Computational Statistics and Data Analysis 173 (2022) 107495

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

A rank-based high-dimensional test for equality of mean vectors

Yanyan Ouyang [a], Jiamin Liu [a,b], Tiejun Tong [c], Wangli Xu [a,*]

[a] Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, PR China
[b] Department of Mathematics, City University of Hong Kong, Hong Kong
[c] Department of Mathematics, Hong Kong Baptist University, Hong Kong

ARTICLE INFO

Article history:
Received 13 April 2021
Received in revised form 1 April 2022
Accepted 1 April 2022
Available online 6 April 2022

Keywords:
Equality of means
High-dimensional data
Wilcoxon signed-rank test
Wilcoxon-Mann-Whitney test

ABSTRACT

The Wilcoxon signed-rank test and the Wilcoxon-Mann-Whitney test are two commonly used rank-based methods for one- and two-sample tests when the one-dimensional data are not normally distributed. The new rank-based nonparametric tests for equality of mean vectors are proposed in the high-dimensional settings. To overcome the technical challenges in data sorting, the new statistics are constructed by taking the sum of the Wilcoxon signed-rank or Wilcoxon-Mann-Whitney test statistics from each dimension of the data. The asymptotic properties of the proposed test statistics are investigated under the null and local alternative hypotheses. Simulation studies show that the new tests perform as well as the state-of-the-art methods when the high-dimensional data are normally distributed, but they turn out to be more powerful when the normality assumption is violated. Finally, the new testing methods are also applied to a human peripheral blood mononuclear cells gene expression data set for demonstrating their usefulness in practice.

1. Introduction

Hypothesis testing for means is one of the most important problems in many practical applications. In this paper, we are interested in high-dimensional temporally independent data and want to test whether the mean vector equals to a given vector $\boldsymbol{\mu_0}$ in one-sample case, or whether the two mean vectors are equal in two-sample case. To be precise, let $\{\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}, i = 1, \ldots, m\}$ and $\{\mathbf{y}_i = (Y_{i1}, \ldots, Y_{ip})^{\mathrm{T}}, i = 1, \ldots, n\}$ be two independent random samples from two distributions with distribution functions $F(\mathbf{x} - \boldsymbol{\mu_1})$ and $F(\mathbf{y} - \boldsymbol{\mu_2})$, where $\boldsymbol{\mu_1} \overset{\text{def}}{=} (\mu_{11}, \ldots, \mu_{1p})^{\mathrm{T}}$ and $\boldsymbol{\mu_2} \overset{\text{def}}{=} (\mu_{21}, \ldots, \mu_{2p})^{\mathrm{T}}$ are the true mean vectors, respectively. Without loss of generality, we also let $\boldsymbol{\mu}_0 = 0$ throughout the paper. Then for one-sample case, the null and alternative hypotheses are

$$H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_0} \quad \text{versus} \quad H_1 : \boldsymbol{\mu_1} \neq \boldsymbol{\mu_0}; \tag{1}$$

and for two-sample case,

$$H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2} \quad \text{versus} \quad H_1 : \boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}. \tag{2}$$

* Corresponding author.
E-mail address: wlxu@ruc.edu.cn (W. Xu).

https://doi.org/10.1016/j.csda.2022.107495
0167-9473/

Hotelling's T^2 test is the most classic method for testing hypotheses (1) and (2) with fixed $p \geq 2$. For one-sample case, Hotelling's T^2 statistic is constructed as $T_1^2 = m\bar{\mathbf{x}}^{\mathrm{T}} S_{\mathrm{x}}^{-1} \bar{\mathbf{x}}$ where $\bar{\mathbf{x}} = m^{-1} \sum_{i=1}^{m} \mathbf{x}_i$ and $S_{\mathrm{x}} = (m-1)^{-1} \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$; and for two-sample case, it is $T_2^2 = mn(m+n)^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})^{\mathrm{T}} S_{\mathrm{w}}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$ where $\bar{\mathbf{y}} = n^{-1} \sum_{i=1}^{n} \mathbf{y}_i$ and $S_{\mathrm{w}} = (m+n-2)^{-1} \{\sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} + \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^{\mathrm{T}}\}$. With the rapid development of high-dimensional data especially in biological sciences (Dan et al., 2008), Hotelling's T^2 test statistics for (1) and (2) are challenged with the so-called "large-p-small-n" problem. Due to the difficulty in solving the inverse of high-dimensional covariance matrices S_{x} and S_{w}, Hotelling's T^2 test will no longer be applicable.

To overcome the problem, various test statistics for (1) and (2) have been proposed for testing high-dimensional data in the recent literature. Using two-sample test for illustration, researchers have made a great effort on improving the sample estimates of the covariance matrix S_{w}. To name a few, Bai and Saranadasa (1996) replaced S_{w} in T_2^2 with the identify matrix I_p. Chen and Qin (2010) modified the test statistics in Bai and Saranadasa (1996) to further relax the restriction on the relationship between the data dimension and sample size. Wang and Shao (2020) extended the U-statistic in Chen and Qin (2010) to high-dimensional stationary process by self-normalization. Other replacements of S_{w} also include $\mathrm{diag}(S_{\mathrm{w}})$ in Srivastava and Du (2008) and $(S_{\mathrm{w}} + \lambda I_p)$ in Chen et al. (2011), where, in particular, the test statistic with $\mathrm{diag}(S_{\mathrm{w}})$ can be viewed as the sum of the squares of Student's t test statistics under the condition of the same variance between the two samples for all p components. Gregory et al. (2015) put forward another test statistic according to the mean of the squares of Student's t test statistics under the condition of different variance between the two samples for all p components, and they further derived the asymptotic normality of their test statistic without restriction on the data distribution. More recently, Hu et al. (2019) proposed a likelihood ratio test for normal mean vectors in high-dimensional data, and their test statistic is given as the sum of likelihood ratio test statistics from each component. For other works related to Hotelling's tests, see, for example, Wu et al. (2006), Srivastava (2007), Yamada and Srivastava (2012), Cai et al. (2013), Dong et al. (2016), Zhao and Xu (2016), Zoh et al. (2018) and Li et al. (2020). In high-dimensional case, some extreme value-type statistics were developed for sparse alternatives, and their limiting distributions were studied by Gaussian approximation with examples including the methods for independent Gaussian process sequences (Chernozhukov et al., 2013) and for weak correlated sequences (Zhang and Cheng, 2018). Moreover, Zhang and Wu (2017) also used Gaussian approximation for mean vector testing in high-dimensional stationary process.

In the era of explosive growth of biological information, the study of gene expression data for diabetes has been of great significance. Nevertheless, as shown in Fig. 7, it is not uncommon that many gene expression data may not follow a normal distribution. According to the official website of the International Diabetes Federation (IDF), there were 463 million adults aged 20-79 suffering from diabetes in the world in 2019; in other words, one in 11 people is diabetic. It was also estimated that the number of diabetic patients will reach 578.4 million by 2030. Also in 2019, about 4.2 million adults died of diabetes or its complications, which is equivalent to one person dying of diabetes every 8 seconds, accounting for 11.3% of all deaths worldwide. Wu et al. (2007) obtained muscle biopsies for 20 insulin sensitive individuals both before and after insulin treatment and identified 779 insulin-responsive genes after microarray data analysis. Wang et al. (2015) developed a high-dimensional nonparametric multivariate test, and then by applying it to the gene expression data in Wu et al. (2007), a total of 954 gene sets out of 2519 candidate genes sets were identified as significant.

In this paper, we focus on whether there are differences in gene expression between healthy people and patients with diabetes. The existing methods based on the normality assumption, however, may not work well or is not robust when the sample sizes are relatively small. To solve the problem, we propose new test statistics by taking the sum of the component-wise Wilcoxon signed-rank statistics for hypothesis (1) in one-sample case, and the component-wise Wilcoxon-Mann-Whitney (WMW) statistics for hypothesis (2) in two-sample case. As we will show, the proposed method has two main advantages: 1) the proposed rank-based method is more robust; and 2) the new test method requires a weaker assumption on the data distribution. Moreover, the asymptotic properties of the proposed tests will also be investigated under both the null and local alternative hypotheses, together with the derivation of the critical values under the null hypothesis.

The rest of the paper is organized as follows. In Section 2, we review the Wilcoxon signed-rank test and propose a rank-based test statistic in high-dimensional case for testing (1). The asymptotic distributions under the null and local alternative hypotheses will also be derived. In Section 3, we review the WMW test and extend it to the setting of multivariate dimension for two-sample case in (2), followed by the derivation of the asymptotic behaviors under the null and local alternative hypotheses. We then conduct extensive numerical studies to evaluate the performance of the proposed rank-based test from Section 3 and compare it with some state-of-the-art methods in Section 4. The proposed test for two-sample case is also applied to a gene expression data set in Section 5. Finally, we conclude the paper in Section 6, and provide the technical results in the Appendix.

2. One-sample test

2.1. Wilcoxon signed-rank test in high dimension

To introduce the procedure for one-sample test in multivariate case clearly, we first review Wilcoxon signed-rank test briefly. In the case of $p = 1$, we suppose $\{\mathbf{x}_i = X_{i1}, i = 1, \ldots, m\}$ is the random sample drawn from the symmetrical distribution $F_1(x - \mu_{11})$ with $E(X_{11}) = \mu_{11}$, where $F_1(\cdot)$ is a continuous distribution function and the corresponding density

function $f_1(x)$ is symmetrical about μ_{11}. For the hypothesis testing problem $H_0 : \mu_{11} = 0$ versus $H_1 : \mu_{11} \neq 0$, we define R^{x}_{i1} as the rank of $|X_{i1}|$ among $\{|X_{i1}|, i = 1, \ldots, m\}$, where $|X_{i1}|$ is the absolute value of X_{i1}. The Wilcoxon signed-rank test is defined as

$$U^1_m \overset{\text{def}}{=} \sum_{i=1}^{m} R^{\mathrm{x}}_{i1} I(X_{i1} > 0), \tag{3}$$

where $I(X_{i1} > 0)$ is the indicator function. Under the null hypothesis that $\mu_{11} = 0$, we have $E(U^1_m) = m(m+1)/4$. Hence, the null hypothesis will be rejected when $\{U^1_m - m(m+1)/4\}^2$ is sufficiently large.

For p-dimensional case, the Wilcoxon signed-rank test for each of p components, U^j_m, can also be defined similarly as (3). That is

$$U^j_m \overset{\text{def}}{=} \sum_{i=1}^{m} R^{\mathrm{x}}_{ij} I(X_{ij} > 0), \quad \text{for } j = 1, 2, \ldots, p,$$

where R^{x}_{ij} is the rank of $|X_{ij}|$ among $\{|X_{ij}|, i = 1, \ldots, m\}$ for fixed j. In parallel to univariate case, the null hypothesis $\mu_{1j} = 0$ will be rejected when $\{U^j_m - E(U^j_m)\}^2$ is sufficiently large with $E(U^j_m) = m(m+1)/4$. Let $W^j_m \overset{\text{def}}{=} \{U^j_m - m(m+1)/4\}^2$, the sum-of-square based test statistic for multivariate test (1) is defined as

$$W_m \overset{\text{def}}{=} p^{-1} \sum_{j=1}^{p} W^j_m = p^{-1} \sum_{j=1}^{p} \Big\{ U^j_m - \frac{m(m+1)}{4} \Big\}^2. \tag{4}$$

We reject H_0 if W_m is sufficiently large.

2.2. Asymptotic null distribution of W_m

In this section, we derive the asymptotic normality of W_m in (4) under H_0 in (1). Firstly, we consider the expectation and variance of W^j_m. Under the null hypothesis, it can be derived that $\{W^j_m, j = 1, \ldots, p\}$ is a series of random variables with same expectation $E(W^j_m)$ and variance $\mathrm{var}(W^j_m)$, where

$$\begin{aligned} &E(W^j_m) = m(m+1)(2m+1)/24 \quad \text{and} \\ &\mathrm{var}(W^j_m) = (6m + 5m^2 - 30m^3 - 25m^4 + 24m^5 + 20m^6)/1440. \end{aligned} \tag{5}$$

To ease subsequent illustration, we define $E(W^j_m) \overset{\text{def}}{=} \xi_m$ and $\mathrm{var}(W^j_m) \overset{\text{def}}{=} \eta_m$.

To prove the asymptotic property of W_m, we also need to consider the dependent structure among $\{V^j_m, j = 1, \ldots, p\}$ with definition $V^j_m \overset{\text{def}}{=} (W^j_m - \xi_m)/\sqrt{\eta_m}$. For any σ-fields $\mathcal{F}$ and $\mathcal{G}$, $\alpha(\mathcal{F}, \mathcal{G}) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}, B \in \mathcal{G}\}$ denotes the strong mixing coefficient. To derive the asymptotic property of the statistic W_m, we assume the following Conditions (C1) and (C2).

(C1) Let $\alpha(r) = \sup\{\alpha(\mathcal{F}^k_1, \mathcal{F}^p_{k+r}) : 1 \le k \le p - r\}$, where $\mathcal{F}^b_a = \sigma\{V^j_m : a \le j \le b\}$. Assume that the stationary sequence $\{V^j_m\}$ satisfies the strong mixing condition such that $\alpha(r) \downarrow 0$ as $r \to \infty$, where $\downarrow$ denotes the monotone decreasing convergence.

(C2) Suppose that $\sum_{r=1}^{\infty} \alpha(r)^{\delta/(2+\delta)} < \infty$ for some $\delta > 0$, and for any $k \ge 0$, $\lim_{p\to\infty} \sum_{j=1}^{p-k} \mathrm{cov}(V^j_m, V^{j+k}_m)/(p-k) = \gamma_1(k)$ exists.

Based on Conditions (C1) and (C2), Theorem 1 shows that the asymptotic distribution of W_m under H_0 in (1) is a normal distribution. In what follows, the symbol $\xrightarrow{D}$ denotes convergence in distribution.

Theorem 1. *Suppose that the sequence $\{V^j_m\}$ is stationary and satisfies Conditions (C1) and (C2). Under H_0 in (1), for any fixed $m \ge 2$, we have*

$$Z_m \overset{\text{def}}{=} \frac{\sqrt{p}(W_m - \xi_m)}{\sqrt{\eta_m}} \xrightarrow{D} N(0, \tau_1^2), \quad \text{as } p \to \infty, \tag{6}$$

where $\tau_1^2 = 1 + 2\sum_{k=1}^{\infty} \gamma_1(k)$.

Let $\mathbf{\Sigma}$ be the $p \times p$ covariance matrix of random vector $\mathbf{x}$. When $\mathbf{\Sigma}$ is a diagonal matrix so that the variables are not correlated, the sequence $\{V^j_m\}$ is stationary and $\gamma_1(k) = 0$ for $k = 1, \ldots, \infty$. Conditions (C1) and (C2) are satisfied naturally and $\tau_1^2 = 1$. Thus, the asymptotic result with the diagonal matrix $\mathbf{\Sigma}$ can be derived from Theorem 1 and is shown in Corollary 1.1.

Corollary 1.1. *When $\mathbf{\Sigma}$ is a diagonal matrix, under H_0 in (1), for any fixed $m \geq 2$, we have*

$$\frac{\sqrt{p}(W_m - \xi_m)}{\sqrt{\eta_m}} \xrightarrow{D} N(0,1), \quad \text{as } p \to \infty.$$

According to Theorem 1, in order to derive the critical value, the parameter τ_1^2 in (6) needs to be estimated. Following Hu et al. (2019), we apply the spectrum analysis to estimate τ_1^2 as follows:

$$\hat{\tau}_1^2 \stackrel{\text{def}}{=} \sum_{0<|k|<L} w(k/L)\hat{\gamma}_1(k) + 1, \tag{7}$$

which depends on estimation $\hat{\gamma}_1(k)$ of autocovariance $\gamma_1(k)$ for $\{V_m^j, j = 1, \ldots, p\}$. Recalling the definition of $\gamma_1(k)$, the estimator of the autocovariance $\gamma_1(k)$ is given by

$$\hat{\gamma}_1(k) \stackrel{\text{def}}{=} (p-k)^{-1} \sum_{j=1}^{p-k} \big(V_m^j - p^{-1}\sum_{j=1}^{p} V_m^j\big)\big(V_m^{j+k} - p^{-1}\sum_{j=1}^{p} V_m^j\big), \quad \text{for } k \geq 1,$$

and we have $\gamma_1(0) = \text{var}(V_m^j) = 1$ under H_0 in (1). The lag-window size L will be specified in numerical analysis. The function $w(\cdot)$ in (7) is lag-window function, and is chosen as the following Parzen window function as Brockwell and Davis (2009),

$$w(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3, & |x| < 1/2, \\ 2(1-|x|)^3, & 1/2 \leq |x| \leq 1, \\ 0, & |x| > 1. \end{cases} \tag{8}$$

2.3. Statistical power under the local alternative

To study the power performance of the proposed statistic for one-sample test, we consider the local alternative that

$$H_1' : \boldsymbol{\mu}_1 = \boldsymbol{\delta}_1/\sqrt{m}, \tag{9}$$

where $\boldsymbol{\delta}_1 = (\delta_{11}, \ldots, \delta_{1p})^{\mathrm{T}}$. We assume the j-th component of $\mathbf{x}$, X_j, has the distribution function $F_j(x - \delta_{1j}/\sqrt{m})$ and the density function $f_j(x - \delta_{1j}/\sqrt{m})$. With the help of Taylor expansion, it can be derived that the expectation of W_m^j under the local alternative (9) is

$$E(W_m^j) = \frac{m(m+1)(2m+1)}{24} + \delta_{1j}^2 m^3 \left[\int_0^\infty \{2F_j(x) - 1\} df_j(x)\right]^2 + o(m^3). \tag{10}$$

By simple algebraic calculation, we can obtain that $\lim_{m\to\infty} \sum_{j=1}^{p} \left\{E(W_m^j) - \xi_m\right\} / \sqrt{\eta_m} = \Delta_1$ and

$$\Delta_1 \stackrel{\text{def}}{=} 6\sqrt{2}\sum_{j=1}^{p} \delta_{1j}^2 \left[\int_0^\infty \{2F_j(x) - 1\} df_j(x)\right]^2. \tag{11}$$

Thus, according to the conclusions in (10) and (11), we can obtain the asymptotic distribution of W_m under the local alternative (9) and the asymptotic power of the level α test, which are summarized in Theorem 2.

Theorem 2. *If the sequence $\{V_m^j\}$ is stationary and satisfies Conditions (C1) and (C2), the asymptotic distribution of Z_m is*

$$Z_m = \frac{\sqrt{p}(W_m - \xi_m)}{\sqrt{\eta_m}} \xrightarrow{D} C_1 \cdot N(0, \tau_1^2) + \frac{\Delta_1}{\sqrt{p}},$$

as $(m, p) \to \infty$, where $C_1 = \sqrt{\tau_1^2 - 1 + \gamma_1(0)}/\tau_1$. Then the asymptotic power of the level α test is

$$\beta(Z_m) = 1 - \Phi\left\{C_1^{-1}\left(Z_{1-\alpha} - \frac{\Delta_1}{\tau_1\sqrt{p}}\right)\right\}, \quad \text{as } (m, p) \to \infty.$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $\Phi(Z_{1-\alpha}) = 1 - \alpha$.

By $\beta(Z_m)$ in Theorem 2, it can be verified that the power function tends to 1 when $\Delta_1/\sqrt{p} \to \infty$ against $(m, p) \to \infty$. Since $|\int_0^\infty \{2F_j(x) - 1\} df_j(x)| \le 1$, we only need $\sqrt{p} = o\left(\sum_{j=1}^{p} \delta_{1j}^2\right)$. Thus for one-sample test, if the true mean vector is sparse, the power function going to 1 requires the signals to be strong such that $\sqrt{p} = o\left(\max_{1\le j\le p}\{\delta_{1j}^2\}\right)$. And if the signals are dense, we only need the weak signals $\delta_{1j}^2 = O(p^\alpha)$ with $\alpha > -1/2$ for each $1 \le j \le p$, which implies $\sqrt{p} = o\left(\sum_{j=1}^{p} \delta_{1j}^2\right)$ and makes the power of Z_m increasing towards 1 as $p \to \infty$.

3. Two-sample test

3.1. Wilcoxon-Mann-Whitney test in high dimension

For two-sample case, we suppose that distribution functions of random vectors $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and $\mathbf{y} = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ are $F(\mathbf{x} - \boldsymbol{\mu}_1)$ and $F(\mathbf{y} - \boldsymbol{\mu}_2)$, respectively. We first review the WMW test briefly. In the case of $p = 1$, we suppose $\{\mathbf{x}_i = X_{i1}, i = 1, \ldots, m\}$ and $\{\mathbf{y}_k = Y_{k1}, k = 1, \ldots, n\}$ are independent samples drawn from distributions with distribution functions $F_1(x - \mu_{11})$ and $F_1(y - \mu_{21})$, respectively. For the testing hypothesis problem that $H_0 : \mu_{11} = \mu_{21}$ versus $H_1 : \mu_{11} \ne \mu_{21}$, the WMW statistic is defined as

$$U_{m,n}^1 \stackrel{\text{def}}{=} \sum_{i=1}^{m} R_{i1}^{\text{xy}} - \frac{m(m+1)}{2}, \tag{12}$$

where R_{i1}^{xy} denotes the rank of X_{i1} in the mixed sample $\{X_{11}, \ldots, X_{m1}; Y_{11}, \ldots, Y_{n1}\}$. Under H_0, we have $E(U_{m,n}^1) = mn/2$. The null hypothesis will be rejected when the value of $(U_{m,n}^1 - mn/2)^2$ is sufficiently large.

For p-dimensional case, the WMW test for each of p components, $U_{m,n}^j$, can be also defined as

$$U_{m,n}^j \stackrel{\text{def}}{=} \sum_{i=1}^{m} R_{ij}^{\text{xy}} - \frac{m(m+1)}{2}, \quad \text{for } j = 1, 2, \ldots, p,$$

where R_{ij}^{xy} is the rank of X_{ij} among $\{X_{1j}, \ldots, X_{mj}; Y_{1j}, \ldots Y_{nj}\}$ for fixed j. In parallel to univariate case, the null hypothesis $\mu_{1j} = \mu_{2j}$ will be rejected when $\{U_{m,n}^j - E(U_{m,n}^j)\}^2$ is sufficiently large with $E(U_{m,n}^j) = mn/2$. Let $W_{m,n}^j \stackrel{\text{def}}{=} (U_{m,n}^j - mn/2)^2$, the proposed statistic is defined as

$$W_{m,n} \stackrel{\text{def}}{=} p^{-1} \sum_{j=1}^{p} W_{m,n}^j = p^{-1} \sum_{j=1}^{p} \left(U_{m,n}^j - \frac{mn}{2}\right)^2. \tag{13}$$

We reject H_0 in (2) when $W_{m,n}$ is sufficiently large.

3.2. Asymptotic null distribution of $W_{m,n}$

In this section, we derive the asymptotic normality of the proposed statistic $W_{m,n}$ in (13) under H_0 in (2). First of all, we calculate the expectation $E(W_{m,n}^j)$ and variance $\text{var}(W_{m,n}^j)$ of $W_{m,n}^j$. By letting $N \stackrel{\text{def}}{=} m + n$, we have

$$\xi_{m,n} \stackrel{\text{def}}{=} E(W_{m,n}^j) = mn(N+1)/12 \quad \text{and}$$
$$\eta_{m,n} \stackrel{\text{def}}{=} \text{var}(W_{m,n}^j) = \{mn(5N+8) - 3N(N+1)\}(N+1)mn/360.$$

Let $V_{m,n}^j \stackrel{\text{def}}{=} (W_{m,n}^j - \xi_{m,n})/\sqrt{\eta_{m,n}}$, we consider the dependence structure of $\{V_{m,n}^j, j = 1, \ldots, p\}$ to investigate the asymptotic property under H_0. The following conditions are assumed for deriving the asymptotic distribution of the test statistic $W_{m,n}$. We use the same notation $\alpha(r)$ as that in Conditions (C1)-(C2) when this does not cause ambiguity.

(C3) Let $\alpha(r) = \sup\{\alpha(\mathcal{F}_1^k, \mathcal{F}_{k+r}^p) : 1 \le k \le p - r\}$, where $\mathcal{F}_a^b = \sigma\{V_{m,n}^j : a \le j \le b\}$. Assume that the stationary sequence $\{V_{m,n}^j\}$ satisfies the strong mixing condition such that $\alpha(r) \downarrow 0$ as $r \to \infty$, where $\downarrow$ denotes the monotone decreasing convergence.

(C4) Suppose that $\sum_{r=1}^{\infty} \alpha(r)^{\delta/(2+\delta)} < \infty$ for some $\delta > 0$, and for any $k \ge 0$, $\lim_{p\to\infty} \sum_{j=1}^{p-k} \text{cov}(V_{m,n}^j, V_{m,n}^{j+k})/(p-k) = \gamma_2(k)$ exists.

Through similar proof techniques in one-sample test, we can derive the following asymptotic normality of the proposed statistic $W_{m,n}$ in (13) under the null hypothesis.

Theorem 3. *Suppose that the sequence $\{V^j_{m,n}\}$ is stationary and satisfies Conditions (C3) and (C4). Under H_0 in (2), for any fixed $m, n \geq 2$, we have*

$$Z_{m,n} \overset{\text{def}}{=} \frac{\sqrt{p}(W_{m,n} - \xi_{m,n})}{\sqrt{\eta_{m,n}}} \xrightarrow{D} N(0, \tau_2^2), \quad \text{as } p \to \infty,$$

where $\tau_2^2 = 1 + 2\sum_{k=1}^{\infty} \gamma_2(k)$.

In parallel to the one-sample case, we have $\gamma_2(0) = \text{var}(V^j_{m,n}) = 1$ under H_0. The estimator of the autocovariance $\gamma_2(k)$ for the sequence $\{V^j_{m,n}\}$ is estimated by

$$\hat{\gamma}_2(k) = (p-k)^{-1} \sum_{j=1}^{p-k} \left(V^j_{m,n} - p^{-1}\sum_{j=1}^{p} V^j_{m,n}\right)\left(V^{j+k}_{m,n} - p^{-1}\sum_{j=1}^{p} V^j_{m,n}\right). \tag{14}$$

Thus, we use the estimator $\hat{\tau}_2^2 = \sum_{0<|k|<L} w(k/L)\hat{\gamma}_2(k) + 1$ for τ_2^2 in order to obtain the critical value.

3.3. Statistical power under the local alternative

To derive the asymptotic power of the proposed statistic for two-sample test, we consider the local alternative that

$$H_1'' : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_2/\sqrt{N}, \tag{15}$$

where $\boldsymbol{\delta}_2 = (\delta_{21}, \ldots, \delta_{2p})^{\mathrm{T}}$. Without loss of generality, we let $\boldsymbol{\mu}_1 = (0, \ldots, 0)^{\mathrm{T}}$, then $\mathbf{x}$ follows the distribution function $F(\mathbf{x})$ and $\mathbf{y}$ follows the distribution function $F(\mathbf{y} + \boldsymbol{\delta}_2/\sqrt{N})$. For j-th dimension, we suppose that the j-th component of $\mathbf{x}$, X_j, follows the distribution function $F_j(x)$ and the j-th component of $\mathbf{y}$, Y_j, follows the distribution function $F_j(y + \delta_{2j}/\sqrt{N})$ and density function $f_j(y + \delta_{2j}/\sqrt{N})$ respectively when this does not cause ambiguity. Invoking Taylor expansion, we can derive the expectation of $W^j_{m,n}$ under the local alternative (15) is

$$E(W^j_{m,n}) = \frac{mn(N+1)}{12} + \frac{m^2n^2}{N}\delta_{2j}^2 E^2\{f_j(X_j)\} + o(N^3). \tag{16}$$

Define $\nu_{m,n} \overset{\text{def}}{=} m/N$ with $\nu_{m,n} \to \nu \in (0,1)$ as $m, n \to \infty$. By algebraic calculation, we have $\lim_{m,n\to\infty} \sum_{j=1}^{p} \left\{E(W^j_{m,n}) - \xi_{m,n}\right\} / \sqrt{\eta_{m,n}} = \Delta_2$, where

$$\Delta_2 \overset{\text{def}}{=} 6\sqrt{2}(1-\nu)\nu \sum_{j=1}^{p} E^2\left\{f_j(X_j)\right\}\delta_{2j}^2. \tag{17}$$

Thus, based on $E(W^j_{m,n})$ in (16), we can obtain the asymptotic power function for $W_{m,n}$ of the level α test given in Theorem 4.

Theorem 4. *If the sequence $\{V^j_{m,n}\}$ is stationary and satisfies Conditions (C3) and (C4),*

$$Z_{m,n} = \frac{\sqrt{p}(W_{m,n} - \xi_{m,n})}{\sqrt{\eta_{m,n}}} \xrightarrow{D} C_2 \cdot N(0, \tau_2^2) + \frac{\Delta_2}{\sqrt{p}},$$

as $(m, n, p) \to \infty$, where $C_2 = \sqrt{\tau_2^2 - 1 + \gamma_2(0)}/\tau_2$. Then the asymptotic power of the level α test is

$$\beta(Z_{m,n}) = 1 - \Phi\left\{C_2^{-1}\left(Z_{1-\alpha} - \frac{\Delta_2}{\tau_2\sqrt{p}}\right)\right\}, \quad \text{as } (m, n, p) \to \infty.$$

By $\beta(Z_{m,n})$ in Theorem 4, it can be verified that the power function tends to 1 when $\Delta_2/\sqrt{p} \to \infty$ against $(m, n, p) \to \infty$. Since $E^2\left\{f_j(X_j)\right\} \leq 1$, we only need $\sqrt{p} = o\left(\sum_{j=1}^{p} \delta_{2j}^2\right)$. Thus for two-sample test, if the true mean vector is sparse, the power function going to 1 requires the signals to be strong such that $\sqrt{p} = o\left(\max_{1\leq j\leq p}\{\delta_{2j}^2\}\right)$. And if the signals are dense, we only need the weak signals $\delta_{2j}^2 = O(p^{\alpha})$ with $\alpha > -1/2$ for each $1 \leq j \leq p$, which implies $\sqrt{p} = o\left(\sum_{j=1}^{p} \delta_{2j}^2\right)$ and makes the power of $Z_{m,n}$ increasing towards 1 as $p \to \infty$.

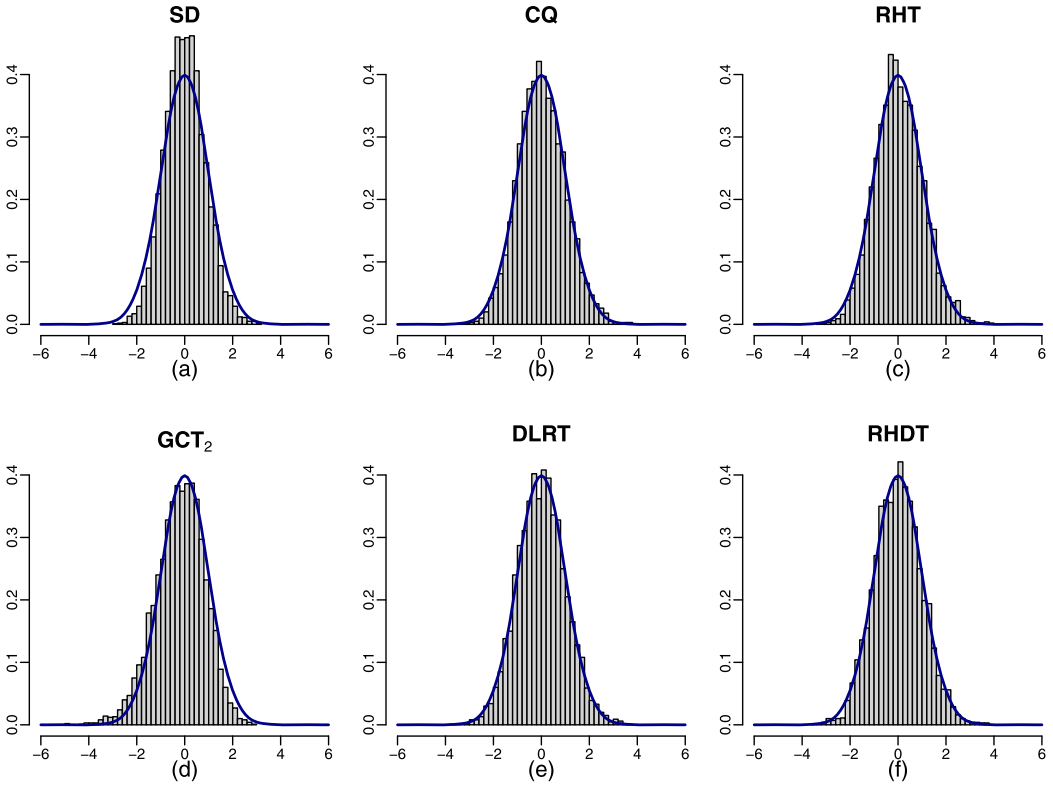

Fig. 1. The simulated null distributions of six test statistics and the line in each subgraph is the probability density function of $N(0,1)$ when the observations follow multivariate normal distribution. And (a)-(f) are corresponding to SD, CQ, RHT, GCT_2, DLRT and RHDT respectively.

4. Simulation study

In this section, we carry out simulation studies to evaluate the performance of the proposed method, which is referred to Rank-based High-Dimensional Test (RHDT). For space consideration, we mainly focus on the testing of mean vectors under two-sample scenarios. The proposed statistic is compared with the following competitors: the test SD in Srivastava and Du (2008), the test CQ in Chen and Qin (2010), the test RHT in Chen et al. (2011), the test GCT_1 and GCT_2 in Gregory et al. (2015), the test DLRT in Hu et al. (2019). According to Gregory et al. (2015), the statistic GCT_1 is adopted in "moderate-p" scenarios restricting p to grow at a rate such that $p = o(m^2)$, while the test statistic GCT_2 is implemented in "large-p" scenarios that allow $p = o(m^6)$.

For sake of comparison convenience, following the setting in Hu et al. (2019), we generate independent samples $\{\mathbf{x}_i, i = 1, 2, \ldots, m\}$ and $\{\mathbf{y}_i, i = 1, 2, \ldots, n\}$ with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrixes $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ from two different distributions, that is, multivariate normal distribution and multivariate t_d distribution with $d = 3$ degree of freedoms. To present the different performance of test methods under various dependent structures, we utilize the following four dependent structures in simulations: (1) independent structure (IND) that $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$; (2) ARMA structure with weak correlation (AR = 0.3) that $\boldsymbol{\Sigma} = (\rho^{|i-j|})_{p\times p}$ with $\rho = 0.3$; (3) ARMA structure with strong correlation (AR = 0.6) that $\boldsymbol{\Sigma} = (\rho^{|i-j|})_{p\times p}$ with $\rho = 0.6$; (4) long range dependent structure (LR) that $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = \{(|i-j|+1)^{2h} + (|i-j|-1)^{2h} - 2|i-j|^{2h}\}/2$ with $h = 0.625$. And we use lag-window size $L = 5$ to estimate τ_2^2 in (14).

4.1. Performance under the null hypothesis

In this subsection, we demonstrate the performance of the proposed statistic RHDT and other five testing methods by the simulated null distributions and the Type I error rates.

At first, we generate samples from multivariate normal distribution and multivariate t_3 distribution with $m = 10, n = 15$ and dimension $p = 500$. Figs. 1–2 display the histograms of six statistics under IND dependent structure. All histograms are obtained by 5000 simulations. The density function for $N(0,1)$ is presented in histograms for comparison. For space consideration, the simulated null distribution of test statistic GCT_1 is not shown, since the test statistic GCT_1 has a similar performance to GCT_2. When samples are from multivariate normal distribution, Fig. 1 shows the simulated null distributions of all statistics agree well with $N(0,1)$. Fig. 2 shows the simulated null distributions when samples are from multivariate

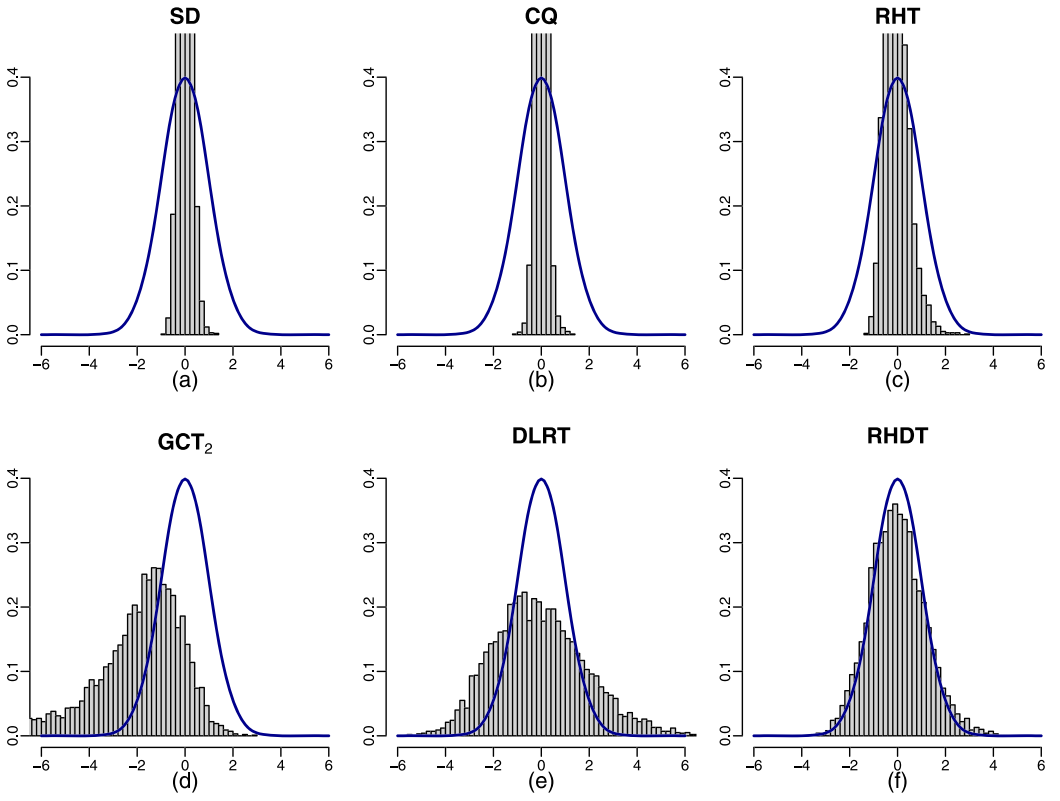

Fig. 2. The simulated null distributions of six test statistics and the line in each subgraph is the probability density function of $N(0,1)$ when the observations follow multivariate t_3 distribution. And (a)-(f) are corresponding to SD, CQ, RHT, GCT_2, DLRT and RHDT respectively.

t_3 distribution, it is obvious that the proposed RHDT still maintains the standard normal distribution. But other competitors can not be fitted well by $N(0,1)$.

For comparison of type I error rates, we generate observations from multivariate normal distribution and multivariate t_3 distribution with different combinations of $(m,n) \times p = \{(3,3) \text{ or } (10,15) \text{ or } (15,15)\} \times \{100 \text{ or } 300 \text{ or } 500\}$. Table 1 shows the simulated results across the four dependence structures for normal data. As we can see in Table 1, the simulated type I error rates of RHDT test are very close to the significance level ($\alpha = 0.05$) under four dependence structures. It is obvious that the tests SD, RHT and GCT_1 can not control the type I error rate when sample sizes are extremely small such as $(m,n) = (3,3)$.

Table 2 shows the sizes of all the tests across the four dependence structures when observations follow multivariate t_3 distribution. The simulated results demonstrate that test RHDT can control the type I error rate pretty well under all scenarios, and the performance of test DLRT is similar to RHDT's. Table 2 also shows that the sizes of SD, CQ and RHT are far less than 0.05 when the sample size is small. It means that these three methods are too conservative for heavy-tailed data. However, under the LR dependence structures, the type I error rates of the SD test reach 0.4 when $(m,n) = (3,3)$. Finally, neither GCT_1 nor GCT_2 can control the type I error for heavy-tailed data.

4.2. Statistical power

For power comparison, the observations are generated from multivariate normal distribution and multivariate t_3 distribution under balanced cases $(m,n) = (15,15)$ and $(m,n) = (50,50)$, and under imbalanced case $(m,n) = (10,15)$. Following the same settings in Hu et al. (2019), we let the mean vectors $\boldsymbol{\mu}_1 = \mathbf{0}_{p\times 1}$ and $\boldsymbol{\mu}_2 = (\mu_{21},\ldots,\mu_{2k},0,\ldots,0)^{\mathrm{T}}$ with $\mu_{21} = \cdots = \mu_{2k} = \eta\sqrt{\chi^2(5)/5}$ and $p = 500$. The proportion $\tau = k/p \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and $\eta = 0.25$ are used to control the differences between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. We mainly focus on how the powers of all methods change with τ under different dependence structures. The simulated results under balanced and imbalanced cases are shown in Figs. 3–5 by 2000 simulations. In each figure, the subgraphs (a)-(d) are the results obtained from multivariate normal distribution, and the subgraphs (e)-(h) are for multivariate t_3 distribution. The four subgraphs in each column correspond to the simulated results under the IND, AR = 0.3, AR = 0.6 and LR dependence structures in turn. Since GCT_1 and GCT_2 can not control the type I error, the simulated powers of GCT_1 and GCT_2 are not shown.

In Fig. 3, we can see that, for normal data under balanced scenarios, the powers of all methods increase to 1 as τ increases to 1. When $\tau \neq 0$, the powers of RHDT are the highest among all methods under all dependence structures. And

Table 1
The sizes of the tests under four dependence structures when observations follow multivariate normal distribution.

	p	(m,n)	SD	CQ	RHT	GCT_1	GCT_2	DLRT	RHDT
IND	100	(3,3)	0.406	0.088	0.112	0.607	0.106	0.060	0.068
		(10,15)	0.043	0.054	0.059	0.080	0.081	0.056	0.067
		(15,15)	0.046	0.059	0.067	0.071	0.097	0.059	0.060
	300	(3,3)	0.366	0.083	0.102	0.962	0.059	0.049	0.054
		(10,15)	0.033	0.050	0.052	0.180	0.071	0.057	0.052
		(15,15)	0.027	0.053	0.053	0.106	0.063	0.050	0.051
	500	(3,3)	0.354	0.084	0.101	0.982	0.036	0.051	0.051
		(10,15)	0.020	0.052	0.056	0.280	0.071	0.054	0.053
		(15,15)	0.029	0.051	0.056	0.155	0.066	0.048	0.052
AR = 0.3	100	(3,3)	0.398	0.079	0.101	0.552	0.115	0.059	0.054
		(10,15)	0.038	0.050	0.057	0.080	0.100	0.052	0.053
		(15,15)	0.041	0.047	0.061	0.071	0.099	0.059	0.067
	300	(3,3)	0.364	0.086	0.098	0.951	0.064	0.056	0.056
		(10,15)	0.029	0.040	0.044	0.148	0.071	0.057	0.057
		(15,15)	0.025	0.045	0.057	0.095	0.079	0.062	0.055
	500	(3,3)	0.325	0.086	0.101	0.983	0.042	0.049	0.045
		(10,15)	0.020	0.047	0.050	0.234	0.072	0.055	0.058
		(15,15)	0.029	0.057	0.057	0.144	0.081	0.064	0.059
AR = 0.6	100	(3,3)	0.359	0.090	0.107	0.408	0.129	0.070	0.060
		(10,15)	0.042	0.052	0.059	0.097	0.133	0.084	0.065
		(15,15)	0.033	0.046	0.051	0.095	0.139	0.080	0.055
	300	(3,3)	0.323	0.084	0.093	0.913	0.085	0.084	0.070
		(10,15)	0.028	0.054	0.057	0.143	0.114	0.085	0.067
		(15,15)	0.026	0.053	0.062	0.098	0.106	0.082	0.066
	500	(3,3)	0.288	0.076	0.091	0.981	0.062	0.068	0.059
		(10,15)	0.020	0.052	0.056	0.195	0.098	0.078	0.071
		(15,15)	0.019	0.048	0.053	0.121	0.093	0.080	0.050
LR	100	(3,3)	0.390	0.071	0.096	0.577	0.106	0.059	0.063
		(10,15)	0.046	0.051	0.057	0.081	0.099	0.062	0.065
		(15,15)	0.042	0.049	0.048	0.071	0.093	0.061	0.070
	300	(3,3)	0.342	0.085	0.099	0.962	0.065	0.058	0.057
		(10,15)	0.036	0.063	0.063	0.167	0.085	0.068	0.069
		(15,15)	0.030	0.051	0.054	0.097	0.086	0.058	0.057
	500	(3,3)	0.323	0.081	0.097	0.987	0.046	0.065	0.050
		(10,15)	0.026	0.047	0.051	0.268	0.067	0.054	0.067
		(15,15)	0.022	0.048	0.044	0.164	0.062	0.053	0.053

RHDT can control the power at 0.05 nearly when $\tau = 0$. The test DLRT has suboptimal performance, and followed by RHT, CQ and SD under all scenarios. Also in Fig. 3, for the multivariate t_3 distribution, the test RHDT also has the superior performance than others when $\tau \neq 0$. Meanwhile, Fig. 3 shows that, under all dependence structures, tests SD and CQ are almost unable to detect difference in mean vectors even when $\tau = 1$. In conclusion, the proposed RHDT has the most stable performance among all five methods, and can also work well for heavy-tailed data such as multivariate t_3 distribution. Fig. 4 indicates there are similar conclusions under the imbalanced scenarios as that under balanced scenarios. The proposed RHDT exhibits superior to SD, CQ, RHT and GCT_2 in many cases. Fig. 5 presents the power curves of the five methods. In the case of multivariate normal distribution, the performance of CQ test will be improved compared with that for the case of small sample size. We also note that its power is close to RHDT test, whereas SD test has the best performance. In contrast, for multivariate t_3 distribution, the performance of CQ test does not improve significantly along with the sample size, which coincides with the heavy-tailed simulation on Pareto(1.5, 1) distribution in Gregory et al. (2015).

Remark 1. For the bandwidth selection, Corollary 1 in Mcmurry and Politis (2010) showed that if $|\gamma_i| = O(p^{-d})$ for some $d > 1$, the optimal bandwidth should satisfy $L = O(p^{(q-1)/(dq)})$ with $1 < q \leq 2$, where γ_i is the estimator of the autocovariance the sequence V_m^j. It can be verified that $0 < (q-1)/(dq) < 1/2$ for (q,d) mentioned above, we thus choose the bandwidth $L = O(p^{1/3})$. To examine how the bandwidth affects the performance of the test statistic, we have conducted a new simulation using multivariate normal distribution with different combinations of $(m,n) \times p \times L = (15, 15) \times 500 \times \{[0.4p^{1/3}], [0.5p^{1/3}], [0.6p^{1/3}], [0.7p^{1/3}], [0.8p^{1/3}], [p^{1/3}], [1.2p^{1/3}]\}$, where $[x]$ is the largest integer smaller than or equal to x. Fig. 6 shows the power curves of RHDT, DLRT and GCT_2 tests as functions of the bandwidth. From the results, it

Table 2
The sizes of the tests under four dependence structures when observations follow multivariate t_3 distribution.

	p	(m,n)	SD	CQ	RHT	GCT_1	GCT_2	DLRT	RHDT
IND	100	(3,3)	0.047	0.002	0.014	0.530	0.179	0.044	0.054
		(10,15)	0.001	0.000	0.008	0.069	0.302	0.092	0.059
		(15,15)	0.000	0.002	0.007	0.072	0.255	0.047	0.066
	300	(3,3)	0.015	0.001	0.006	0.947	0.150	0.076	0.053
		(10,15)	0.000	0.000	0.004	0.143	0.397	0.221	0.070
		(15,15)	0.000	0.000	0.001	0.112	0.358	0.041	0.059
	500	(3,3)	0.006	0.000	0.003	0.982	0.171	0.115	0.062
		(10,15)	0.000	0.000	0.003	0.221	0.457	0.316	0.074
		(15,15)	0.000	0.000	0.000	0.137	0.432	0.044	0.056
AR = 0.3	100	(3,3)	0.069	0.006	0.013	0.509	0.197	0.048	0.064
		(10,15)	0.002	0.003	0.012	0.077	0.291	0.103	0.068
		(15,15)	0.000	0.001	0.011	0.070	0.267	0.041	0.053
	300	(3,3)	0.015	0.001	0.004	0.942	0.184	0.076	0.067
		(10,15)	0.000	0.000	0.009	0.141	0.381	0.229	0.076
		(15,15)	0.000	0.000	0.001	0.095	0.344	0.041	0.048
	500	(3,3)	0.005	0.000	0.002	0.983	0.196	0.117	0.061
		(10,15)	0.000	0.000	0.004	0.225	0.446	0.302	0.080
		(15,15)	0.000	0.000	0.000	0.136	0.418	0.049	0.051
AR = 0.6	100	(3,3)	0.062	0.009	0.022	0.401	0.191	0.060	0.065
		(10,15)	0.003	0.005	0.021	0.108	0.320	0.123	0.060
		(15,15)	0.001	0.004	0.021	0.090	0.276	0.067	0.067
	300	(3,3)	0.018	0.001	0.009	0.887	0.191	0.091	0.065
		(10,15)	0.000	0.001	0.005	0.116	0.377	0.202	0.074
		(15,15)	0.001	0.001	0.002	0.101	0.340	0.068	0.062
	500	(3,3)	0.010	0.000	0.009	0.973	0.188	0.122	0.074
		(10,15)	0.000	0.001	0.006	0.169	0.403	0.268	0.077
		(15,15)	0.000	0.000	0.000	0.124	0.393	0.079	0.062
LR	100	(3,3)	0.403	0.028	0.031	0.516	0.136	0.040	0.056
		(10,15)	0.030	0.015	0.035	0.077	0.276	0.051	0.067
		(15,15)	0.024	0.021	0.040	0.067	0.278	0.043	0.068
	300	(3,3)	0.430	0.020	0.026	0.958	0.123	0.060	0.058
		(10,15)	0.022	0.013	0.022	0.135	0.480	0.049	0.053
		(15,15)	0.017	0.014	0.021	0.106	0.450	0.051	0.070
	500	(3,3)	0.446	0.017	0.020	0.987	0.099	0.071	0.057
		(10,15)	0.013	0.011	0.020	0.232	0.624	0.051	0.056
		(15,15)	0.010	0.013	0.021	0.150	0.580	0.043	0.056

is evident that their performance is quite stable and not sensitive to the choice of the bandwidth. For the simulations in Section 4, we choose bandwidth $L = 5$ following the setting in Hu et al. (2019).

5. Application to real data analysis

As described in the introduction, study of gene expression data can lead to important insights into diabetes biology. In this section, we apply the proposed method to a human peripheral blood mononuclear cells (PBMCs) gene expression data set. The gene data set is available from the Gene Expression Omnibus online pathway databases (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE142153). The data set contains cells from 10 healthy controls and 23 patients with diabetic nephropathy. Fig. 7 indicates that the data set can not show normal distribution in each dimension. Considering that GCT_1 can not control the type I error rate when sample sizes are extremely small, we apply other tests to test the equality of global PBMCs gene expression between healthy controls and patients with diabetic nephropathy.

Following the setting in Hu et al. (2019), in order to compare the performance of the tests, we screen the most significant p genes by performing two-sample t tests for each dimension, and then use $p = 100$ to calculate the experience ability. The empirical powers for the tests applied to both the healthy controls and patients with diabetic nephropathy are given in Table 3 on 1000 simulation runs. In each run, we randomly select n_1 and n_2 samples from healthy controls and patients with diabetic nephropathy without replacement, respectively. When the sample size is small like $n_1 = 3$ and $n_2 = 3$, the empirical power of the RHDT test is higher than the SQ, CQ, RHT and GCT_2 tests. And the DLRT test also works well. The results are similar in unbalanced samples when $n_1 = 3$ and $n_2 = 5$. With the increase of sample size, all the tests work well under the setting with $n_1 = 5$ and $n_2 = 5$.

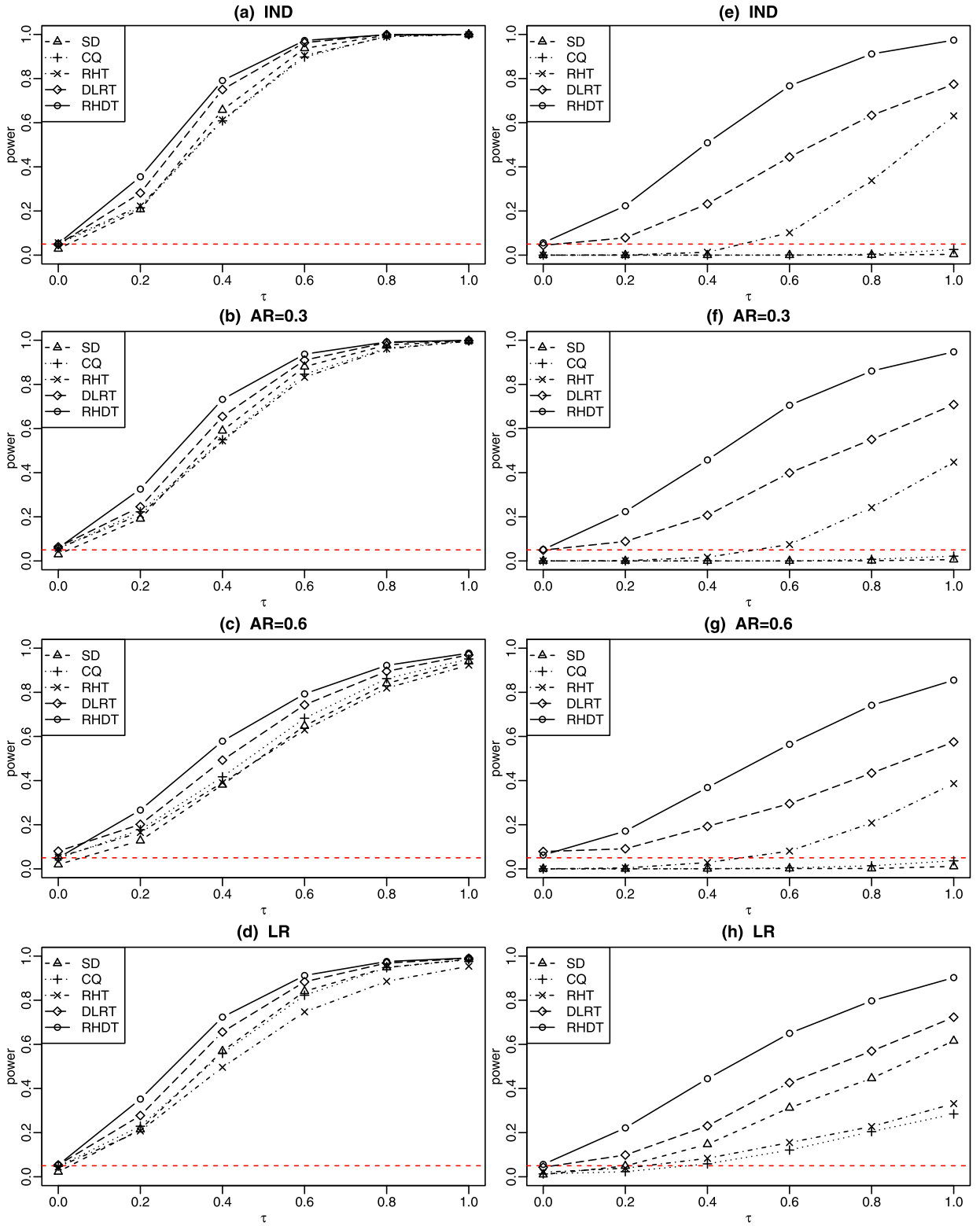

Fig. 3. Power curves of the five methods against the proportion of nonzero mean differences τ under IND, AR $= 0.3$, AR $= 0.6$ and LR dependence with $(m, n) = (15, 15)$. The subgraphs (a)-(d) and (e)-(h) present the simulated results from multivariate normal distribution and multivariate t_3 distribution, respectively.

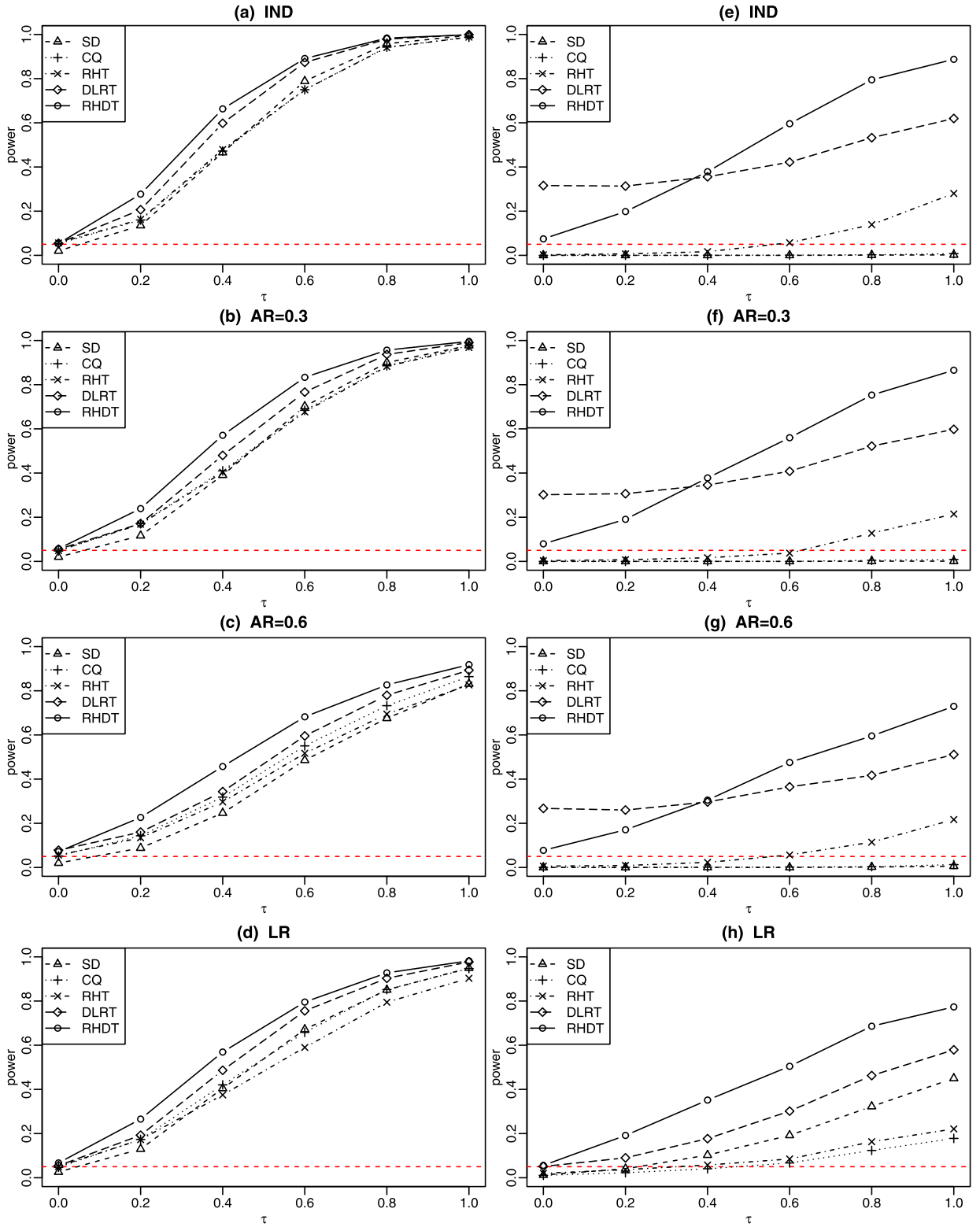

Fig. 4. Power curves of the five methods against the proportion of nonzero mean differences τ under IND, AR $= 0.3$, AR $= 0.6$ and LR dependence with $(m, n) = (10, 15)$. The subgraphs (a)-(d) and (e)-(h) present the simulated results from multivariate normal distribution and multivariate t_3 distribution, respectively.

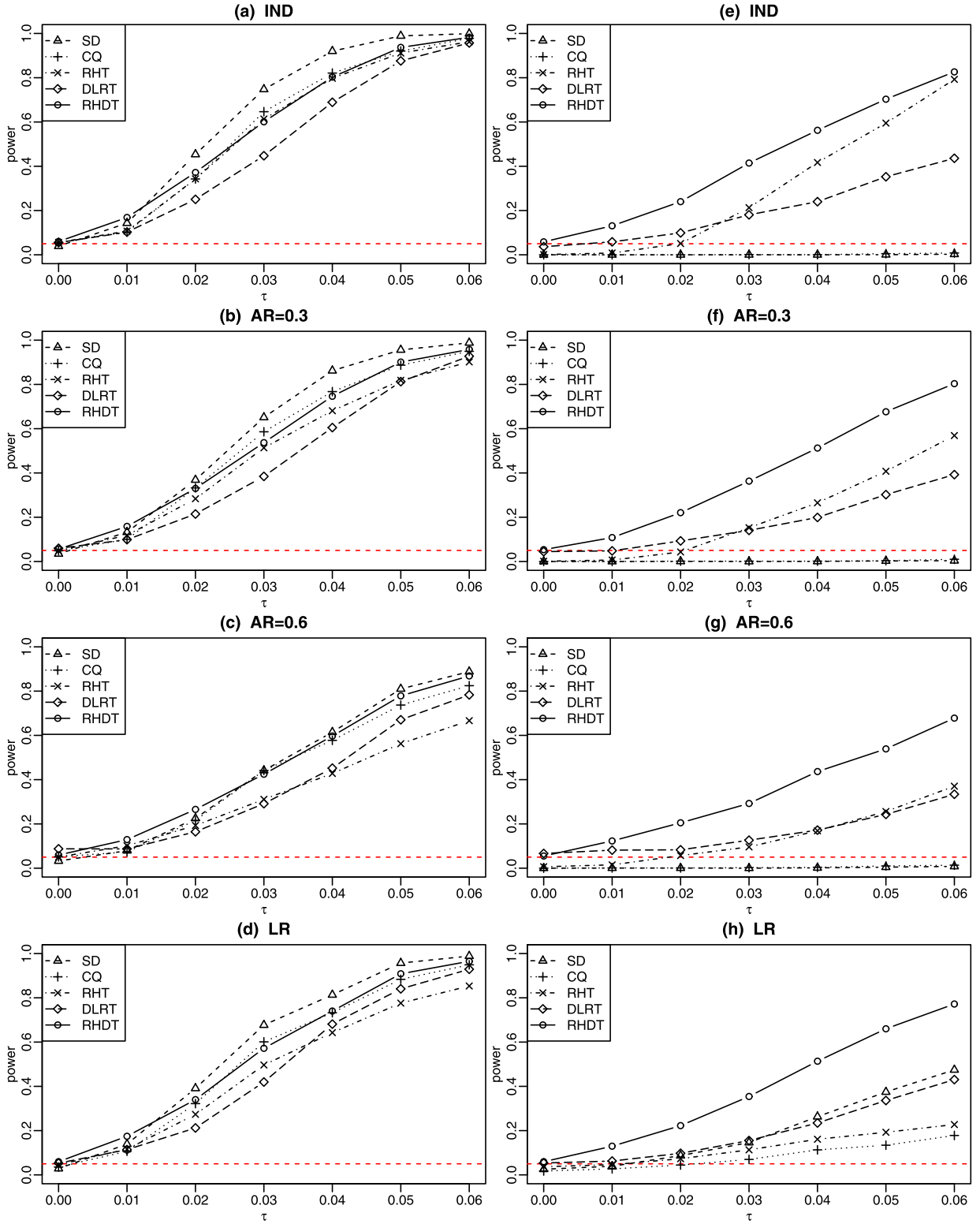

Fig. 5. Power curves of the five methods against the proportion of nonzero mean differences τ under IND, AR $= 0.3$, AR $= 0.6$ and LR dependence with $(m, n) = (50, 50)$. The subgraphs (a)-(d) and (e)-(h) present the simulated results from multivariate normal distribution and multivariate t_3 distribution, respectively.

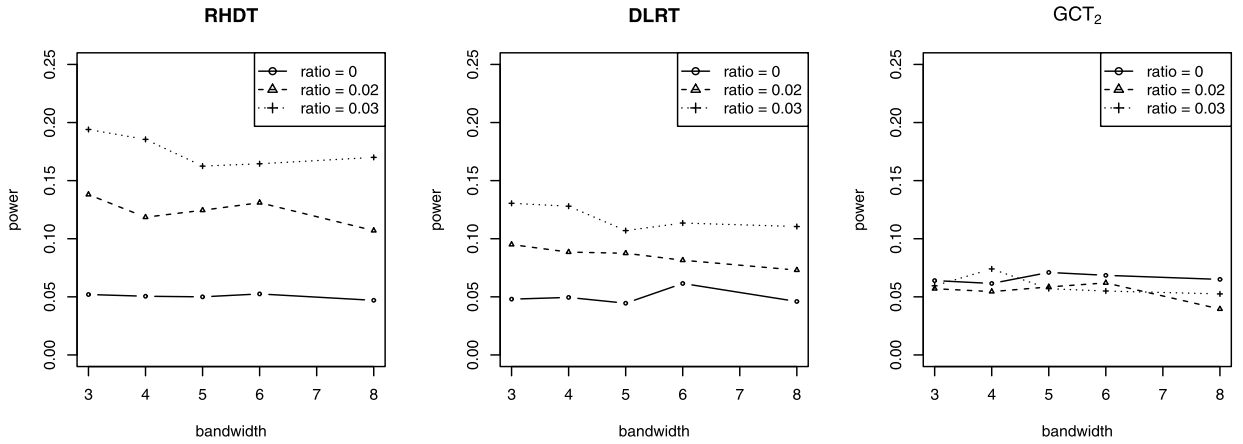

Fig. 6. Power curves of three methods against different bandwidths under IND dependence with $(m, n) = (15, 15)$.

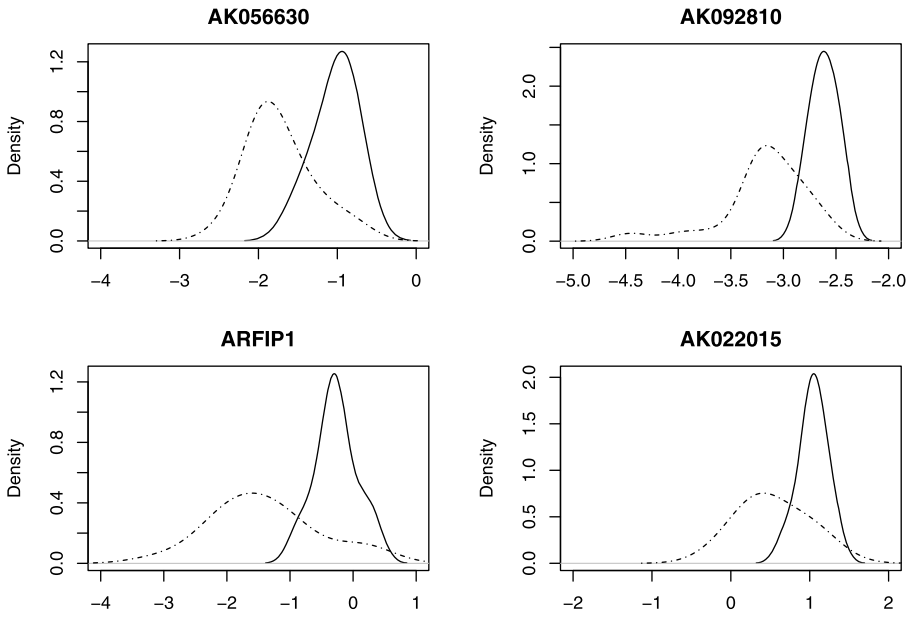

Fig. 7. The density curves of healthy controls (solid line) and patients with diabetic nephropathy (dot-dash line) on the top 4 significant gene sets AK056630, AK092810, ARFIP1 and AK022015.

Table 3
Data analysis results for PBMCs gene expression data set with $p = 100$.

n_1	n_2	SD	CQ	RHT	GCT_2	DLRT	RHDT
3	3	0.777	0.771	0.781	0.767	0.929	0.934
3	5	0.757	0.892	0.763	0.926	0.971	0.984
5	5	0.951	0.973	0.927	0.996	1.000	1.000

6. Conclusion

Hotelling's T^2 test is the most commonly used method for testing mean vectors in multivariate statistics. It is known however that, when the dimension of the data exceeds the number of samples, Hotelling's T^2 test will no longer be applicable due to the singularity of the sample covariance matrix. Some modified methods have made efforts to replace the inverse of the sample covariance matrix S_W^{-1} in Hotelling's T^2 test statistics, while Gregory et al. (2015) used the mean of the squares of Student's t test statistics in all p components. Hu et al. (2019) proposed a likelihood ratio test under the assumption of normal data.

In this paper, we propose a rank-based mean test, which combines the results of nonparametric tests in each dimension. The method can be applied to one-sample or two-sample in high-dimensional situation, which is based on the Wilcoxon signed-rank test or the WMW test. Our new method has no requirements for data distribution and has a wider scope of application. We also establish the asymptotic normality of the proposed test statistic under the null and local alternative hypotheses. In simulation section, the proposed RHDT test can work as well as other methods for normal data, and has the most efficient performance when observations follow multivariate t_3 distribution. In real data analysis, the proposed test works also equally.

Acknowledgements

The authors thank the editor, the associate editor, and two reviewers for their constructive comments that led to a substantial improvement of the paper. Wangli Xu's research was supported by Beijing Natural Science Foundation (No Z200001), National Natural Science Foundation of China (No 11971478) and Public Computing Cloud Platform, Renmin University of China. Tiejun Tong's research was supported by the General Research Funds (HKBU12303918, HKBU12303421), the Initiation Grants for Faculty Niche Research Areas (RC-FNRA-IG/20-21/SCI/03) of Hong Kong Baptist University, and the National Natural Science Foundation of China (1207010822).

Appendix

The proofs of Theorems 1–4 are presented in Appendices A–D, respectively.

Appendix A. Proof of Theorem 1

To investigate the asymptotic property of W_m under H_0 in (1), we first calculate the expectation and variance of W_m^j. Recalling the definition of W_m^j, we have $E(W_m^j) = \mathrm{var}(U_m^j)$. It indicates that $E(W_m^j) = m(m+1)(2m+1)/24$. For the variance of W_m^j, we have

$$\begin{aligned}\mathrm{var}(W_m^j) &= E\{(W_m^j)^2\} - E^2(W_m^j)\\ &= E\{(U_m^j)^4\} - 4E\{(U_m^j)^3\}E(U_m^j) + 8E\{(U_m^j)^2\}E^2(U_m^j) - 4E^4(U_m^j) - E^2\{(U_m^j)^2\}.\end{aligned}$$

We have the result in (5) by algebraic calculations.

Based on the definitions of ξ_m and η_m, $V_m^j = (W_m^j - \xi_m)/\sqrt{\eta_m}$ implies the sequence $\{V_m^j, j = 1, \ldots, p\}$ is a series of random variables with mean zero and variance one. Invoking the central limit theorem under the strong mixing conditions (see Corollary 5.1 in Hall and Heyde (1980)), when sequence $\{V_m^j, j = 1, \ldots, p\}$ satisfies Conditions (C1) and (C2), we only need to show that $E|V_m^j|^{2+\delta} < \infty$ for any m and fixed $0 < \delta < \infty$ to prove

$$\frac{\sum_{j=1}^p V_m^j}{\tau_1\sqrt{p}} \xrightarrow{D} N(0,1), \quad \text{as } p \to \infty. \tag{A.1}$$

Based on the definition of V_m^j, we have

$$E|V_m^j|^{2+\delta} = \frac{E|W_m^j - \xi_m|^{2+\delta}}{\eta_m^{(2+\delta)/2}}.$$

Noting that the Wilcoxon signed test statistic $U_m^j \le m(m+1)/2$, we have $W_m^j \le \{m(m+1)\}^2/16$. Moreover, we have $E|W_m^j - \xi_m|^{2+\delta} \le |\{m(m+1)\}^2/16 - \xi_m|^{2+\delta}$. Thus, for any fixed $m \ge 2$, we have

$$E|V_m^j|^{2+\delta} \le \left[\frac{\{m(m^2-1)(3m+2)\}}{48\sqrt{\eta_m}}\right]^{2+\delta} < \infty.$$

As a result, (A.1) has been proved. Moreover, we have

$$\frac{\sum_{j=1}^p V_m^j}{\tau_1\sqrt{p}} = \frac{\sqrt{p}(W_m - \xi_m)}{\tau_1\sqrt{\eta_m}} \xrightarrow{D} N(0,1), \quad \text{as } (m,p) \to \infty.$$

This completes the proof of Theorem 1. □

Appendix B. Proof of Theorem 2

To prove Theorem 2, we first calculate the expectation of W_m^j under the local alternative (9). For fixed j, we suppose $\{X_{(k)j}, k = 1, \ldots, m\}$ is the order sample for $\{X_{ij}, i = 1, \ldots, m\}$ ranking by the absolute value $|X_{ij}|$. The Wilcoxon signed-rank statistic for the j-th dimension can be equivalently rewritten as

$$U_m^j = \sum_{k=1}^{m} kI(X_{(k)j} > 0), \tag{B.1}$$

and we have $E\{I(X_{(k)j} > 0)\} = \Pr(X_{(k)j} > 0)$. To ease subsequent illustration, we define $\Pr(X_{(k)j} > 0) \overset{\text{def}}{=} c_{kj}$ and $A_j(x) \overset{\text{def}}{=} F_j(x - \delta_{1j}/\sqrt{m}) - F_j(-x - \delta_{1j}/\sqrt{m})$. By algebraic derivation, we have

$$c_{kj} = \frac{m!}{(k-1)!(m-k)!} \int_0^{\infty} \{A_j(x)\}^{k-1} \left\{1 - A_j(x)\right\}^{m-k} \mathrm{d}\{F_j(x - \delta_{1j}/\sqrt{m})\}.$$

For the expectation of W_m^j under the local alternative H_1' in (9), by simple transformation, we have $E(W_m^j) = \mathrm{var}(U_m^j) + \left\{E(U_m^j) - m(m+1)/4\right\}^2$. In the following, we calculate the expectation and variance of U_m^j under the local alternative H_1' in (9). Based on the transformation of U_m^j in (B.1), it can be derived that

$$\begin{aligned}
E(U_m^j) &= \sum_{k=1}^{m} kE\{I(X_{(k)j} > 0)\} = \sum_{k=1}^{m} kc_{kj}, \\
E(U_m^j)^2 &= E\Big\{ \sum_{k=1}^{m} k^2 I(X_{(k)j} > 0) + \sum_{k_1=1}^{m} \sum_{k_2 \neq k_1}^{m} k_1 k_2 I(X_{(k_1)j} > 0) I(X_{(k_2)j} > 0) \Big\} \\
&= \sum_{k=1}^{m} k^2 c_{kj} + \sum_{k_1=1}^{m} \sum_{k_2 \neq k_1}^{m} k_1 k_2 c_{k_1 j} c_{k_2 j}
\end{aligned}$$

and $\mathrm{var}(U_m^j) = E(U_m^j)^2 - E^2(U_m^j) = \sum_{k=1}^{m} k^2 c_{kj}(1 - c_{kj})$. Furthermore, it follows that

$$E(W_m^j) = \sum_{k=1}^{m} k^2 c_{kj}(1 - c_{kj}) + \Big\{ \sum_{k=1}^{m} kc_{kj} - \frac{m(m+1)}{4} \Big\}^2. \tag{B.2}$$

Next, we simplify $E(W_m^j)$ in (B.2) by simplifying c_{kj}. Based on Taylor expansions with first order for $F_j(x - \delta_{1j}/\sqrt{m})$ and $F_j(-x - \delta_{1j}/\sqrt{m})$, we have

$$A_j(x) = F_j(x) - \frac{\delta_{1j}}{\sqrt{m}} f_j(x) - F_j(-x) + \frac{\delta_{1j}}{\sqrt{m}} f_j(-x) + O(m^{-1}).$$

The symmetry of distribution $F(\cdot)$ implies that $F_j(-x) = 1 - F_j(x)$ and $f_j(-x) = f_j(x)$, thus we have $A_j(x) = 2F_j(x) - 1 + O(m^{-1})$ and the last term $O(m^{-1})$ can be omitted when $m \to \infty$. Moreover, c_{kj} can be simplified as

$$\begin{aligned}
c_{kj} &= k\binom{m}{k} \int_0^{\infty} \{2F_j(x) - 1\}^{k-1} \left\{2 - 2F_j(x)\right\}^{m-k} d\{F_j(x) - \frac{\delta_{1j}}{\sqrt{m}} f_j(x)\} \\
&= \frac{1}{2} - \frac{\delta_{1j} k}{\sqrt{m}} \binom{m}{k} \int_0^{\infty} \{2F_j(x) - 1\}^{k-1} \{2 - 2F_j(x)\}^{m-k} df_j(x) + R_m \\
&= \frac{1}{2} - B_{jk} + R_m,
\end{aligned}$$

where

$$B_{jk} \overset{\text{def}}{=} \frac{\delta_{1j} k}{\sqrt{m}} \binom{m}{k} \int_0^{\infty} \{2F_j(x) - 1\}^{k-1} \{2 - 2F_j(x)\}^{m-k} df_j(x),$$

and the term R_m can be omitted when $m \to \infty$. The remaining terms involved in the below are also represented by R_m, but the actual meaning of each line is different. Furthermore, we have $c_{kj}(1-c_{kj}) = 1/4 - B_{jk}^2 + R_m$. The first term of $E(W_m^j)$ in equation (B.2) is

$$\begin{aligned}\sum_{k=1}^{m} k^2 c_{kj}(1-c_{kj}) &= \sum_{k=1}^{m} \frac{k^2}{4} - \sum_{k=1}^{m} k^2 (B_{jk})^2 + R_m \\ &= \frac{m(m+1)(2m+1)}{24} - \sum_{k=1}^{m} k^2 (B_{jk})^2 + R_m,\end{aligned}$$

where $R_m = o(m^3)$. In what follows, we prove that $\sum_{k=1}^{m} k^2 (B_{jk})^2 = o(m^3)$. Notice that $-1/2 \le B_{jk} < 0$ and $\sum_{k=1}^{m} k^2 B_{jk}$ is equal to

$$\begin{aligned}&\frac{\delta_{1j}}{\sqrt{m}} \int_0^{\infty} \sum_{k=1}^{m} k^3 \binom{m}{k} \{2F_j(x)-1\}^{k-1} \{2-2F_j(x)\}^{m-k} df_j(x) \\ &= \frac{\delta_{1j}}{\sqrt{m}} \int_0^{\infty} [m(m-1)(m-2)\{2F_j(x)-1\}^2 + 3m(m-1)\{2F_j(x)-1\} + m] df_j(x) \\ &= O(m^{5/2}).\end{aligned}$$

It follows that $\sum_{k=1}^{m} k^2 B_{jk} < 0$ and $\sum_{k=1}^{m} k^2 |B_{jk}| = o(m^3)$. Since $|B_{jk}| < 1/2$, we have

$$\sum_{k=1}^{m} k^2 B_{jk}^2 = \sum_{k=1}^{m} k^2 |B_{jk}|^2 < \frac{1}{2} \sum_{k=1}^{m} k^2 |B_{jk}| = o(m^3).$$

Thus, the proof of $\sum_{k=1}^{m} k^2 (B_{jk})^2 = o(m^3)$ has been completed. Furthermore, we can obtain that

$$\sum_{k=1}^{m} k^2 c_{kj}(1-c_{kj}) = \frac{m(m+1)(2m+1)}{24} + o(m^3).$$

Meanwhile, the term $\sum_{k=1}^{m} k c_{kj}$ in (B.2) can be simplified as

$$\begin{aligned}&\sum_{k=1}^{m} \frac{k}{2} - \sum_{k=1}^{m} k B_{jk} + R_m \\ &= \frac{m(m+1)}{4} - \frac{\delta_{1j}}{\sqrt{m}} \int_0^{\infty} \sum_{k=1}^{m} k^2 \binom{m}{k} \{2F_j(x)-1\}^{k-1} \{2-2F_j(x)\}^{m-k} df_j(x) + R_m \\ &= \frac{m(m+1)}{4} - \frac{\delta_{1j}}{\sqrt{m}} \int_0^{\infty} [m(m-1)\{2F_j(x)-1\} + m] df_j(x) + R_m.\end{aligned}$$

Thus, we have

$$E(W_m^j) = \frac{m(m+1)(2m+1)}{24} + \Big(\int_0^{\infty} \{2F_j(x)-1\} df_j(x) \Big)^2 \delta_{1j}^2 m^3 + o(m^3).$$

In the following, we derive the asymptotic property for W_m under the local alternative hypothesis in (9). First of all, we have

$$\begin{aligned}\frac{\sqrt{p}(W_m - \xi_m)}{\sqrt{\eta_m}} &= \frac{\sqrt{p}\{W_m - \xi_m - p^{-1}\sum_{j=1}^{p} E(W_m^j) + p^{-1}\sum_{j=1}^{p} E(W_m^j)\}}{\sqrt{\eta_m}} \\ &= \frac{\sqrt{p}\{W_m - p^{-1}\sum_{j=1}^{p} E(W_m^j)\}}{\sqrt{\eta_m}} + \frac{\sqrt{p}\{p^{-1}\sum_{j=1}^{p} E(W_m^j) - \xi_m\}}{\sqrt{\eta_m}}.\end{aligned}$$

For the first item in the equality above, we have

$$\frac{\sqrt{p}\{W_m - p^{-1}\sum_{j=1}^{p} E(W_m^j)\}}{\sqrt{\eta_m}} = \frac{\sum_{j=1}^{p}\{V_m^j - E(V_m^j)\}}{\sqrt{p}}$$
$$= \frac{\sum_{j=1}^{p}\{V_m^j - E(V_m^j)\}}{\sqrt{\mathrm{var}(\sum_{j=1}^{p} V_m^j)}} \frac{\sqrt{\mathrm{var}(\sum_{j=1}^{p} V_m^j)}}{\sqrt{p}}.$$

Under Conditions (C1)-(C2) and the definition of τ_1^2, we have $p^{-1}\,\mathrm{var}(\sum_{j=1}^{p} V_m^j) = \tau_1^2 - 1 + \gamma_1(0)$. Recalling the definition of Δ_1 in (11), we have

$$\frac{\sqrt{p}(W_m - \xi_m)}{\sqrt{\eta_m}} \xrightarrow{D} C_1 \cdot N(0, \tau_1^2) + \frac{\Delta_1}{\sqrt{p}}, \quad \text{as } p \to \infty,$$

where C_1 is defined in Theorem 2. Finally, this yields the asymptotic power function of Z_m as

$$\beta(Z_m) = \Pr\left\{\frac{\sqrt{p}(W_m - \xi_m)}{\tau_1\sqrt{\eta_m}} > Z_{1-\alpha} \mid H_1' \text{ is true}\right\}$$
$$= 1 - \Phi\left\{C_1^{-1}\left(Z_{1-\alpha} - \frac{\Delta_1}{\tau_1\sqrt{p}}\right)\right\}. \qquad \square$$

Appendix C. Proof of Theorem 3

In parallel to the proof of Theorem 1, the derivation of the asymptotic distribution of $W_{m,n}$ under H_0 in (2) can be divided into two parts. At first, we calculate the expectation and variance of $W_{m,n}^j$. And then, we verify that the sequence $\{V_{m,n}^j,\ j = 1, \ldots, p\}$ satisfies the condition of central limit theorem under the strong mixing conditions.

To calculate the expectation and variance of $W_{m,n}^j$, we consider the probability distribution of ranks R_{ij}^{xy}. For ease of syntax and without loss of generality, we ignore the subscript j and the superscript xy of R_{ij}^{xy} hereafter since the distributions of R_{ij}^{xy} are same for $j = 1, \ldots, p$. For a fixed j and mutually different indexes i_1, i_2, i_3, i_4, we let $\{R_{i_1}, R_{i_2}, R_{i_3}, R_{i_4}\}$ denote the ranks of $\{X_{i_1 j}, X_{i_2 j}, X_{i_3 j}, X_{i_4 j}\}$ among $\{X_{1j}, \ldots, X_{mj}, Y_{1j}, \ldots Y_{nj}\}$. Let $r_1, r_2, r_3, r_4 = 1, 2, \ldots, N$ and be mutually different, under the null hypothesis that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, we have

$$\Pr(R_{i_1} = r_1, \ldots, R_{i_s} = r_s) = 1/\prod_{t=1}^{s}(N - t + 1), \quad s = 1, 2, 3, 4.$$

Thus, we can obtain that $E(R_{i_1}) = (N+1)/2$, $E(R_{i_1}^2) = (N+1)(2N+1)/6$ and $E(R_{i_1}R_{i_2}) = (N+1)(3N+2)/12$. Recalling the definition of $W_{m,n}^j$, we have $E(W_{m,n}^j) = \mathrm{var}\left(\sum_{i_1=1}^{m} R_{i_1}\right)$. Furthermore,

$$E(W_{m,n}^j) = \sum_{i_1=1}^{m} E(R_{i_1}^2) + \sum_{i_1=1}^{m}\sum_{i_2\neq i_1}^{m} E(R_{i_1}R_{i_2}) - \left\{mE(R_{i_1})\right\}^2 = \frac{mn(N+1)}{12}.$$

On the other hand, we have $E(R_{i_1}^3) = N(N+1)^2/4$, $E(R_{i_1}R_{i_2}^2) = N(N+1)^2/6$, $E(R_{i_1}R_{i_2}R_{i_3}) = N(N+1)^2/8$ and

$$E(R_{i_1}^4) = (N+1)(2N+1)(3N^2+3N-1)/30,$$
$$E(R_{i_1}R_{i_2}^3) = (N+1)(15N^3+21N^2-4)/120,$$
$$E(R_{i_1}^2R_{i_2}^2) = (N+1)(2N-1)(2N+1)(5N+6)/180,$$
$$E(R_{i_1}R_{i_2}R_{i_3}^2) = (N+1)(30N^3+35N^2-11N-12)/360$$
$$\text{and } E(R_{i_1}R_{i_2}R_{i_3}R_{i_4}) = (N+1)(15N^3+15N^2-10N-8)/240.$$

Furthermore, we can derive that

$$E\Big(\sum_{i=1}^{m} R_i\Big)^3 = N(N+1)^2m^2(m+1)/8 \quad \text{and}$$
$$E\Big(\sum_{i=1}^{m} R_i\Big)^4 = m(N+1)\{mN(5N^2+9N+2) + 2m^2N(15N^2+25N+8)$$
$$+m^3(15N^3+15N^2-10N-8) - 2N^2(N+1)\}/240.$$

For the variance of $W^j_{m,n}$, we have

$$\begin{aligned}\mathrm{var}(W^j_{m,n}) &= E\Big\{\sum_{i=1}^{m} R_i - \frac{m(N+1)}{2}\Big\}^4 - \{E(W^j_{m,n})\}^2 \\ &= E\Big(\sum_{i=1}^{m} R_i\Big)^4 - 2m(N+1)E\Big(\sum_{i=1}^{m} R_i\Big)^3 + \frac{3m^2(N+1)^2}{2}E\Big(\sum_{i=1}^{m} R_i\Big)^2 \\ &\quad - \frac{m^3(N+1)^3}{2}E\Big(\sum_{i=1}^{m} R_i\Big) + \frac{m^4(N+1)^4}{16} - \frac{m^2n^2(N+1)^2}{144}.\end{aligned}$$

By algebraic calculation, we have $\mathrm{var}(W^j_{m,n}) = \{mn(5N+8) - 3N(N+1)\}(N+1)mn/360$.

Below, we verify the sequence $\{V^j_{m,n},\ j = 1, \ldots, p\}$ satisfies the condition of central limit theorem under the strong mixing condition. Namely, we need to prove that $E|V^j_{m,n}|^{2+\delta} < \infty$ for any m, n and fixed $0 < \delta < \infty$. Based on the definition of $V^j_{m,n}$, we have

$$E|V^j_{m,n}|^{2+\delta} = \frac{E|W^j_{m,n} - \xi_{m,n}|^{2+\delta}}{\eta^{(2+\delta)/2}_{m,n}}.$$

Notice that the WMW test statistic $U^j_{m,n} \le mn$, thus, we have $W^j_{m,n} \le (mn/2)^2$. Moreover, we have $E|W^j_{m,n} - \xi_{m,n}|^{2+\delta} \le |m^2n^2/4 - \xi_{m,n}|^{2+\delta}$. Thus, for any fixed $m, n \ge 2$, we have

$$E|V^j_{m,n}|^{2+\delta} \le \Big\{\frac{mn(3mn - N - 1)}{12\sqrt{\eta_{m,n}}}\Big\}^{2+\delta} < \infty.$$

Finally, by the central limit theorem under the strong mixing condition (see Corollary 5.1 in Hall and Heyde (1980)), the proof of Theorem 3 is completed. □

Appendix D. Proof of Theorem 4

To prove Theorem 4, we first calculate the expectation of $W^j_{m,n}$, namely, $\mathrm{var}(U^j_{m,n})$, under the local alternative H''_1 in (15). In order to calculate the variance of $U^j_{m,n}$, we rewrite $U^j_{m,n}$ as

$$U^j_{m,n} = \sum_{i=1}^{m}\sum_{k=1}^{n} I(Y_{kj} < X_{ij}).$$

For ease of subsequent illustration, we define

$$\begin{aligned}&e_j \overset{\mathrm{def}}{=} E\{F_j(X_j + \delta_{2j}/\sqrt{N})\}, \\ &g_j \overset{\mathrm{def}}{=} E\{F_j(X_j + \delta_{2j}/\sqrt{N})\}^2 \quad \text{and} \\ &h_j \overset{\mathrm{def}}{=} E\{1 - F_j(X_j + \delta_{2j}/\sqrt{N})\}^2.\end{aligned} \tag{D.1}$$

For any $1 \le i \le m$ and $1 \le k \le n$, we have $E\left\{I(Y_{kj} < X_{ij})\right\} = \Pr(Y_j < X_j) = e_j$. Thus

$$E(U^j_{m,n}) = \sum_{k=1}^{n}\sum_{i=1}^{m} E\left\{I(Y_{kj} < X_{ij})\right\} = mne_j. \tag{D.2}$$

For any $1 \le i_1 \ne i_2 \le m$ and $1 \le k_1 \ne k_2 \le n$, we have

$$\begin{aligned}E\left\{I(Y_{k_1 j} < X_{i_1 j})I(Y_{k_2 j} < X_{i_1 j})\right\} &= E\{\Pr(Y_j < X_j \mid X_j)\}^2 = g_j, \\ E\left\{I(Y_{k_1 j} < X_{i_1 j})I(Y_{k_1 j} < X_{i_2 j})\right\} &= E\{\Pr(X_j > Y_j \mid Y_j)\}^2 = h_j \\ \text{and } E\left\{I(Y_{k_1 j} < X_{i_1 j})I(Y_{k_2 j} < X_{i_2 j})\right\} &= E^2\{\Pr(Y_j < X_j \mid X_j)\} = e_j^2.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}E(U^j_{m,n})^2 &= E\Big\{\sum_{k=1}^{n}\sum_{i=1}^{m} I(Y_{kj} < X_{ij})\Big\}^2 \\ &= mne_j + mn(n-1)g_j + mn(m-1)h_j + mn(m-1)(n-1)e_j^2.\end{aligned} \tag{D.3}$$

Combining (D.2) and (D.3), we can obtain that

$$\mathrm{var}(U^j_{m,n}) = mne_j - mn(N-1)e_j^2 + mn(n-1)g_j + mn(m-1)h_j. \tag{D.4}$$

Furthermore, the expectation of $W^j_{m,n}$ under the local alternative H_1'' in (15) can be derived as

$$E(W^j_{m,n}) = mn\left\{(1-mn)e_j + \tfrac{m-1}{n-1}e_j^2 + (n-1)g_j + (m-1)h_j + mn/4\right\}.$$

Let the function $f(x)$ be differentiable with the first derivative $f'(x)$. Then by the Taylor expansion, we have

$$F_j\Big(x + \frac{\delta_{2j}}{\sqrt{N}}\Big) = F_j(x) + f_j(x)\frac{\delta_{2j}}{\sqrt{N}} + f_j'(x)\frac{\delta_{2j}^2}{2N} + O(N^{-3/2}).$$

Thus,

$$\begin{aligned} e_j &= \int_{-\infty}^{\infty}\Big\{F_j(x) + f_j(x)\frac{\delta_{2j}}{\sqrt{N}} + f_j'(x)\frac{\delta_{2j}^2}{2N} + O(N^{-3/2})\Big\}\mathrm{d}F_j(x) \\ &= \frac{1}{2} + E\left\{f_j(X_j)\right\}\frac{\delta_{2j}}{\sqrt{N}} + E\left\{f_j'(X_j)\right\}\frac{\delta_{2j}^2}{2N} + O(N^{-3/2}). \end{aligned} \tag{D.5}$$

By the same way, we can transfer g_j and h_j into

$$\begin{aligned} g_j &= \frac{1}{3} + E\left\{F_j(X_j)f_j(X_j)\right\}\frac{\delta_j}{\sqrt{N}} + O(N^{-1}) \quad \text{and} \\ h_j &= \frac{1}{3} + E[\left\{F_j(X_j) - 2\right\}f_j(X_j)]\frac{\delta_j}{\sqrt{N}} + O(N^{-1}). \end{aligned} \tag{D.6}$$

Hence, plugging equations (D.5)-(D.6) into (D.4), we can simplify $E(W^j_{m,n})$ to (16) and obtain that $E(W^j_{m,n}) - \xi_{m,n} = E^2\left\{f_j(X_j)\right\}m^2n^2\delta_{2j}^2/N + O\left(N^{5/2}\right)$.

Next, we investigate the asymptotic normality of $W_{m,n}$ under the local alternative H_1'' in (15). First of all, we have

$$\frac{\sqrt{p}(W_{m,n} - \xi_{m,n})}{\sqrt{\eta_{m,n}}} = \frac{\sqrt{p}\left\{W_{m,n} - p^{-1}\sum_{j=1}^p E(W^j_{m,n})\right\}}{\sqrt{\eta_{m,n}}} + \frac{\sum_{j=1}^p\{E(W^j_{m,n}) - \xi_{m,n}\}}{\sqrt{p\eta_{m,n}}}. \tag{D.7}$$

Under Conditions (C3)-(C4) and the definition of τ_2^2, we have $p^{-1}\,\mathrm{var}(\sum_{j=1}^p V^j_{m,n}) = \tau_2^2 - 1 + \gamma_2(0)$. Thus, we can prove that the first term in (D.7) converges to $C_2 \cdot N(0,1)$ in distribution through similar techniques as in Theorem 2. As $(m,n,p) \to \infty$, we can prove that the limit of the second term in (D.7) is $p^{-1/2}\Delta_2$, where Δ_2 is defined in (17). As a result,

$$\frac{\sqrt{p}(W_{m,n} - \xi_{m,n})}{\sqrt{\eta_{m,n}}} \xrightarrow{D} C_2 \cdot N(0,\tau_2^2) + \frac{\Delta_2}{\sqrt{p}}, \quad \text{as } (m,n,p) \to \infty.$$

This further leads to the power function of $W_{m,n}$ as

$$\begin{aligned} \beta(W_{m,n}) &= \Pr\Big\{\frac{\sqrt{p}(W_{m,n} - \xi_{m,n})}{\tau_2\sqrt{\eta_{m,n}}} > Z_{1-\alpha} \mid H_1'' \text{ is true}\Big\} \\ &= 1 - \Phi\Big\{C_2^{-1}\Big(Z_{1-\alpha} - \frac{\Delta_2}{\tau_2\sqrt{p}}\Big)\Big\}. \qquad \square \end{aligned}$$

References

Bai, Z.D., Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. Stat. Sin. 6, 311–329.

Brockwell, P.J., Davis, R.A., 2009. Time Series: Theory and Methods. Springer Series in Statistics. Springer, New York.

Cai, T., Liu, W., Xia, Y., 2013. Two-sample test of high dimensional means under dependence. J. R. Stat. Soc., Ser. B 76, 349–372.

Chen, L.S., Paul, D., Prentice, R.L., Wang, P., 2011. A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. J. Am. Stat. Assoc. 106, 1345–1360.

Chen, S.X., Qin, Y.L., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. Ann. Stat. 38, 808–835.

Chernozhukov, V., Chetverikov, D., Kato, K., 2013. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann. Stat. 41, 2786–2819.

Dan, N., Recknor, J., Reecy, J.M., 2008. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. Bioinformatics 24, 192–201.

Dong, K., Pang, H., Tong, T., Genton, M.G., 2016. Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. J. Multivar. Anal. 143, 127–142.

Gregory, K.B., Carroll, R.J., Baladandayuthapani, V., Lahiri, S.N., 2015. A two-sample test for equality of means in high dimension. J. Am. Stat. Assoc. 110, 837–849.
Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and Its Application. Academic Press, New York.
Hu, Z., Tong, T., Genton, M.G., 2019. Diagonal likelihood ratio test for equality of mean vectors in high-dimensional data. Biometrics 75, 256–267.
Li, H., Aue, A., Paul, D., Peng, J., Wang, P., 2020. An adaptable generalization of Hotelling's T^2 test in high dimension. Ann. Stat. 48, 1815–1847.
Mcmurry, T.L., Politis, D.N., 2010. Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. J. Time Ser. Anal. 31, 471–482.
Srivastava, M.S., 2007. Multivariate theory for analyzing high dimensional data. J. Japan Statist. Soc. 37, 53–86.
Srivastava, M.S., Du, M., 2008. A test for the mean vector with fewer observations than the dimension. J. Multivar. Anal. 99, 386–402.
Wang, L., Peng, B., Li, R., 2015. A high-dimensional nonparametric multivariate test for mean vector. J. Am. Stat. Assoc. 110, 1658–1669.
Wang, R., Shao, X., 2020. Hypothesis testing for high-dimensional time series via self-normalization. Ann. Stat. 48, 2728–2758.
Wu, X., Wang, J., Cui, X., Maianu, L., Rhees, B., Rosinski, J., So, W.V., Willi, S.M., Osier, M.V., Hill, H.S., et al., 2007. The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle. Endocrine 31, 5–17.
Wu, Y., Genton, M., Stefanski, L., 2006. A multivariate two-sample mean test for small sample size and missing data. Biometrics 62, 877–885.
Yamada, T., Srivastava, M., 2012. A test for multivariate analysis of variance in high dimension. Commun. Stat., Theory Methods 41, 2602–2615.
Zhang, D., Wu, W.B., 2017. Gaussian approximation for high dimensional time series. Ann. Stat. 45, 1895–1919.
Zhang, X., Cheng, G., 2018. Gaussian approximation for high dimensional vector under physical dependence. Bernoulli 24, 2640–2675.
Zhao, J., Xu, X., 2016. A generalized likelihood ratio test for normal mean when p is greater than n. Comput. Stat. Data Anal. 99, 91–104.
Zoh, R.S., Sarkar, A., Carroll, R.J., Mallick, B.K., 2018. A powerful bayesian test for equality of means in high dimensions. J. Am. Stat. Assoc. 113, 1733–1741.