based on $\mathcal{Y} \setminus \{\mathbf{Y}_i\}$. In [45,46], the jackknife pseudo-values are unbiased estimators of $\mathbf{u}_\gamma$, i.e.,

$$\mathbb{E} \hat{\mathbf{u}}_{\gamma,i} = \mathbf{u}_\gamma. \quad (5)$$

According to [47], the jackknife pseudo-values are not only uncorrelated but also asymptotically independent. Hence, $\mathbf{u}_\gamma$ can be estimated by a sample mean of jackknife pseudo-values,

$$\hat{\mathbf{u}}_\gamma = n_\gamma^{-1} \sum_{i=1}^{n_\gamma} \hat{\mathbf{u}}_{\gamma,i} \quad (6)$$

Further, we provide the variance estimators for $\sqrt{n} \hat{\mathbf{u}}_{\gamma,s}$ as follows:

$$\hat{\sigma}_{\gamma,ss} = (n_\gamma - 1)^{-1} \sum_{i=1}^{n_\gamma} (\hat{\mathbf{u}}_{\gamma,i,s} - \hat{\mathbf{u}}_{\gamma,s})^2, \quad \gamma = 1, 2, \quad s = 1, \dots, q. \quad (7)$$

Remark 1. The variance estimator in (7) is also the delete-1 jackknife estimator for $\sqrt{n} \tilde{\mathbf{u}}_{\gamma,s}$. As long as the $\tilde{u}_{\gamma,s}$ is a smooth function of the observations, this jackknife estimator is consistent.

Hence, for the one sample testing problem (2), we construct the following test statistic:

$$T_0 = \max_{1 \leq s \leq q} \frac{|\hat{u}_{1,s} - u_{0,s}|}{\sqrt{\hat{\sigma}_{1,ss}/n_1}}. \quad (8)$$

For the two sample testing problem (3), we set the test statistic as follows:

$$T_d = \max_{1 \leq s \leq q} \frac{|\hat{u}_{1,s} - \hat{u}_{2,s}|}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}}. \quad (9)$$

Remark 2. If $\mathbf{u}_\gamma$ represents the mean of $\mathbf{X}$, the jackknife pseudo-values reduce to the independent sample observations, and the test statistics simplify to the sample mean test in [12,13].

In high dimensions, for the centered independent random vectors $\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\} \in \mathbb{R}^d$, ref. [48] derives that the distribution of $\max_{1 \leq j \leq d} n^{-1/2} \sum_{i=1}^n Z_{i,j}$ can be approximated by the maximum of the sum of the Gaussian random vectors with the same covariance matrices as $\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$. Meanwhile, ref. [48] proposes a multiplier bootstrap procedure to obtain the Gaussian random vectors. Motivated by [48], we apply the multiplier bootstrap procedure to the asymptotically independent jackknife pseudo-values $\hat{u}_{\gamma,i}$. By this procedure, one can approximate the distribution for T_o and T_d in (8) and (9). Specifically, set $\xi_{\gamma,1}^b, \xi_{\gamma,2}^b, \dots, \xi_{\gamma,n_\gamma}^b$ as a sequence of iid $N(0, 1)$ with $\gamma = 1, 2$ and $b = 1, 2, \dots, B$. The b -th multiplier bootstrap samples of $\hat{u}_{\gamma,i}$ are $\{\xi_{\gamma,1}^b \hat{u}_{\gamma,1}, \dots, \xi_{\gamma,n_\gamma}^b \hat{u}_{\gamma,n_\gamma}\}$. Correspondingly, the b -th multiplier bootstrap sample of $\hat{u}_{\gamma,s}$ is as follows:

$$\hat{u}_{\gamma,s}^b = \frac{1}{n_\gamma} \sum_{i=1}^{n_\gamma} \xi_{\gamma,i}^b (\hat{u}_{\gamma,i,s} - \hat{u}_{\gamma,s}) \quad \text{for } \gamma = 1, 2, \quad \text{and } s = 1, 2, \dots, q. \quad (10)$$

Based on $\hat{u}_{\gamma,s}^b$, we define the b -th bootstrap sample of the test statistics T_o and T_d by the following:

$$T_o^b = \max_{1 \leq s \leq q} \frac{|\hat{u}_{1,s}^b - u_{0,s}|}{\sqrt{\sigma_{1,ss}/n_1}}, \quad T_d^b = \max_{1 \leq s \leq q} \frac{|\hat{u}_{1,s}^b - \hat{u}_{2,s}^b|}{\sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}}.$$

Based on these multiplier bootstrap samples, the critical value and p -value for T_o and T_d can be estimated by

$$\hat{t}_\alpha^o = \inf \left\{ t \in \mathbb{R} : \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{T_o^b \leq t\} > 1 - \alpha \right\}, \quad \hat{t}_\alpha^d = \inf \left\{ t \in \mathbb{R} : \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{T_d^b \leq t\} > 1 - \alpha \right\}.$$

After obtaining the critical value, we obtain the test for the hypothesis in (2) and (3) as follows:

$$\phi_\alpha^o = \mathbf{1}\{T_o > \hat{t}_\alpha^o\}, \quad \phi_\alpha^d = \mathbf{1}\{T_d > \hat{t}_\alpha^d\}.$$

Thus, we reject (2) if and only if $\Phi_\alpha^o = 1$, and reject (3) if and only if $\Phi_\alpha^d = 1$. Correspondingly, we can construct the p -value estimator of T_o and T_d as

$$\hat{P}_o = \frac{1}{B+1} \sum_{b=1}^B \mathbf{1}\{T_o^b > T_o\}, \quad \hat{P}_d = \frac{1}{B+1} \sum_{b=1}^B \mathbf{1}\{T_d^b > T_d\}. \quad (11)$$

For a given significance level α , we then reject the $\mathbf{H}_0$ of (2) if and only if $\hat{P}_o \leq \alpha$, and reject (3) if and only if $\hat{P}_d \leq \alpha$.

3. Theoretical Analysis

In this section, we justify the validity of the multiplier bootstrap method used in the last section and study the empirical size and power of our proposed tests. Before presenting the detailed theoretical results, some mild assumptions are introduced.

Assumption 1. *There exists $0 < \delta < 1/7$ such that $\log(q) = O(n^\delta)$. For the one-sample problem, $n = n_1$. For the two-sample problem, $n = \max(n_1, n_2)$, where the two sample sizes are comparable $n_1 \asymp n_2$.*

Next, we will introduce additional notations to present the assumptions for the kernel functions. Specifically, based on $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_{n_1}\}$, set the centered kernel function and its derivation as follows:

$$\begin{aligned} \Psi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) &= \Phi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) - u_{\gamma,s}, \\ g_s(\mathbf{X}_i) &= \mathbf{E}(\Psi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) | \mathbf{X}_i). \end{aligned} \quad (12)$$

Further, set

$$\begin{aligned}\Psi(X_{i_1}, \dots, X_{i_m}) &= (\Psi_1(X_{i_1}, \dots, X_{i_m}), \dots, \Psi_q(X_{i_1}, \dots, X_{i_m})), \\ \mathbf{g}(X_i) &= (g_1(X_i), \dots, g_q(X_i)).\end{aligned}$$

Analogously, we can construct vectors $\Psi(Y_{i_1}, \dots, Y_{i_m})$ and $\mathbf{g}(Y_i)$ based on $\mathcal{Y} = \{Y_1, \dots, Y_{n_2}\}$.

Assumption 2. For any indexes $0 < i_1, \dots, i_m < n_1$ and $0 < j_1, \dots, j_m < n_2$,

$$\begin{aligned}\max_{1 \leq s \leq q} \mathbb{E}[\exp(|\Psi_s(X_{i_1}, \dots, X_{i_m})|/K)] &\leq 2, \\ \max_{1 \leq s \leq q} \mathbb{E}[\exp(|\Psi_s(Y_{i_1}, \dots, Y_{i_m})|/K)] &\leq 2.\end{aligned}$$

Assumption 3. There exists a positive constant b , such that $\mathbb{E}[|\mathbf{v}^\top \mathbf{g}(X)|^2] \geq b$ and $\mathbb{E}[|\mathbf{v}^\top \mathbf{g}(Y)|^2] \geq b$, for any $\mathbf{v} \in \mathcal{V}_{s_0}$, where $\mathcal{V}_{s_0} = \{\mathbf{v} \in \mathbb{S}^{q-1} : \|\mathbf{v}\|_0 \leq s_0\}$.

Assumption 4. There exists a constant $Q \geq 1$ such that $\mathbb{E}[|g_s(X)|^{2+m}] \leq Q^m \mathbb{E}[|g_s(Y)|^{2+m}] \leq Q^m$ holds for all $1 \leq s \leq q$, and $m = 1, 2$.

This assumption specifies the relationship between the sample size n and the parameter dimension q . Assumption 1 permits the parameter dimension q and sample size n go to infinity as long as $\log(q) = O(n^\delta)$ holds. Additionally, it requires the sample size of each group to be in the same order. Assumption 2 requires that the centered kernel functions $\Psi_s(X_{i_1}, \dots, X_{i_m})$ and $\Psi_s(Y_{i_1}, \dots, Y_{i_m})$ follow a sub-exponential distribution. This assumption is mild, as the bounded Ψ_s (such as useful rank-based U -statistic) satisfies this condition. Assumption 3 excludes the degenerate U -statistics and requires that the inner product of $\mathbf{g}(X)$ $\mathbf{g}(Y)$ and $\mathbf{v} \in \mathcal{V}_{s_0}$ is not degenerate. Assumption 4 include mild moment conditions. These Assumptions are crucial for the high-dimensional central limiting theorem.

The multiplier bootstrap method plays an important role in our tests. Based on the above assumptions, we justify the validity of the multiplier bootstrap procedure used in Section 2 by the following theorem:

Theorem 1. Set $\mathcal{X} = \{X_1, \dots, X_{n_1}\}$ and $\mathcal{Y} = \{Y_1, \dots, Y_{n_2}\}$. Suppose Assumptions 1–4 hold. Under $\mathbf{H}_0$ of (2), we have

$$\sup_{z \in (0, \infty)} \left| \mathbb{P}(T_o \leq z) - \mathbb{P}(T_o^b \leq z | \mathcal{X}) \right| = o_p(1), \quad \text{as } n \rightarrow \infty. \quad (13)$$

Under $\mathbf{H}_0$ of (3),

$$\sup_{z \in (0, \infty)} \left| \mathbb{P}(T_d \leq z) - \mathbb{P}(T_d^b \leq z | \mathcal{X}, \mathcal{Y}) \right| = o_p(1), \quad \text{as } n \rightarrow \infty. \quad (14)$$

The detailed proof of Theorem 1 is presented in Appendix A.3. This theorem ensures that the distribution of our tests can be well approximated by their corresponding multiplier bootstrap's distribution under $\mathbf{H}_0$.

Theorem 2. Suppose Assumptions 1–4 hold. Under $\mathbf{H}_0$ of (2), we have

$$\mathbf{P}_{\mathbf{H}_0}(\Phi_\alpha^o = 1) \rightarrow \alpha \quad \text{and} \quad \text{as } n, B \rightarrow \infty. \quad (15)$$

Under $\mathbf{H}_0$ of (3), we have

$$\mathbf{P}_{\mathbf{H}_0}(\Phi_\alpha^d = 1) \rightarrow \alpha \quad \text{and} \quad \text{as } n, B \rightarrow \infty. \quad (16)$$

Given the pre-specified level α , Theorem 2 establishes that the size of T_o and T_d are well under control. We omit the proof procedure, as this theorem can be viewed as a consequence of Theorem 1. Specifically, according to Theorem 1, the distribution of our tests T_o and T_d can be well approximated by that of T_o^b and T_d^b . Further, by the Dvoretzky–Kiefer–Wolfowitz inequality of the Massart version, we have

$$\sup \left| \widehat{F}_{T_o^b}(z) - F_{T_o^b}(z) \right| \rightarrow 0, \quad \sup \left| \widehat{F}_{T_d^b}(z) - F_{T_d^b}(z) \right| \rightarrow 0 \quad \text{as } n, B \rightarrow \infty,$$

where $\widehat{F}_{T_o^b}(z)$ and $\widehat{F}_{T_d^b}(z)$ are the empirical distributions used to obtain the critical values in the test procedures, while $F_{T_o^b}(z)$ and $F_{T_d^b}(z)$ represent the theoretical distributions in Theorem 1. Thus, Theorem 2 is established by combining these results. In addition to asymptotic size control, we investigate the power properties of T_o and T_d in the subsequent theorem.

Theorem 3. Suppose Assumptions 1–4 hold, and let $\epsilon_n = o(1)$ with $\epsilon_n \sqrt{\log q} \rightarrow \infty$ as $n, q \rightarrow \infty$. Further, assume $\log(q) = o(n^{1/2})$. For one sample test we have

$$\mathbf{P}(\Phi_\alpha^o = 1) \rightarrow 1 \quad \text{as } n, q, B \rightarrow \infty, \quad (17)$$

if $\mathbf{H}_1$ of (2) holds with

$$\max_{1 \leq s \leq q} \left| \frac{u_{1,s} - u_{0,s}}{\sqrt{\sigma_{1,ss}/n_1}} \right| \geq (1 + \epsilon_n) n^{-1/2} \left(\sqrt{2 \log(q)} + \sqrt{2 \log(1/\alpha)} \right). \quad (18)$$

For two sample tests, we have

$$\mathbf{P}(\Phi_\alpha^d = 1) \rightarrow 1 \quad \text{as } n, q, B \rightarrow \infty, \quad (19)$$

if $\mathbf{H}_1$ of (3) holds with

$$\max_{1 \leq s \leq q} \left| \frac{u_{1,s} - u_{2,s}}{\sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}} \right| \geq (1 + \epsilon_n) \left(\sqrt{2 \log(q)} + \sqrt{2 \log(1/\alpha)} \right). \quad (20)$$

The detailed proof for this theorem is provided in Appendix A.4. This result demonstrates that, under certain mild conditions, the power of our proposed test converges to 1.

4. Simulation Study

In this section, we conduct comprehensive simulations to investigate the empirical performance of our proposed test. As discussed in Section 2, for the sample mean test our method reduces to what was introduced by [12,13]. Therefore, we focus on testing the Kendall tau correlation matrix and compare our method with several established methods. Specifically, we consider the sample covariance-based methods from [49,50], as well as the U -statistic-based Kendall tau correlation test from [42]. For simplicity, we refer to these methods as CLX, HD, and ZU, respectively, and denote our method as UJB.

We generate two sets of independent random samples, $\{\mathbf{X}_i\}_{i=1}^{n_1}$ and $\{\mathbf{Y}_i\}_{i=1}^{n_2}$, where $\mathbf{X}_i = \Sigma_1^{*1/2} \mathbf{Z}_i^{(1)}$ and $\mathbf{Y}_i = \Sigma_2^{*1/2} \mathbf{Z}_i^{(2)}$. Here, $\mathbf{Z}_i^{(\gamma)} = (Z_{i1}^{(\gamma)}, \dots, Z_{id}^{(\gamma)})^T$, where $\gamma = 1, 2$, and $Z_{i1}^{(\gamma)}, \dots, Z_{id}^{(\gamma)}$ are independent and identically distributed (i.i.d.) random variables with variances $\sigma_{Z,\gamma}^2$. To mimic practical scenarios, we assume Z_{ij}^γ are samples from the following three models:

- **Model 1 (Gamma Distribution):** Let $Z_{ij}^\gamma \sim \text{Gamma}(4, 10)$, for $j = 1, \dots, d$.
- **Model 2 (Zero-Inflated Poisson Distribution):** Let $Z_{ij}^\gamma \sim \text{Pois}(1000)$ with probability 0.15, and $Z_{ij}^\gamma = 0$ with probability 0.85, for $j = 1, \dots, d$.
- **Model 3 (Student's t Distribution):** Let $Z_{ij}^\gamma \sim t_3(\mu)$, where μ is drawn from $\text{Unif}(-2, 2)$, for $j = 1, \dots, d$.

Thus, the covariance matrix of $\mathbf{X}$ is $\Sigma_1 = \sigma_{Z,1}^2 \Sigma_1^*$, and the covariance matrix of $\mathbf{Y}$ is $\Sigma_2 = \sigma_{Z,2}^2 \Sigma_2^*$. Under the null hypothesis, we assume $\Sigma_1^* = \Sigma_2^* = \Sigma^*$. We consider the following four covariance structures for Σ^* :

- **Case I (Block):** Set $\Sigma^* = \mathbf{D}^{1/2} \mathbf{R} \mathbf{D}^{1/2}$, where $\mathbf{D}$ is a diagonal matrix with i.i.d. entries drawn from $\text{Unif}(0.5, 2.5)$. The matrix $\mathbf{R} = (r_{k\ell})_{1 \leq k, \ell \leq d}$ represents a block correlation structure, with $r_{kk} = 1$ for all k , $r_{k\ell} = 0.55$ for $10(p-1) + 1 \leq k \neq \ell \leq 10p$ (for $p = 1, \dots, \lfloor d/10 \rfloor$), and $r_{k\ell} = 0$ otherwise.
- **Case II (Decay):** Set $\Sigma^* = (\sigma_{k\ell})_{1 \leq k, \ell \leq d}$, where $\sigma_{k\ell} = 0.6^{|k-\ell|}$.
- **Case III (Non-Sparse):** Define $\Sigma^* = \mathbf{D}^{1/2} (\mathbf{F} + \mathbf{U} \mathbf{U}^T) \mathbf{D}^{1/2}$, where $\mathbf{F} = (f_{k\ell})_{1 \leq k, \ell \leq d}$ with $f_{kk} = 1$ and $f_{k,k+1} = f_{\ell+1,\ell} = 0.5$. The matrix $\mathbf{D}$ is diagonal, with entries drawn from $\text{Unif}(1, 6)$, and $\mathbf{U}$ is uniformly distributed on the Stiefel manifold $\mathcal{V}(d, k_0)$, i.e., $\mathbf{U} \in \mathbb{R}^{d \times k_0}$ and $\mathbf{U}^T \mathbf{U} = \mathbf{I}_{k_0}$, where $\mathbf{I}_{k_0}$ is the identity matrix of dimension k_0 (set $k_0 = 10$).
- **Case IV (Long-range dependence):** Set $\Sigma^* = (\sigma_{k\ell})_{1 \leq k, \ell \leq d}$, where $\sigma_{kk} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(1, 2)$, $\sigma_{k\ell} = f(|k - \ell|)$ with $f(x) = \{(x+1)^{1.7} + (x-1)^{1.7} - 2x^{1.7}\}/2$.

Table 1 shows the empirical sizes across Models 1–3 and Cases I–III. Overall, all methods maintain sizes close to the nominal level of $\alpha = 0.05$, which indicates good control of Type I error rates. In Models 1 and 2, the empirical sizes of the CLX and HD methods remain stable around the nominal level. ZU and UJB also perform well, although UJB slightly exceeds the nominal size in some instances. In Model 3, which employs a heavy-tailed t -distribution, the empirical size of UJB continues to align closely with the nominal level. In contrast, the empirical sizes for CLX and HD are more conservative, highlighting the robustness of UJB in the presence of heavy-tailed data.

Table 1. Empirical sizes of Model 1–3 with Case I–IV based on 500 replications, with $\alpha = 0.05$.

		Model 1			Model 2			Model 3		
$n_1 = n_2 = 100$	d	30	40	50	30	40	50	30	40	50
Case I	CLX	0.04	0.03	0.03	0.03	0.05	0.03	0.01	0.01	0.01
	HD	0.05	0.04	0.05	0.04	0.06	0.05	0.03	0.02	0.02
	ZU	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	UJB	0.06	0.07	0.07	0.05	0.06	0.07	0.06	0.06	0.07
Case II	CLX	0.04	0.04	0.05	0.04	0.05	0.05	0.02	0.01	0.01
	HD	0.05	0.05	0.05	0.06	0.05	0.06	0.03	0.02	0.01
	ZU	0.03	0.03	0.04	0.05	0.04	0.04	0.03	0.03	0.04
	UJB	0.06	0.08	0.07	0.09	0.08	0.07	0.08	0.05	0.07
Case III	CLX	0.04	0.03	0.04	0.05	0.04	0.07	0.01	0.01	0.01
	HD	0.04	0.03	0.04	0.06	0.05	0.07	0.01	0.01	0.01
	ZU	0.04	0.05	0.03	0.05	0.04	0.05	0.05	0.03	0.06
	UJB	0.07	0.08	0.07	0.07	0.07	0.08	0.06	0.05	0.07
Case IV	CLX	0.03	0.05	0.04	0.06	0.04	0.04	0.02	0.02	0.02
	HD	0.05	0.05	0.05	0.07	0.06	0.06	0.03	0.02	0.03
	ZU	0.02	0.02	0.03	0.02	0.03	0.03	0.02	0.02	0.02
	UJB	0.07	0.08	0.09	0.09	0.09	0.09	0.07	0.09	0.08

Under the alternative hypothesis, we introduce a random symmetric matrix $\Delta = (\delta_{k\ell}) \in \mathbb{R}^{d \times d}$ with exactly eight nonzero entries. Among these, four entries are randomly selected from

the upper triangle of Δ , with magnitudes generated from the uniform distribution on $(0, \sigma_{\max}^2)$, where $\sigma_{\max}^2$ is the maximum diagonal value of Σ^* . The remaining four entries are determined by symmetry. We then define $\tilde{\Sigma}_1 = \Sigma^* + \delta \mathbf{I}$ and $\tilde{\Sigma}_2 = \Sigma^* + \Delta + \delta \mathbf{I}$, where $\delta = |\min\{\lambda_{\min}(\Sigma^* + \Delta), \lambda_{\min}(\Sigma^*)\}| + 0.05$. These matrices, $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$, are used to generate samples for $\mathbf{X}$ and $\mathbf{Y}$ under the alternative hypothesis.

The empirical power performance across Models 1–3 and Cases I–IV is summarized in Table 2. For Models 1 and 2, which involve Gamma and zero-inflated Poisson distributions, the data quickly approximate normality. This allows the CLX and HD methods to perform comparably to UJB and ZU. However, UJB and the rank-based ZU tests demonstrate a slight advantage. In Model 3, where the data follow a heavy-tailed t -distribution, UJB and ZU significantly outperform CLX and HD, highlighting the effectiveness of rank-based correlation tests for heavy-tailed distributions. Across all cases, UJB maintains comparable or slightly superior performance to ZU, highlighting the strength and robustness of our proposed method.

Table 2. Empirical powers of Model 1–3 with Case I–IV based on 500 replications, with $\alpha = 0.05$.

		Model 1			Model 2			Model 3		
$n_1 = n_2 = 100$	d	30	40	50	30	40	50	30	40	50
Case I	CLX	0.75	0.60	0.52	0.98	0.83	0.57	0.49	0.25	0.09
	HD	0.75	0.62	0.54	0.99	0.83	0.60	0.52	0.26	0.11
	ZU	1.00	0.99	0.86	1.00	0.99	0.87	1.00	0.99	0.73
	UJB	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.98
Case II	CLX	0.62	0.59	0.59	0.77	0.71	0.67	0.30	0.18	0.09
	HD	0.65	0.61	0.61	0.79	0.73	0.68	0.34	0.20	0.10
	ZU	0.78	0.78	0.61	0.84	0.79	0.62	0.75	0.56	0.51
	UJB	1.00	0.99	0.97	1.00	1.00	0.95	0.98	0.97	0.87
Case III	CLX	0.98	1.00	0.98	1.00	1.00	1.00	0.71	0.61	0.55
	HD	0.98	1.00	0.98	1.00	1.00	1.00	0.73	0.63	0.57
	ZU	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	UJB	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Case IV	CLX	0.91	0.36	0.62	0.95	0.43	0.75	0.42	0.29	0.19
	HD	0.92	0.39	0.64	0.95	0.45	0.76	0.47	0.32	0.21
	ZU	1.00	0.81	0.97	1.00	0.89	0.98	1.00	0.59	0.92
	UJB	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

5. Real Data Analysis

In this section, we apply our proposed method to explore potential dependency differences within the Wnt signaling pathway, which is associated with lung cancer as well as other cancers, including gastric and breast cancer [51–53]. The dataset used for this analysis is publicly available through the Gene Expression Omnibus (GEO) repository (<https://www.ncbi.nlm.nih.gov/geo/>, accessed on 25 November 2024) under accession number GDS2771. It contains 22,283 microarray-derived gene expression measurements from large airway epithelial cells in 97 patients diagnosed with lung cancer and 90 control patients without lung cancer.

In this study, we focus on a subset of 119 genes within the Wnt signaling pathway. As the number of potential dependencies grows quadratically with the number of genes, reliable statistical inference becomes challenging with a limited sample size. To address this, we apply individual t -tests to each gene to assess differential expression between the two groups. This conservative approach reduces dimensionality while retaining potentially relevant genes for further analysis. Ultimately, we select 32 genes with statistically significant expression differences for the dependency analysis. This selected subset includes genes previously identified as important in lung cancer research, such as WNT1, WNT2, WNT5A [51], FDZ [54], and RHOA [55]. Based on this subset, we use the testing methods

outlined in Section 4 to test for the equality in the correlation matrices of these 32 genes between lung cancer patients and control patients. Given that the CLX and HD methods are conservative for heavy-tailed data, we transformed the gene expression levels by applying a logarithmic scale and then standardized each gene within its respective group.

All methods reject the null hypothesis, providing strong evidence against equal dependency structures. Specifically, the p -values for our proposed method (UJB) and the comparison methods (CLX and HD) are 0.019, 0.046, and 0.029, respectively. This finding suggests that the dependency structure within the Wnt signaling pathway is likely distinct between lung cancer patients and controls. Thus, beyond differential gene expression, changes in dependency structures among genes in the Wnt pathway may offer new insights into the mechanisms underlying lung cancer progression. For example, ref. [56] shows that WNT5A–RHOA signaling drives tumorigenesis and represents a therapeutic target in small-cell lung cancer. It also highlights the importance of gene dependency structures in the Wnt pathway for understanding lung cancer progression.

6. Discussion

In this paper, we introduce a novel approach that reformulates the testing of U -statistic-type parameters into a sample mean testing problem by using jackknife pseudo-values. This transformation allows us to apply established methods for sample mean testing, simplifying the test process while maintaining the flexibility and power inherent to U -statistic-based inference. Moreover, we obtain critical values for the test by using multiplier bootstrap and ensuring the validity of this process.

Our simulation study involving samples from various distributions with different covariance structures, demonstrates the robustness and adaptability of the proposed method. Theoretical analysis further confirms the validity of our approach. However, a notable limitation lies in the computational cost of the jackknife procedure, which increases with the sample size. This can become a significant challenge as the number of samples grows too large. Future work could explore strategies to improve computational efficiency or alternative approaches that preserve the statistical accuracy of the test while reducing its computational burden.

In addition to addressing computational challenges, another possible work direction lies in addressing size distortions. Some empirical sizes in Table 1 deviate from the nominal significance level of 5%, particularly for Model 3. To improve the empirical performance, future work could incorporate methods like simulation-based calibration to better approximate null distributions [57], imputation-based techniques for enhanced accuracy [58], or adjustments to test statistics to address structural dependencies [59].

Author Contributions: Methodology, M.Z.; validation, L.J.; writing—original draft preparation, M.Z.; writing—review and editing, L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 12101412) to M.Z., and the National Philosophy and Social Science Foundation of China (No. 23BTJ046) to L.J.

Data Availability Statement: The real data can be obtained from the Gene Expression Omnibus (GEO) repository (<https://www.ncbi.nlm.nih.gov/geo/>, accessed on 25 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of the Main Theorem

Appendix A.1. Useful Lemmas

To establish the proof of the main results, some useful lemmas are required. Firstly, some additional notations are introduced. Let $X_1, \dots, X_n$ be independent centered random vectors in $\mathbb{R}^d$ with the covariance matrix Σ_X , and $X_i = (X_{i,1}, \dots, X_{i,d})$ with $i = 1, \dots, n$.

Set $\mathbf{Z}_i (i = 1, \dots, n)$ to be independent Gaussian random vectors in $\mathbb{R}^d$ with the same mean vector and covariance matrix as $\mathbf{X}_i$. Further, set

$$M_k = \max_{1 \leq j \leq d} (\mathbf{E}[|X_{i,j}^k|]^{1/k}).$$

Define $u_X(\zeta)$ as the infimum over all numbers u such that

$$\mathbf{P}\left(|X_{i,j}| \leq u(\mathbf{E}[X_{i,j}^2])^{1/2}, 1 \leq i \leq n, 1 \leq j \leq d\right) \geq 1 - \zeta.$$

Similarly, $u_Z(\zeta)$ are the corresponding quantities for the analog Gaussian case, namely with $(X_i)_{i=1}^n$ is replaced by $(Z_i)_{i=1}^n$ in the above definition. According to [48], the following Kolmogorov–Smirnov distance

$$\rho = \sup_{z \in \mathbb{R}} \left| \mathbf{P}\left(\left\|n^{-1/2} \sum_{i=1}^n \mathbf{X}_i\right\|_{\infty} \leq z\right) - \mathbf{P}\left(\left\|n^{-1/2} \sum_{i=1}^n \mathbf{Z}_i\right\|_{\infty} \leq z\right) \right|,$$

can be bounded as follows.

Lemma A1 (Gaussian approximation, Theorem 2.2 of [48]). *Suppose there exist some positive constants $c_1 < c_2 < \infty$ such that $c_1 \leq \sigma_{X,k,k} \leq c_2$ for all $1 \leq k \leq d$. For any $0 < \zeta < 1$, we have*

$$\rho \leq C \left\{ n^{-1/8} (M_3^{3/4} \vee M_4^{1/2}) (\log(dn/\zeta))^{7/8} + n^{-1/2} (\log(dn/\zeta))^{3/2} u(\zeta) + \zeta \right\}, \quad (\text{A1})$$

where C is a positive constant that only depends on c_1 and c_2 , and $u(t) = \max\{u_X(t), u_G(t)\}$.

For all $i \leq n, j \leq d$, set $\mathbf{P}(|X_{i,j}| > s) \leq \exp(-c_1 s)$. Then, $u(\zeta)$ in (A1) can be bounded by $u(\zeta) \leq \log(dn/\zeta)$. Hence, by taking $\zeta = n^{-1/2}$, we have the bound for ρ as follows

$$\rho = O_p \left\{ n^{-1/8} (\log(dn))^{7/8} + n^{-1/2} (\log(dn))^{5/2} \right\}. \quad (\text{A2})$$

In addition, suppose $\mathbf{X}_1$ and $\mathbf{X}_2$ are two centered Gaussian random vectors in $\mathbb{R}^d$, with the covariance matrices as Σ_1 and Σ_2 , respectively. Ref. [48] proposed to bound the Kolmogorov–Smirnov distance of the maximum of these two different Gaussian distributions by the following lemma.

Lemma A2 (Comparison of the distributions of Gaussian maxima, Lemma 3.1 of [48]). *Suppose all the diagonal elements of Σ_1 are bounded away from 0 and ∞ . Then, we have*

$$\sup_{x \in \mathbb{R}} \left| \mathbf{P}(\|\mathbf{X}_1\|_{\infty} \leq x) - \mathbf{P}(\|\mathbf{X}_2\|_{\infty} \leq x) \right| \leq C \Delta^{1/3} (1 \vee \log(d/\Delta))^{2/3},$$

where $\Delta = \|\Sigma_1 - \Sigma_2\|_{\infty}$, C is a positive constant.

Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be independent random vectors, $\Phi(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m})$ is a kernel function of $m (m < n)$ order. Define the corresponding U -statistic as follows:

$$U = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} \Phi(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}).$$

Lemma A3 (Hoeffding, 1963). *If the kernel function $\Phi(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m})$ is bounded,*

$$\mathbf{P}(U - \mathbf{E}(U) \geq t) \leq \exp(-2kt^2 / (b - a)^2),$$

where a and b are the lower and upper bound of $\Phi(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m})$, respectively, and $k = \lfloor n/m \rfloor$.

Lemma A4. $\mathbf{Z} = (Z_1, \dots, Z_d)^\top$ is a random vector with marginal distribution $N(0, \sigma^2)$. For any $t > 0$, we have

$$\mathbb{E} \left[\max_{1 \leq j \leq d} |Z_j| \right] \leq \frac{\log(2d)}{t} + \frac{t\sigma^2}{2}.$$

By $\Phi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m})$ is the m -order symmetric kernel function, and combining the definition in (12), we set the covariance matrices of $mg(\mathbf{X}_i)$ and $mg(\mathbf{Y}_i)$ are $\Sigma_1 = (\sigma_{1,st})$ and $\Sigma_2 = (\sigma_{2,st})$, respectively. Further, the corresponding correlation matrices are defined as follows:

$$\mathbf{R}_\gamma = \text{Diag}(\Sigma_\gamma)^{-1/2} \Sigma_\gamma \text{Diag}(\Sigma_\gamma)^{-1/2}, \quad \gamma = 1, 2.$$

Specifically, the (s, t) -th entry of $\mathbf{R}_\gamma$ is given by the following:

$$r_{\gamma,st} = \frac{\sigma_{\gamma,st}}{\sqrt{\sigma_{\gamma,ss}\sigma_{\gamma,tt}}}, \quad 1 \leq s, t \leq q.$$

The following lemma provides an upper bound for the estimation errors of the sample covariance matrix $\hat{\Sigma}_\gamma = (\hat{\sigma}_{\gamma,st})$ and its correlation matrix $\hat{\mathbf{R}}_\gamma = (\hat{r}_{\gamma,st})$.

Lemma A5. Suppose the Assumptions 1–4 hold. When d, T are sufficiently large, we have that

$$\max_{\gamma=1,2} \left(\|\hat{\Sigma}_\gamma - \Sigma_\gamma\|_\infty, \|\hat{\mathbf{R}}_\gamma - \mathbf{R}_\gamma\|_\infty \right) = O_p \left(\sqrt{\frac{\log^3(qn)}{n}} \right). \quad (\text{A3})$$

Appendix A.2. Proof of Lemma A5

Proof. To prove Lemma A5, it suffices to show that the upper bound in (A3) holds for both $\gamma = 1$ and $\gamma = 2$ individually. According to similar definitions and assumptions, it is easy to address the upper bound with similar arguments. Without loss of generality, we will show $\gamma = 1$. According to the definition,

$$\Sigma_1 = m^2 \mathbb{E} \left[(\mathbf{g}(\mathbf{X}_i))^\top \mathbf{g}(\mathbf{X}_i) \right].$$

By the definition of $\|\hat{\Sigma}_1 - \Sigma_1\|_\infty$, we need to bound the following equation,

$$\begin{aligned} \max_{1 \leq s, t \leq q} |\hat{\sigma}_{1,st} - \sigma_{1,st}| &= \max_{1 \leq s, t \leq q} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[(\hat{u}_{1,is} - \hat{u}_{1,s})(\hat{u}_{1,it} - \hat{u}_{1,t}) - m^2 \mathbb{E}(g_s(\mathbf{X}_i)g_t(\mathbf{X}_i)) \right] \right| \\ &\leq \underbrace{\max_{1 \leq s, t \leq q} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[(\hat{u}_{1,is} - \hat{u}_{1,s})(\hat{u}_{1,it} - \hat{u}_{1,t}) - m^2 \frac{1}{n_1} \sum_{i=1}^{n_1} (g_s(\mathbf{X}_i)g_t(\mathbf{X}_i)) \right] \right|}_{L_1} \\ &\quad + \underbrace{\max_{1 \leq s, t \leq q} \left| m^2 \frac{1}{n_1} \sum_{i=1}^{n_1} (g_s(\mathbf{X}_i)g_t(\mathbf{X}_i) - m^2 \mathbb{E}(g_s(\mathbf{X}_i)g_t(\mathbf{X}_i))) \right|}_{L_2}. \end{aligned}$$

By Theorem 6 in [60], and the sub-exponential assumptions for $g_s(\mathbf{X}_i)$, it is easy to obtain the following upper bound for L_2 :

$$L_2 \leq Cm^2 \sqrt{\frac{\log(qn_1)}{n_1}} + C_1 m^2 \frac{\log^2(qn_1)}{n_1}. \quad (\text{A4})$$

For L_1 ,

$$L_1 = \underbrace{\max_{1 \leq s, t \leq q} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[(\hat{u}_{1, is} - u_{1, s})(\hat{u}_{1, it} - u_{1, t}) \right] - m^2 \frac{1}{n_1} \sum_{i=1}^{n_1} \left[g_s(\mathbf{X}_i) g_t(\mathbf{X}_i) \right] \right|}_{L_{11}} + \underbrace{\max_{1 \leq s, t \leq q} \left| (\hat{u}_{1s} - u_{1s})(\hat{u}_{1t} - u_{1t}) \right|}_{L_{12}}.$$

To bound L_1 , we bound L_{11} and L_{12} , respectively. For L_{12} , it is obvious that

$$\begin{aligned} \mathbf{P}(L_{12} > z) &\leq q^2 \max_{1 \leq s, t \leq q} \mathbf{P}(|(\hat{u}_{1s} - u_{1s})(\hat{u}_{1t} - u_{1t})| > z) \\ &\leq 2q^2 \max_{1 \leq s \leq q} \mathbf{P}(|\hat{u}_{1s} - u_{1s}| > \sqrt{z}). \end{aligned} \quad (\text{A5})$$

Hence, the key to bound L_{12} is to study $\max_{1 \leq s \leq q} \mathbf{P}(|\hat{u}_{1s} - u_{1s}| > \sqrt{z})$. According to the definition of $\hat{u}_{1s}$ in (4), and with some simple calculations, we have the following:

$$\hat{u}_{1s} - u_{1s} = \tilde{u}_{1s} - u_{1s}. \quad (\text{A6})$$

Next, we introduce

$$\tilde{u}'_{1s} - u_{1s} = \left(\frac{n_1}{m} \right)^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n_1} V_{1s}^{i_1, \dots, i_m} - E_{1s}, \quad (\text{A7})$$

where

$$\begin{aligned} V_{1s}^{i_1, \dots, i_m} &= \Psi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) \mathbb{I}\{|\Psi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m})| \leq B_n\}, \\ E_{1s} &= \mathbf{E}(\Psi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) \mathbb{I}\{|\Psi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m})| \leq B_n\}), \end{aligned} \quad (\text{A8})$$

with threshold $B_n = C \log(qn)$. By the triangle inequality, we have

$$|\tilde{u}_{1s} - u_{1s}| \leq |\tilde{u}_{1s} - \tilde{u}'_{1s}| + |\tilde{u}'_{1s} - u_{1s}|.$$

According to the definition of $\tilde{u}'_{1s} - u_{1s}$ in (A7), and by the triangle inequality, we have

$$|\tilde{u}_{1s} - \tilde{u}'_{1s}| \leq \underbrace{\left| \tilde{u}_{1s} - u_{1s} - \left(\frac{n_1}{m} \right)^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n_1} V_{1s}^{i_1, \dots, i_m} \right|}_{L_{2s}} + E_{1s}.$$

For any $\delta > 0$, by choosing proper C , we have $E_{1s} \prec (qn)^{-\delta}$. By setting $z \succ (qn)^{-\delta}$, we have

$$\max_{1 \leq s \leq q} \mathbf{P}(|\tilde{u}_{1s} - u_{1s}| > z) \leq \max_{1 \leq s \leq q} (\mathbf{P}(|\tilde{u}'_{1s} - u_{1s}| > z/3) + \mathbf{P}(L_{2s} > z/3)).$$

Using the exponential inequality for bounded U -statistics, we have

$$\max_{1 \leq s \leq q} \mathbf{P}(|\tilde{u}'_{1s} - u_{1s}| > z/3) \leq C \exp\left(-C_1 n_1 z^2 / B_n^2\right).$$

By Assumption 2, we also have

$$\max_{1 \leq s \leq q} \mathbf{P}(L_{2s} > z/3) \leq C n_1^m \exp(-C_1 B_n).$$

Combining these results, we have

$$\max_{1 \leq s \leq q} \mathbf{P}(|\tilde{u}_{1s} - u_{1s}| > z) \leq C \exp(-C_1 n_1 z^2 / B_n^2) + C n_1^m \exp(-C_1 B_n). \quad (\text{A9})$$

Therefore, for a sufficiently large n_1 , we have

$$L_{12} \leq \log^3(qn_1)/n_1, \quad (\text{A10})$$

with probability $1 - Cn_1^{-1}$.

For L_{11} , using the definition of $\hat{u}_{1,i,s} - u_{1s}$, we simplify the expression by applying the following Hoeffding decomposition. Specifically,

$$\tilde{u}_{1s} - u_{1s} = \frac{m}{n_1} \sum_{k=1}^n g(X_k) + \binom{n_1}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n_1} \varphi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}),$$

where $\varphi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) = \Phi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) - u_{1s} - \sum_{\ell=1}^n g(X_\ell)$, $s = 1, \dots, q$. Similarly, for $\tilde{u}_{1s}^{-i} - u_{1s}$, we have

$$\tilde{u}_{1s}^{-i} - u_{1s} = \frac{m}{n_1 - 1} \sum_{\substack{\ell=1 \\ \ell \neq i}}^n g(X_\ell) + \binom{n_1 - 1}{m}^{-1} \sum_{\substack{1 \leq i_1 < \dots < i_m \leq n_1 \\ i_j \neq i, j=1, \dots, m}} \varphi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}).$$

Set

$$\begin{aligned} \Delta_{1,i,s} = & n_1 \binom{n_1}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n_1} \varphi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}) - \\ & (n_1 - 1) \binom{n_1 - 1}{m}^{-1} \sum_{\substack{1 \leq i_1 < \dots < i_m \leq n_1 \\ i_j \neq i, j=1, \dots, m}} \varphi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m}). \end{aligned}$$

According to the definition of $\hat{u}_{1,i,s} - u_{1s}$, we have

$$\hat{u}_{1,i,s} - u_{1s} = m g_s(\mathbf{X}_i) + \Delta_{1,i,s}, \text{ for } s = 1, \dots, q.$$

Hence, plugging these into L_{11} , and performing some calculations, we obtain

$$L_{11} = \max_{1 \leq s, t \leq q} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} [m g_s(\mathbf{X}_i) \Delta_{1,i,t} + m g_t(\mathbf{X}_i) \Delta_{1,i,s} + \Delta_{1,i,s} \Delta_{1,i,t}] \right|.$$

By triangle inequality and Cauchy–Swartz inequality, we have

$$L_{11} \leq 2 \underbrace{\sqrt{\max_{1 \leq s \leq q} \frac{1}{n_1} \sum_{i=1}^{n_1} (m g_s(\mathbf{X}_i))^2}}_{L_{11,1}} \sqrt{\max_{1 \leq t \leq q} \frac{1}{n_1} \sum_{i=1}^{n_1} \Delta_{1,i,t}^2} + \underbrace{\max_{1 \leq t \leq q} \frac{1}{n_1} \sum_{i=1}^{n_1} \Delta_{1,i,t}^2}_{L_{11,2}}.$$

Next, we bound $L_{11,1}$ and $L_{11,2}$, respectively. By Assumption 4, it's obvious that $L_{11,1}$ bounded. For $L_{11,2}$,

$$\begin{aligned} \Delta_{1,i,s} = & m \underbrace{\left(\frac{n_1 - 1}{m - 1} \right)^{-1} \sum_{\substack{1 \leq i_1 < \dots < i_{m-1} \leq n_1 \\ i_j \neq i, j=1, \dots, m-1}} \varphi_s(\mathbf{X}_i, \mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_{m-1}})}_{\Delta_{11,i,s}} \\ & + (m - 1) \underbrace{\left(\frac{n_1 - 1}{m} \right)^{-1} \sum_{\substack{1 \leq i_1 < \dots < i_m \leq (n_1 - 1) \\ i_j \neq i, j=1, \dots, m}} \varphi_s(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_m})}_{\Delta_{12,i,s}}. \end{aligned}$$

Given $\mathbf{X}_i$, $\varphi_s(\mathbf{X}_i, \mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_{m-1}})$ can be treated as a symmetric kernel function of zero mean and $m - 1$ order, and $\Delta_{11,i,s}$ is a U -statistic. Analogously, $\Delta_{12,i,s}$ is a U -statistic with a kernel function of zero mean and m order. Using the comparable technique employed to bound L_{12} , we can introduce a threshold kernel and exponential inequality for U -statistic. Thus, given $\mathbf{X}_i$, for sufficiently large n_1 , we have

$$\begin{aligned} \max_{1 \leq s \leq q} \Delta_{11,i,s}^2 & \leq C \frac{\log^3(qn_1)}{n_1}, \quad \max_{1 \leq s \leq q} \Delta_{12,i,s}^2 \leq C \frac{\log^3(qn_1)}{n_1}, \\ \max_{1 \leq s \leq q} \Delta_{11,i,s} \Delta_{12,i,s} & \leq C \frac{\log^3(qn_1)}{n_1}, \end{aligned}$$

with probability $1 - Cn_1^{-1}$. Hence, $L_{11,2} \leq C \log^3(qn_1)/n_1$ with probability $1 - Cn_1^{-1}$. Furthermore, for sufficiently large n_1 ,

$$L_{11} \leq C \sqrt{\frac{\log^3(qn_1)}{n_1}}, \quad (\text{A11})$$

with probability $1 - Cn_1^{-1}$. Combining the bound of L_{12} in (A10), it is straightforward to obtain

$$L_1 \leq C \sqrt{\frac{\log^3(qn_1)}{n_1}}. \quad (\text{A12})$$

Therefore, combining the bound of L_2 in (A4), we have

$$\max_{1 \leq s, t \leq q} |\hat{\sigma}_{1,st} - \sigma_{1,st}| = O_p \left(\sqrt{\frac{\log^3(qn_1)}{n_1}} \right). \quad (\text{A13})$$

Next, we show the bound for $\max_{1 \leq s, t \leq q} |\hat{r}_{1,st} - r_{1,st}|$. According to the definition of $r_{1,st}$ and the triangle inequality, we have

$$\begin{aligned} |\hat{r}_{1,st} - r_{1,st}| & = \left| \frac{\hat{\sigma}_{1,st}}{\sqrt{\hat{\sigma}_{1,ss}\hat{\sigma}_{1,tt}}} - \frac{\sigma_{1,st}}{\sqrt{\sigma_{1,ss}\sigma_{1,tt}}} \right| \\ & \leq \underbrace{\left| \frac{\hat{\sigma}_{1,st}}{\sqrt{\hat{\sigma}_{1,ss}\hat{\sigma}_{1,tt}}} - \frac{\hat{\sigma}_{1,st}}{\sqrt{\sigma_{1,ss}\sigma_{1,tt}}} \right|}_{A_1} + \underbrace{\left| \frac{\hat{\sigma}_{1,st}}{\sqrt{\sigma_{1,ss}\sigma_{1,tt}}} - \frac{\sigma_{1,st}}{\sqrt{\sigma_{1,ss}\sigma_{1,tt}}} \right|}_{A_2}. \end{aligned}$$

Hence, we need to bound A_1 and A_2 , respectively. For A_1 , considering $|\hat{r}_{1,st}| \leq 1$, we have

$$A_1 = \left| \frac{\hat{\sigma}_{1,st}}{\sqrt{\hat{\sigma}_{1,ss}\hat{\sigma}_{1,tt}}} \right| \left| 1 - \frac{\sqrt{\hat{\sigma}_{1,ss}\hat{\sigma}_{1,tt}}}{\sqrt{\sigma_{1,ss}\sigma_{1,tt}}} \right| \leq \frac{|\hat{\sigma}_{1,ss}\hat{\sigma}_{1,tt} - \sigma_{1,ss}\sigma_{1,tt}|}{\sigma_{1,ss}\sigma_{1,tt}}.$$

By Assumptions 3 and 4, there are constants b and B , such that $0 < b \leq \sigma_{1,ss} \leq B < \infty$ for $s = 1, \dots, q$. Hence, we have

$$A_1 \leq b^{-2} \max_{1 \leq s \leq q} |\hat{\sigma}_{1,ss} - \sigma_{1,ss}|^2 + 2Bb^{-2} \max_{1 \leq s \leq q} |\hat{\sigma}_{1,ss} - \sigma_{1,ss}|.$$

For A_2 , we have

$$A_2 \leq b^{-1} \max_{1 \leq s, t \leq q} |\hat{\sigma}_{1,st} - \sigma_{1,st}|.$$

Therefore, we have

$$\max_{1 \leq s, t \leq q} |\hat{r}_{1,st} - r_{1,st}| = O_p \left(\sqrt{\frac{\log^3(qn_1)}{n_1}} \right). \quad (\text{A14})$$

By combining the bounding results from (A13) and (A14), we complete the proof of Lemma A5. $\square$

Appendix A.3. Proof of Theorem 1

Proof. The proof procedures for (13) and (14) are similar. Without loss of generality, we provide the specific proof process for (14).

According to the definition of T_d in (9), we introduce an oracle test statistic with known variances

$$\tilde{T}_d = \max_{1 \leq s \leq q} \frac{\hat{u}_{1,s} - \hat{u}_{2,s}}{\sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}}.$$

As the jackknife pseudo-values are asymptotically independent, and by Lemma A1, we have

$$\sup_{z \in (0, \infty)} \left| \mathbf{P}(\tilde{T}_d \leq z) - \mathbf{P}(\|G\|_\infty \leq z) \right| = o_p(1), \quad \text{as } n \rightarrow \infty, \quad (\text{A15})$$

where G is a Gaussian distribution random vector defined as $G \sim N_q(\mathbf{0}, \mathbf{R}_{12})$ with $\mathbf{R}_{12} = \text{Diag}(\boldsymbol{\Sigma}_{12})^{-1/2} \boldsymbol{\Sigma}_{12} \text{Diag}(\boldsymbol{\Sigma}_{12})^{-1/2}$. Here, $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_1/n_1 + \boldsymbol{\Sigma}_2/n_2$, where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the covariance matrices of $mg(\mathbf{X}_i)$ and $mg(\mathbf{Y}_i)$, respectively.

$$T_d \leq \tilde{T}_d + \underbrace{\max_{1 \leq s \leq q} \left| \frac{\hat{u}_{1,s} - \hat{u}_{2,s}}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}} - \frac{\hat{u}_{1,s} - \hat{u}_{2,s}}{\sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}} \right|}_L, \quad (\text{A16})$$

$$L \leq \underbrace{\max_{1 \leq s \leq q} \left| \frac{\hat{u}_{1,s} - \hat{u}_{2,s}}{\sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}} \right|}_{L_1} \underbrace{\max_{1 \leq s \leq q} \left| \frac{\sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}} - 1 \right|}_{L_2}.$$

For $y \in (0, 1]$, with a simple calculation, we have

$$\mathbf{P}(L_2 \leq y) \geq \mathbf{P} \left(\max_{1 \leq s \leq q} \left| \frac{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}}{\sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}} - 1 \right| \leq y/2 \right). \quad (\text{A17})$$

By Assumptions 3 and 4, there exist positive constants b and B , such that $0 < b \leq \sigma_{\gamma,ss} \leq B < \infty$. According to Assumption 1, the sample size n_1 and n_2 is in the same order. Thus,

$$\max_{1 \leq s \leq q} \left| \frac{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2} - 1 \right| \leq \max_{1 \leq s \leq q} |\sigma_{1,ss} - \hat{\sigma}_{1,ss}| + \max_{1 \leq s \leq q} |\sigma_{2,ss} - \hat{\sigma}_{2,ss}|.$$

Further, by Lemma A5 we have

$$\max_{1 \leq s \leq q} \left| \frac{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2} - 1 \right| = O_p \left(\sqrt{\frac{\log^3(qn)}{n}} \right). \quad (\text{A18})$$

Hence, we have $L_2 = O_p \left(\sqrt{\log^3(qn)/n} \right)$. For L_1 , there are similar arguments for L_{12} in Lemma A5. By Hoeffding inequality, we have

$$\mathbf{P}(L_1 > \varepsilon) \leq Cq \exp(-C_1 n \varepsilon^2).$$

Hence, by combining (A16), we have

$$\sup_{z \in (0, \infty)} \left| \mathbf{P}(T_d \leq z) - \mathbf{P}(\tilde{T}_d \leq z) \right| = o_p(1), \quad \text{as } n \rightarrow \infty. \quad (\text{A19})$$

As $\xi_{\gamma,1}^b, \xi_{\gamma,2}^b, \dots, \xi_{\gamma,n_\gamma}^b$ is a sequence of iid $N(0, 1)$, and according to the definition of $\hat{u}_{\gamma,s}^b$ in (10), given $\mathcal{X}$ and $\mathcal{Y}$, we obtain $\hat{u}_\gamma^b = \left(\hat{u}_{\gamma,1}^b, \dots, \hat{u}_{\gamma,q}^b \right)^\top \sim N(u_\gamma, \hat{\Sigma}_\gamma/n_\gamma)$. By Lemma A1, we have

$$\sup_{z \in (0, \infty)} \left| \mathbf{P}(T_d^b \leq z | \{\mathcal{X}, \mathcal{Y}\}) - \mathbf{P}(\|G\|_\infty \leq z) \right| = o_p(1), \quad \text{as } n \rightarrow \infty. \quad (\text{A20})$$

By triangle inequality and combining the results of (A15), (A19) and (A20), we have

$$\sup_{z \in (0, \infty)} \left| \mathbf{P}(T_d \leq z) - \mathbf{P}(T_d^b \leq z | \{\mathcal{X}, \mathcal{Y}\}) \right| = o_p(1), \quad \text{as } n \rightarrow \infty. \quad (\text{A21})$$

Therefore, Theorem 1 is proved. $\square$

Appendix A.4. Proof of Theorem 3

Proof. The proof procedures for (17) and (19) are similar. Without loss of generality, we outline the specific proof process for (19). Following the approach in Theorem 2, we define the oracle critical value t_α^d from the theoretical distribution $F_{T_d^b}(z)$,

$$t_\alpha^d = \inf \left\{ t \in \mathbb{R} : \mathbf{P}(T_d \leq t) > 1 - \alpha \right\}.$$

The critical value $\hat{t}_\alpha^d$ used in the test Φ_α^d serves as the bootstrap estimator of t_α^d . As $B \rightarrow \infty$, we obtain

$$\lim_{n, B \rightarrow \infty} \left| \mathbf{P}(T_d > t_\alpha^d) - \mathbf{P}(T_d > \hat{t}_\alpha^d) \right| = o_p(1).$$

Therefore, to prove Theorem 3, it suffices to show that the lower bound of $\mathbf{P}(T_d^b \leq t)$ approaches 1 as $n \rightarrow \infty$.

First, we establish the lower bound for t_α^d . As demonstrated in the proof for Theorem 1, given $\mathcal{X}$ and $\mathcal{Y}$, we find that $(\hat{u}_{1,s}^b - \hat{u}_{2,s}^b) / \sqrt{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2}$ follows the standard normal distribution. According to Lemma A4, by setting $\sigma = 1$ and $t = \sqrt{2 \log q}$, we have

$$\mathbf{E}[T_d^b | \mathcal{X}, \mathcal{Y}] \leq \sqrt{\log(2q)} (1 + (\log(2q))^{-1}).$$

By Theorem 5.8 of [61], it follows that

$$\mathbf{P}(T_d^b \geq \mathbf{E}[T_d^b | \mathcal{X}, \mathcal{Y}] + z | \mathcal{X}, \mathcal{Y}) \leq \exp(-z^2/2).$$

Thus, we have

$$t_\alpha^d \leq \sqrt{\log(2q)} (1 + (\log(2q))^{-1}) + \sqrt{2 \log(1/\alpha)}. \quad (\text{A22})$$

Consequently, we have

$$\mathbf{P}(T_d > t_\alpha^d) \geq \underbrace{\mathbf{P}(T_d > \sqrt{\log(2q)} (1 + (\log(2q))^{-1}) + \sqrt{2 \log(1/\alpha)})}_L. \quad (\text{A23})$$

Next, we focus on establishing the lower bound for L . Under $\mathbf{H}_1$, there exists some $s \in \{1, \dots, q\}$, $u_{1,s} \neq u_{2,s}$. Set,

$$T_d' = \max_{1 \leq s \leq q} \frac{|\hat{u}_{1,s} - \hat{u}_{2,s} - u_{1,s} + u_{2,s}|}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}}.$$

By the triangle inequality we have

$$T_d \geq \max_{1 \leq s \leq q} \frac{|u_{1,s} - u_{2,s}|}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}} - T_d'.$$

Define the subset

$$\mathcal{E}_0(x) = \left\{ \max_{1 \leq s \leq q} \left| \frac{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2} - 1 \right| \leq x \right\}.$$

According to Lemma A5 and (A18), with a probability of at least $1 - Cn^{-1}$, we have

$$\max_{1 \leq s \leq q} \left| \frac{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}{\sigma_{1,ss}/n_1 + \sigma_{2,ss}/n_2} - 1 \right| \leq C \sqrt{\frac{\log^3(qn)}{n}}.$$

Thus, we have $\mathbb{P}(\mathcal{E}_0(x)^c) < n^{-1}$. Under $\mathcal{E}_0(x)$, we have

$$T_d \geq \underbrace{\frac{1}{1+x} \max_{1 \leq s \leq q} \frac{|u_{1,s} - u_{2,s}|}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}}}_{L_1} - T_d'.$$

Consequently, we have

$$L \geq \mathbf{P}\left(L_1 > (1 + (\log(q))^{-1}) \left(\sqrt{2 \log(q)} + \sqrt{2 \log(1/\alpha)} \right), \mathcal{E}_0(x)\right).$$

Considering the signal size requirements in (20), by choosing z satisfying $(1 + \varepsilon_n) = (1 + x)(1 + z + \{2 \log q\}^{-1})$, we have

$$L \geq \mathbf{P}\left(T'_d < z\sqrt{2 \log(q)}\right).$$

By the triangle inequality,

$$\mathbf{P}\left(T'_d > z\sqrt{2 \log(q)}\right) \leq \max_{1 \leq s \leq q} \mathbf{P}\left(\frac{|\hat{u}_{1,s} - \hat{u}_{2,s} - u_{1,s} + u_{2,s}|}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}} > z\sqrt{2 \log(q)}\right). \quad (\text{A24})$$

Thus, we have

$$L \geq 1 - \underbrace{\max_{1 \leq s \leq q} \mathbf{P}\left(\frac{|\hat{u}_{1,s} - \hat{u}_{2,s} - u_{1,s} + u_{2,s}|}{\sqrt{\hat{\sigma}_{1,ss}/n_1 + \hat{\sigma}_{2,ss}/n_2}} > z\sqrt{2 \log(q)}\right)}_{L_2}.$$

With similar arguments to obtain the upper bound for L_{12} in the proof of Lemma A5, we have $L_2 \rightarrow 0$ as $n \rightarrow \infty$. Further, we have $L \rightarrow 1$ with a probability of 1 as $n, q, B \rightarrow \infty$, i.e., $\mathbf{P}(\Phi_\alpha^d = 1) \rightarrow 1$, Theorem 3 is proved. $\square$

References

1. Hu, R.; Qiu, X.; Glazko, G.; Klebanov, L.; Yakovlev, A. Detecting Intergene Correlation Changes in Microarray Analysis: A New Approach to Gene Selection. *BMC Bioinform.* **2009**, *10*, 20.
2. Hu, R.; Qiu, X.; Glazko, G. A New Gene Selection Procedure Based on the Covariance Distance. *Bioinformatics* **2010**, *26*, 348–354.
3. Shaw, P.; Greenstein, D.; Lerch, J.; Clasen, L.; Lenroot, R.; Gogtay, N.; Evans, A.; Rapoport, J.; Giedd, J. Intellectual Ability and Cortical Development in Children and Adolescents. *Nature* **2006**, *440*, 676–679.
4. Shedden, K.; Chen, W.; Kuick, R.; Ghosh, D.; Macdonald, J.; Cho, K.R.; Giordano, T.J.; Gruber, S.B.; Fearon, E.R.; Taylor, J.M.G.; et al. Comparison of Seven Methods for Producing Affymetrix Expression Scores Based on False Discovery Rates in Disease Profiling Data. *BMC Bioinform.* **2005**, *6*, 1–12.
5. Dubois, P.C.; Trynka, G.; Franke, L.; Hunt, K.A.; Romanos, J.; Curtotti, A.; Zhernakova, A.; Heap, G.A.R.; Adány, R.; Aromaa, A.; et al. Multiple Common Variants for Celiac Disease Influencing Immune Gene Expression. *Nat. Genet.* **2010**, *42*, 295–302.
6. Huang, Y.; Li, C.; Li, R.; Yang, S. An Overview of Tests on High-Dimensional Means. *J. Multivar. Anal.* **2022**, *188*, 104813.
7. Bai, Z.; Saranadasa, H. Effect of High Dimension: By an Example of a Two Sample Problem. *Stat. Sin.* **1996**, *6*, 311–329.
8. Srivastava, M.S.; Du, M. A Test for the Mean Vector with Fewer Observations than the Dimension. *J. Multivar. Anal.* **2008**, *99*, 386–402.
9. Srivastava, M.S. A Test for the Mean Vector with Fewer Observations than the Dimension Under Non-Normality. *J. Multivar. Anal.* **2009**, *100*, 518–532.
10. Chen, S.X.; Qin, Y.L. A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing. *Ann. Stat.* **2010**, *38*, 808–835.
11. Li, H.; Aue, A.; Paul, D.; Peng, J.; Wang, P. An Adaptable Generalization of Hotelling's T^2 Test in High Dimension. *Ann. Stat.* **2020**, *48*, 1815–1847.
12. Cai, T.T.; Liu, W.; Xia, Y. Two-Sample Test of High Dimensional Means Under Dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76*, 349–372.
13. Chang, J.; Zheng, C.; Zhou, W.X.; Zhou, W. Simulation-Based Hypothesis Testing of High Dimensional Means Under Covariance Heterogeneity. *Biometrics* **2017**, *73*, 1300–1310.
14. Xue, K.; Yao, F. Distribution and Correlation-Free Two-Sample Test of High-Dimensional Means. *Ann. Stat.* **2020**, *48*, 1304–1328.
15. Xu, G.; Lin, L.; Wei, P.; Pan, W. An Adaptive Two-Sample Test for High-Dimensional Means. *Biometrika* **2016**, *103*, 609–624.
16. Zhou, C.; Zhang, X.; Zhou, W.; Liu, H. A Unified Framework for Testing High Dimensional Parameters: A Data-Adaptive Approach. *arXiv* **2018**, arXiv:1808.02648.
17. Feng, L.; Jiang, T.; Li, X.; Liu, B. Asymptotic independence of the sum and maximum of dependent random variables with applications to high-dimensional tests. *Stat. Sin.* **2024**, *34*, 1745–1763.
18. Lopes, M.E.; Jacob, L.J.; Wainwright, M.J. A More Powerful Two-Sample Test in High Dimensions Using Random Projection. *arXiv* **2012**, arXiv:1108.2401v2.
19. Srivastava, R.; Li, P.; Ruppert, D. RAPTT: An Exact Two-Sample Test in High Dimensions Using Random Projections. *J. Comput. Graph. Stat.* **2016**, *25*, 954–970.
20. Liu, W.; Yu, X.; Zhong, W.; Li, R. Projection Test for Mean Vector in High Dimensions. *J. Am. Stat. Assoc.* **2024**, *119*, 744–756.

21. Zhou, C.; Kong, X.B. Testing of High Dimensional Mean Vectors via Approximate Factor Model. *J. Stat. Plan. Inference* **2015**, *167*, 216–227.
22. Zhang, M.; Zhou, C.; He, Y.; Zhang, X. Adaptive Test for Mean Vectors of High-Dimensional Time Series Data with Factor Structure. *J. Korean Stat. Soc.* **2018**, *47*, 450–470.
23. He, Y.; Zhang, M.; Zhang, X.; Zhou, W. High-Dimensional Two-Sample Mean Vectors Test and Support Recovery with Factor Adjustment. *Comput. Stat. Data Anal.* **2020**, *151*, 107004.
24. Ma, H.; Feng, L.; Wang, Z.; Bao, J. Adaptive testing for alphas in conditional factor models with high dimensional assets. *J. Bus. Econ. Stat.* **2024**, *42*, 1356–1366.
25. Johnstone, I.M. On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *Ann. Stat.* **2001**, *29*, 295–327.
26. Soshnikov, A. A Note on Universality of the Distribution of the Largest Eigenvalues in Certain Sample Covariance Matrices. *J. Stat. Phys.* **2002**, *108*, 1033–1056.
27. Péché, S. Universality Results for the Largest Eigenvalues of Some Sample Covariance Matrix Ensembles. *Probab. Theory Relat. Fields* **2009**, *143*, 481–516.
28. Birke, M.; Dette, H. A Note on Testing the Covariance Matrix for Large Dimension. *Stat. Probab. Lett.* **2005**, *74*, 281–289.
29. Srivastava, M.S. Some Tests Concerning the Covariance Matrix in High Dimensional Data. *J. Jpn. Stat. Soc.* **2005**, *35*, 251–272.
30. Chen, S.X.; Zhang, L.X.; Zhong, P.S. Tests for High-Dimensional Covariance Matrices. *J. Am. Stat. Assoc.* **2010**, *105*, 810–819.
31. Schott, J.R. A Test for the Equality of Covariance Matrices When the Dimension Is Large Relative to the Sample Sizes. *Comput. Stat. Data Anal.* **2007**, *51*, 6535–6542.
32. Srivastava, M.S.; Yanagihara, H. Testing the Equality of Several Covariance Matrices with Fewer Observations than the Dimension. *J. Multivar. Anal.* **2010**, *101*, 1319–1329.
33. Li, J.; Chen, S.X. Two Sample Tests for High-Dimensional Covariance Matrices. *Ann. Stat.* **2012**, *40*, 908–940.
34. Cai, T.T.; Ma, Z. Optimal Hypothesis Testing for High Dimensional Covariance Matrices. *Bernoulli* **2013**, *19*, 2359–2388.
35. Cai, T.T.; Jiang, T. Limiting Laws of Coherence of Random Matrices With Applications to Testing Covariance Structure and Construction of Compressed Sensing Matrices. *Ann. Stat.* **2011**, *39*, 1496–1525.
36. Qiu, Y.; Chen, S.X. Test for Bandedness of High-Dimensional Covariance Matrices and Bandwidth Estimation. *Ann. Stat.* **2012**, *40*, 1285.
37. Cai, T.T.; Zhang, A. Inference for High-Dimensional Differential Correlation Matrices. *J. Multivar. Anal.* **2016**, *143*, 107–126.
38. Zheng, S.; Cheng, G.; Guo, J.; Zhu, H. Test for High Dimensional Correlation Matrices. *Ann. Stat.* **2019**, *47*, 2887–2921.
39. Zhou, W. Asymptotic Distribution of the Largest Off-Diagonal Entry of Correlation Matrices. *Trans. Am. Math. Soc.* **2007**, *359*, 5345–5363.
40. Bao, Z.; Lin, L.C.; Pan, G.; Zhou, W. Spectral Statistics of Large Dimensional Spearman’s Rank Correlation Matrix and Its Application. *Ann. Stat.* **2015**, *43*, 2588–2623.
41. Han, F.; Chen, S.; Liu, H. Distribution-Free Tests of Independence in High Dimensions. *Biometrika* **2017**, *104*, 813–828.
42. Zhou, C.; Han, F.; Zhang, X.S.; Liu, H. An Extreme-Value Approach for Testing the Equality of Large U-Statistic Based Correlation Matrices. *Bernoulli* **2019**, *25*, 1472–1503.
43. He, Y.; Xu, G.; Wu, C.; Pan, W. Asymptotically Independent U-Statistics in High-Dimensional Testing. *Ann. Stat.* **2021**, *49*, 154.
44. Chen, X. Gaussian and Bootstrap Approximations for High-Dimensional U-Statistics and Their Applications. *Ann. Stat.* **2018**, *46*, 642–678.
45. Tukey, J.W. Bias and Confidence in Not Quite Large Samples. *Ann. Math. Stat.* **1958**, *29*, 614.
46. Quenouille, M.H. Notes On Bias In Estimation. *Biometrika* **1956**, *3–4*, 3–4.
47. Shi, X. The Approximate Independence of Jackknife Pseudo-Values and the Bootstrap Methods. *J. Wuhan Univ. Hydraul. Electr. Eng.* **1984**, *2*, 83–90.
48. Chernozhukov, V.; Chetverikov, D.; Kato, K. Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors. *Ann. Stat.* **2013**, *41*, 2786–2819.
49. Cai, T.; Liu, W.; Xia, Y. Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings. *J. Am. Stat. Assoc.* **2013**, *108*, 265–277.
50. Chang, J.; Zhou, W.; Zhou, W.X.; Wang, L. Comparing Large Covariance Matrices Under Weak Conditions on the Dependence Structure and Its Application to Gene Clustering. *Biometrics* **2017**, *73*, 31–41.
51. Mazieres, J.; He, B.; You, L.; Xu, Z.; Jablons, D.M. Wnt Signaling in Lung Cancer. *Cancer Lett.* **2005**, *222*, 1–10.
52. Clements, W.M.; Wang, J.; Sarnaik, A.; Kim, O.J.; Macdonald, J.; Fenogliopreiser, C.; Groden, J.; Lowy, A.M. Beta-Catenin Mutation Is a Frequent Cause of Wnt Pathway Activation in Gastric Cancer. *Cancer Res.* **2002**, *62*, 3503–6.
53. Howe, L.R.; Brown, A.M. Wnt Signaling and Breast Cancer. *Cancer Biol. Ther.* **2004**, *3*, 36–41.
54. Corda, G.; Sala, A. Non-Canonical WNT/PCP Signalling in Cancer: Fzd6 Takes Centre Stage. *Oncogenesis* **2017**, *6*, e364.
55. Rapp, J.; Jaromi, L.; Kvell, K.; Miskei, G.; Pongracz, J.E. WNT Signaling Lung Cancer Is No Exception. *Respir. Res.* **2017**, *18*, 167.
56. Kim, K.B.; Kim, D.W.; Kim, Y.; Tang, J.; Kirk, N.; Gan, Y.; Kim, B.; Fang, B.; Park, J.I.; Zheng, Y.; et al. WNT5A–RHOA signaling is a driver of tumorigenesis and represents a therapeutically actionable vulnerability in small cell lung cancer. *Cancer Res.* **2022**, *82*, 4219–4233.
57. Lloyd, C.J. Estimating test power adjusted for size. *J. Stat. Comput. Simul.* **2005**, *75*, 921–933.
58. Cuparić, M.; Milošević, B. To impute or to adapt? model specification tests’ perspective. *Stat. Pap.* **2024**, *65*, 1021–1039.

59. Papadimitriou, C.K.; Meintanis, S.G.; Andrade, B.B.; Tsionas, M.G. Specification tests for normal/gamma and stable/gamma stochastic frontier models based on empirical transforms. *Econom. Stat.* **2024**, *in press*.
60. Delaigle, A.; Hall, P.; Jin, J. Robustness and Accuracy of Methods for High Dimensional Data Analysis Based on Student's T-Statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2011**, *73*, 283–301.
61. Boucheron, L.G.S.; Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*; Oxford University Press: Oxford, UK, 2013.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.