Statistica Sinica **34** (2024), 1951-1971
doi:https://doi.org/10.5705/ss.202022.0143

AN ADAPTIVE WEIGHTED COMPONENT TEST FOR HIGH-DIMENSIONAL MEANS

Yidi Qu, Lianjie Shu and Jinfeng Xu*

The University of Hong Kong, University of Macau and City University of Hong Kong

Abstract: Two recent streams of two-sample tests for high-dimensional data are the sum-of-squares-based and supremum-based tests. The former is powerful against dense differences in two population means, and the latter is powerful against sparse differences. However, the level of sparsity and signal strength are often unknown, in practice, making it unclear which type of test to use. Here, we propose an adaptive weighted component test that provides good power against a variety of alternative hypotheses with unknown sparsity levels and varying signal strengths. The basic idea is to first allocate different weights to components with varying magnitudes in a sum-of-squares-based test, and then to combine multiple weighted component tests to make the underlying test adaptive to different sparsity levels of the mean differences. We examine the asymptotic properties of the proposed test, and use numerical comparisons to demonstrate the superior performance of the proposed test across a spectrum of situations.

Key words and phrases: High-dimensional test, Huber's weight function, testing equality of mean vectors, weighted components.

1. Introduction

In real applications, it is often desirable to test whether the mean vectors of two populations are the same. This can be formulated as a hypothesis testing problem as follows:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ versus } H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2,$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ denote the two population mean vectors. To fix the notation, let $\{\boldsymbol{X}_{1i}\}_{i=1}^{n_1}$ and $\{\boldsymbol{X}_{2j}\}_{j=1}^{n_2}$ be independent and identically distributed (i.i.d.) samples from two populations with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and $p \times p$ covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Here, n_1 and n_2 represent the size of the first and second samples, respectively. Denote n as the sum of sample sizes, that is, $n = n_1 + n_2$.

In low-dimensional cases, that is, $p \ll n$, several methods have been developed to test the difference in the mean vectors between two populations. For example, the classical T^2 test of Hotelling (1931) has desirable properties and

*Corresponding author.

satisfactory power in conventional low-dimensional cases. However, with rapid advances in sensing technology and data acquisition systems, high-dimensional data are becoming more common, where the dimension of the data can exceed the number of sampled observations, that is, $p > n$, leading to the so-called "large-p-small-n" problem. For example, genetic data may contain thousands of DNA segments from only a few hundred patients (Chen and Qin (2010)). In a 200mm fabrication line investigated by Kumar et al. (2011), which produces 250 chips per wafer in lots of 25 wafers, the manufactured product with 22 layers can involve 524 processing steps, with more than 21,710 process variables.

In high-dimensional cases, traditional multivariate two-sample tests, such as the T^2 test, either cannot be applied directly or their power is too low. For example, the T^2 test statistic is undefined when p is larger than n, because it involves inverting the $p \times p$ sample covariance matrix, which is singular. Even when the T^2 test is defined, its detection power decreases as the dimension p increases. As shown theoretically in Fan (1996), the standard Wald, score, and likelihood ratio tests may have power that decrease in terms of the type-I error rate as p increases, even for the simple one-sample test on the mean of a normal distribution with a known covariance matrix.

Various two-sample tests for high-dimensional data have been proposed, and can be grouped into two categories: sum-of-squares-based tests, and supremum-based tests. The first category is motivated by the L_2-type distance between two mean vectors, where all entries are considered. Several researchers have attempted to extend the T^2 statistic to the case of $p > n$ by replacing the sample covariance matrix with a nonsingular matrix. For example, Bai and Saranadasa (1996) propose a straightforward procedure (referred to here as the BS test), in which they replace the sample covariance matrix with an identity matrix. In order to simplify the theoretical derivation, Chen and Qin (2010) suggest a test (the CQ test) that removes the cross-product terms from the BS test. To account for possibly varying variances of the components of the data, one can replace the sample covariance matrix with a diagonal version; see, for example, Srivastava and Du (2008), Srivastava (2009), and Srivastava and Kubokawa (2013). In order to avoid a full estimation of the covariance matrix, Gregory et al. (2015) propose a generalized component test (GCT) that assumes that the p components admit a logical ordering such that the dependence between components is related to their displacement. Moreover, to accommodate strongly spiked eigenvalues (SSE) in high-dimensional data, Aoshima and Yata (2018) and Ishii, Yata and Aoshima (2019) propose distance-based tests that use the estimated eigen-structures, and obtain their limiting distributions. Zhang et al. (2020) propose a Welch–Satterthwaite χ^2-type test to further relax the restrictive assumptions on the covariance structure. Other approaches use the random projection method Srivastava, Li and Ruppert (2016)), interpoint distance (Biswas and Ghosh (2014)), and spatial sign ranks (Wang, Peng and Li (2015),

Chakraborty and Chaudhuri (2017)). The second category is motivated by the L_∞-type distance between two mean vectors, where only the largest deviation is used. A sample of research in this category includes Chang et al. (2017) and the CLX test proposed by Cai, Liu and Xia (2014).

However, these two streams of tests are designed for extreme situations. The first category is particularly efficient in the dense case, in which almost all of the components in the two mean vectors exhibit some differences. In contrast, the second category is efficient in the sparse case in which a few leading components in the two mean vectors suffer from substantial changes. As a result, no single test performs relatively well in both cases.

In reality, the sparsity level of the mean differences, that is, the number of zero elements in $\boldsymbol{\mu_1}-\boldsymbol{\mu_2}$, is often unknown. Furthermore, the sparsity level may lie somewhere between the two extreme cases, neither dense nor sparse. Therefore, it is unclear how to choose a powerful test from the above two categories when the sparsity level of the mean differences is unknown. Moreover, most of the above tests assume that the signal strength (or magnitude) is equal for each component of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. In order to remove the assumptions of a known sparsity level of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and an equal shift magnitude in each component, we require a flexible two-sample test for comparing high-dimensional mean vectors. Motivated by this, we develop a robust two-sample test for high-dimensional mean vectors with unknown sparsity levels and varying magnitudes of the mean differences.

The proposed test compresses two steps. The first introduces a robust weighting function capable of allocating different weights to components of varying magnitudes in a sum-of-squares-based test. This naturally generalizes the GCT, with equal weights on each component as a special case. Intuitively, this improves the test power when the mean differences have different magnitudes by putting relatively large weights onto leading components, and relatively small weights onto small components. The second step combines the multiple weighted component tests (WCTs) from the first step to select the most powerful test from the candidate tests. This second step makes the proposed test adaptive to different sparsity levels of mean differences, and is similar to the idea of the adaptive sum-of-powers test (ASPU test) of Xu et al. (2016). For simplicity, we denote the proposed adaptive WCT as AWCT throughout the remainder of the paper.

Note that our approach differs from the ASPU test in two important aspects. First, the proposed approach dynamically allocates weights to components based on their magnitudes. In contrast, the ASPU test always puts the same weight on each component in each individual sum-of-powers-type test. In this sense, the proposed approach is more flexible, because it is more reasonable to assume that the components have different shifts in magnitudes in practice. Second, although both the ASPU and the AWCT tests combine multiple individual tests to improve the test power when the sparsity level of the signal is unknown, the

individual tests work differently. The individual sum-of-powers test in the ASPU test adjusts the power for detecting sparse or dense signals by tuning the power index of the distances. In contrast, the individual WCT test does so by tuning the weighting parameter of a robust weight function, such as Huber's function. Therefore, the proposed approach is expected to provide overall good test power when the components have varying magnitudes of mean shifts, in addition to its robustness to the sparsity level of the signals.

The remainder of the paper is organized as follows. Section 2 describes the AWCT statistic in detail. Section 3 derives its asymptotic properties. Section 4 presents an extensive simulation study of the AWCT, comparing its performance with that of the BS, CQ, GCT, CLX, and ASPU tests in terms of power and maintenance of the nominal size. Section 5 presents two real examples. Concluding remarks are presented in Section 6. Proofs of our asymptotic theories are provided in the Supplementary Material.

2. Test Statistics

For samples $\{\boldsymbol{X}_{ki}\}_{i=1}^{n_k}$, where $k = 1, 2$, denote X_{ki}^j as the jth component ($j = 1, \ldots, p$) of the ith observation in sample k. Denote $s_{k,jj}^2 = \sum_{i=1}^{n_k}(X_{ki}^j - \bar{X}_k^j)^2/n_k$ as the sample variance of the jth component for the kth sample, where $\bar{X}_k^j = \sum_{i=1}^{n_k} X_{ki}^j/n_k$. Define t_j^2 as

$$t_j^2 = \frac{(\bar{X}_1^j - \bar{X}_2^j)^2}{s_{1,jj}^2/n_1 + s_{2,jj}^2/n_2},$$

which then converges to a χ_1^2 distribution as $n_1, n_2 \to \infty$ under the null hypothesis.

The statistic t_j^2 tests the mean difference in the jth component. To consider all signal information, one can compute the sum of t_j^2 over all components, as in the GCT statistic, for $j = 1, \ldots, p$. However, the components often have varying magnitudes. Thus, it is reasonable to assign larger weights to large components to improve the power of the test statistic. For this purpose, we establish the WCT statistics, as follows:

$$T_{WCT} = \sum_{j=1}^{p} \frac{\omega_j t_j^2}{p}, \tag{2.1}$$

where ω_j is the weight allocated to t_j. Clearly, the WCT statistic is a natural generalization of the GCT statistic, because it allows us to assign different weights to each of the components t_j. When ω_j is fixed as a constant, an equal weight is assigned to all components. In this case, the WCT performs in essentially the same way as the GCT.

Different weighting functions can be used. Here, we consider weights motivated by robust procedures such as Huber's function (Dutter and Huber

(1981)) and Welsch's function (Holland and Welsch (1977)). For the sake of simplicity, we restrict our discussion to Huber's weight function:

$$\omega_j = \begin{cases} 1-(1-\kappa)R/t_j^2, & t_j < -\sqrt{R}, \\ \kappa, & -\sqrt{R} \le t_j \le \sqrt{R}, \\ 1-(1-\kappa)R/t_j^2, & t_j > \sqrt{R}, \end{cases}$$

where $\kappa \in (0,1]$, and R is a positive threshold that determines whether the component t_j^2 is too large.

Note that when $R \to \infty$, $\omega_j = \kappa$; that is, the same weight is allocated to t_j^2 along each component. Therefore, the value of R should not be too high in practice in order to adaptively allocate weights to the components. In robust weight functions, the value of R is often chosen based on the rule of thumb $R \in [2.5, 3.5]$ (Capizzi and Masarotto (2003)). By doing so, the random variable t_j^2 has a small probability of exceeding R. Note that t_j^2 converges to a χ_1^2 distribution as $n_1, n_2 \to \infty$ under the null hypothesis. For a χ_1^2 random variable, there is only a 11.38% probability of it exceeding $R = 2.5$, and a 6.13% probability of exceeding $R = 3.5$. In this study, we choose $R \in [2.5, 3.5]$, focusing on $R = 3$ for simplicity.

The parameter κ controls the relative weight allocated to the component t_j^2. To illustrate the effect of κ, Figure 1 plots the weight ω_j as a function of t_j for different values of κ when $R = 3$, showing that a smaller κ value allocates relatively small weights to smaller components t_j^2, but relatively large weights to larger components t_j^2. When κ increases, the differences in the weights for all the components tends to decrease. Consider two extreme cases. When $\kappa \to 0$, $\omega_j \to 0$ for $t_j^2 \le R$ and $\omega_j = 1 - R/t_j^2$ for $t_j^2 > R$. This implies that we consider only the extremely large components t_j^2 in the WCT statistic, and ignore the other components. In this case, one can expect the WCT to perform like the CLX test, which has good test power in the case of sparse signals. On the other hand, when $\kappa = 1$, $\omega_j = 1$, for $j = 1, 2, \ldots, p$. In this case, the same weight is used for all the components. Therefore, one can expect the WCT to perform essentially like the GCT, which has good test power in the case of dense signals.

Therefore, the parameter κ has an important effect on the power of the WCT. The WCT statistic in Equation (2.1) can be rewritten as

$$T_{WCT}(\kappa) = \sum_{j=1}^{p} \omega_j(\kappa) \frac{t_j^2}{p}.$$

Whether $T_{WCT}(\kappa)$ is powerful depends on the unknown sparsity level, that is, the pattern of nonzero signals. To provide overall good power, one can incorporate multiple testing in the procedure so that at least one yields a high power for a particular application with unknown truth. This can be achieved by combining

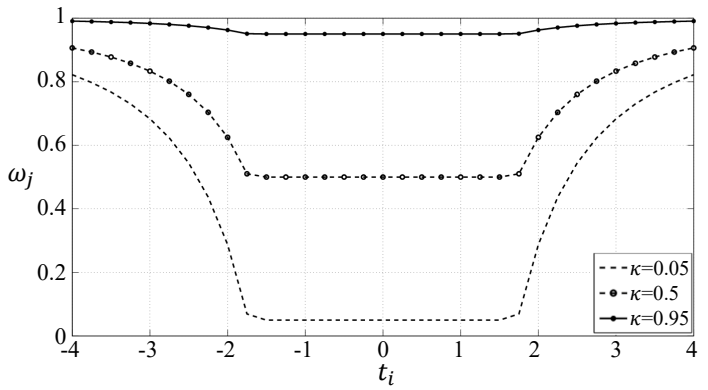

Figure 1. Plot of ω_j under different values of κ when $R = 3$.

multiple WCTs, as follows:

$$T_{AWCT} = T_{WCT}\left(\underset{0\le\kappa\le 1}{\operatorname{argmin}} P(\kappa)\right),$$

where $P(\kappa)$ is the p-value of the $T_{WCT}(\kappa)$ test. The idea of taking the minimum p-value to approximate the maximum power is widely used; see, for example, Xu et al. (2016) and Yu et al. (2009).

In practice, we need to choose candidate values for κ for the proposed test in order to improve the test performance when the sparsity level of the signal is unknown. In principle, there are many candidate values for κ. However, this greatly complicates the underlying test for only a marginally improvement in the test power. To achieve a trade-off between simplicity and test power, we choose three candidate values of $\kappa \in \Gamma = \{0.05, 0.5, 0.95\}$, aimed at detecting very sparse, not-that-sparse, and dense shifts in the mean differences, respectively. However, other choices of candidate values for κ can be analyzed similarly. As shown later, $\kappa \in \Gamma = \{0.05, 0.5, 0.95\}$ provides an overall good power under a wide variety of alternative hypotheses when the sparsity level is unknown.

3. Main Results

3.1. Asymptotic theory

For a set of multivariate random vectors $\boldsymbol{Z}$ and integers $a < b$, let $\mathcal{F}_a^b$ be the σ field generated by $\{Z^j : j \in [a, b]\}$, that is, $\mathcal{F}_a^b = \sigma\left\{Z^a, Z^{a+1}, \ldots, Z^b\right\}$, where Z^j denotes the jth element of $\boldsymbol{Z}$. For all positive integers $s < p$, the strong mixing coefficients are defined as

$$\alpha_Z(s) = \sup_{1\le k\le p-s} \left\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+s}^p\right\}.$$

Similar to the assumptions made in Xu et al. (2016), the following Conditions are assumed to derive the asymptotic distribution of T_{WCT}:

C.1 There exists some constant B such that

$$B^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}_1), \lambda_{\min}(\boldsymbol{\Sigma}_2), \lambda_{\max}(\boldsymbol{\Sigma}_1), \lambda_{\max}(\boldsymbol{\Sigma}_2) \leq B,$$

where $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the minimum and maximum eigenvalues, respectively, of a matrix $\boldsymbol{A}$. In addition, the correlations are bounded away from -1 and 1, that is,

$$\max_{k=1,2;1\leq i\neq j\leq p} \frac{|\sigma_{k,ij}|}{(\sigma_{k,ii}\sigma_{k,jj})^{1/2}} < 1-\eta,$$

for some $\eta > 0$.

C.2 $\{(X_{ki}^j, i = 1, \ldots, n_k) : j \geq 1\}$ is α-mixing, for $k = 1, 2$, and $\alpha_{\boldsymbol{X}}(s) \leq M\delta^s$, for $\delta \in (0, 1)$ and some constant M.

C.3 $n_1/n_2 \to c \in (0, \infty)$ and $p = o(n^2)$; $\max_{1\leqslant j\leqslant p} E[\exp\{h(X_{k1}^j - \mu_k^j)^2\}] < \infty$ for $h \in [-M, M]$ and $k = 1, 2$, where μ_k^j denotes the jth element of $\boldsymbol{\mu}_k$.

C.1 and C.3 are assumptions on the eigenvalues and covariance, respectively, needed to establish the weak convergence of the WCT statistic and its joint asymptotic normality. C.2 is a commonly used mixing condition that assumes weak dependence for data sets with components that admit an ordering in time, space, or some other index, such that their dependence diminishes as the components become further apart. For example, measurements for methylation values are taken along a chromosome. The location of each measurement is recorded, providing an index over which dependence can be modeled. Under C.1–C.3, the asymptotic normality of the test statistic T_{WCT} and its asymptotic joint distribution are derived in Theorems 1 and 2, respectively.

Theorem 1. *Assume that Conditions* C.1–C.3 *hold. Under* H_0*, we have*

$$\frac{\sqrt{p}(T_{WCT} - \nu)}{\zeta} \to^d N(0, 1)$$

as $p \to \infty$*, where* $\nu = E(T_{WCT})$ *and* $\zeta^2 = p \cdot Var(T_{WCT})$ *are stated in Propositions* 1 *and* 2*.*

Proof. See the Appendix.

Theorem 2. *Assume that Conditions* C.1–C.3 *hold. Under* H_0*, for* $\boldsymbol{\Gamma} = \{\kappa_1, \kappa_2, \ldots, \kappa_d\} \in [0, 1]^d$ *(*$d < \infty$*), we have*

$$\sqrt{p}(T_{WCT}(\boldsymbol{\Gamma}) - \boldsymbol{\nu}(\boldsymbol{\Gamma}))^T \to^d N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = (r_{st})$ *with* $r_{ss} = \zeta_s^2 = pVar(T_{WCT}(\kappa_s))$ *for* $1 \le s \le d$, *and* $r_{st} = \gamma_{st}^2 = pCov(T_{WCT}(\kappa_s), T_{WCT}(\kappa_t))$ *for* $s \ne t \in \{1, 2, \ldots, d\}$.

Proof. See the Appendix.

Denote $I_j = I(t_j^2 \le R)$, and rewrite the mean of the T_{WCT} statistic as $\nu = \sum_{j=1}^p \nu_j/p$, where $\nu_j = E(\omega_j t_j^2)$. The following approximation holds for ν, ζ^2, and γ_{st}^2 under $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Proposition 1. *Under* $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, *we have*

$$\begin{aligned}\nu_j &= E\left\{I_j \kappa t_j^2 + (1 - I_j)(1 - (1-\kappa)R t_j^{-2}) t_j^2\right\} \\ &= (1-\kappa)\left\{\int_0^R F(x)dx - R\right\} + \int_0^\infty x f(x)dx + O\left(\frac{1}{n}\right),\end{aligned}$$

where $F(x)$ *and* $f(x)$ *denote the cumulative distribution function and probability density function, respectively, of the* χ_1^2 *distribution. Thus, the term* $\int_0^\infty x f(x)dx$ *is equal to one and is replaced by one in the following.*

According to Proposition 1, we estimate ν by $\hat{\nu} = (1-\kappa)\{\int_0^R F(x)dx - R\} + 1$. The consistency of $\hat{\nu}$ is shown in the Supplementary Material. Then, denoting $K_i = (\kappa - 1) I_i t_i^2 + t_i^2 + (1-\kappa) R I_i$, we have, $\zeta^2 = p^{-1} Var(\sum_{j=1}^p \omega_j t_j^2) = p^{-1} \sum_{j=1}^p Var\{K_j\} + p^{-1} \sum_{i \ne j} Cov\{K_i, K_j\}$.

Proposition 2. *Assume that Conditions* C.1–C.3 *hold. Under* H_0, *we have*

$$\begin{aligned}\varsigma^2 &= Var\{K_j\} \\ &= \int_0^R (1-\kappa)(R-x)\left[(1-\kappa)(R-x) + 2x\right] f(x)dx + \int_0^\infty x^2 f(x)dx \\ &\quad -(\kappa-1)^2\left\{\int_0^R F(x)dx\right\}^2 - 2(\kappa-1)\int_0^R F(x)dx - 1 + O\left(\frac{1}{n}\right).\end{aligned}$$

Note that $Cov\{K_i, K_j\} = \rho_{ij}\varsigma^2$, where $\rho_{ij} = Corr(K_i, K_j)$, which can be estimated by

$$\hat{\rho}_{ij} = \frac{\sum_{l=1}^{p-|i-j|}(K_l - \bar{K})(K_{l+|i-j|} - \bar{K})}{\sum_{l=1}^p (K_l - \bar{K})^2}, \qquad i, j = 1, 2, \ldots, p,$$

where $\bar{K} = \sum_{l=1}^p K_l/p$.

We estimate ζ^2 by

$$\hat{\zeta}^2 = \varsigma^2 + \frac{\sum_{i \ne j} \mathfrak{p}(|i-j|/L)\hat{\rho}_{ij}\varsigma^2}{p},$$

where $\mathfrak{p}(x)$ is a piecewise function of x such that $\mathfrak{p}(0) = 1$, $|\mathfrak{p}(x)| \le 1$ for all x, and $\mathfrak{p}(x) = 0$ for $|x| > 1$, and L is a user-selected lag window size. Here, we use

the Parzen window (Brockwell and Davis (2013)), that is,

$$\mathfrak{p}(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3, & |x| < \frac{1}{2}, \\ 2(1-|x|)^3, & \frac{1}{2} \le x \le 1, \\ 0, & |x| > 1. \end{cases}$$

The consistency of $\hat{\zeta}^2$ is shown in the Supplementary Material.

To derive the asymptotic joint distribution of the test statistics $T_{WCT}(\kappa)$, we need the following result to approximate the covariance $\gamma_{st}^2 = Cov(T_{WCT}(\kappa_s), T_{WCT}(\kappa_t))$.

Proposition 3. *Assume that Conditions* C.1–C.3 *hold. Under* H_0*, for* $0 \le \kappa_s, \kappa_t \le 1$*, we have*

$$\gamma_{st}^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{Cov(K_i(\kappa_s), K_j(\kappa_t))}{p},$$

where $K_i(\kappa) = (\kappa - 1)I_i t_i^2 + t_i^2 + (1-\kappa)RI_i$*. For* $i = j$*,*

$$\begin{aligned} \varsigma'^2 &= Cov(K_i(\kappa_s), K_i(\kappa_t)) \\ &= \int_0^R \left[(1-\kappa_s)(1-\kappa_t)(R-x)^2 + (2-\kappa_s-\kappa_t)(R-x)x\right] f(x)dx \\ &\quad + \int_0^\infty x^2 f(x)dx - (1-\kappa_s)(1-\kappa_t)\left\{\int_0^R F(x)dx\right\}^2 \\ &\quad -(2-\kappa_s-\kappa_t)\int_0^R F(x)dx - 1 + O\left(\frac{1}{n}\right). \end{aligned}$$

For $i \neq j$*,* $Cov\{K_i(\kappa_s), K_j(\kappa_t)\} = \varrho_{ij}\varsigma'^2$*, where* $\varrho_{ij} = Corr(K_i(\kappa_s), K_j(\kappa_t))$ *is estimated by*

$$\begin{aligned} \hat{\varrho}_{ij} = \sum_{l=1}^{p-|i-j|} &\left[(K_l(\kappa_s) - \bar{K}(\kappa_s))(K_{l+|i-j|}(\kappa_t) - \bar{K}(\kappa_t))\right. \\ &\left.+(K_l(\kappa_t) - \bar{K}(\kappa_t))(K_{l+|i-j|}(\kappa_s) - \bar{K}(\kappa_s))\right] \\ &\left[2\sum_{l=1}^{p}(K_l(\kappa_s) - \bar{K}(\kappa_s))(K_l(\kappa_t) - \bar{K}(\kappa_t))\right]^{-1}, \end{aligned}$$

for $i, j = 1, 2, \ldots, p$*, where* $\bar{K}(\kappa) = \sum_{l=1}^{p} K_l(\kappa)/p$*.*

Finally, we estimate γ_{st}^2 by

$$\hat{\gamma}_{st}^2 = \varsigma'^2 + \sum_{i \neq j} \frac{\mathfrak{p}(|i-j|/L)\hat{\varrho}_{ij}\varsigma'^2}{p}.$$

3.2. Asymptotic type-I error and power analysis

Denote $T = \sqrt{p}(T_{WCT} - \nu)/\zeta$. Assuming that Conditions C.1–C.3 hold, the asymptotic type-I error of the AWCT test based on $\mathbf{\Gamma} = \{\kappa_1, \kappa_2, \ldots, \kappa_d\} \in [0,1]^d$ $(d < \infty)$ can be calculated as

$$\begin{aligned} p &= pr(T_{AWCT} > C | H_0 \text{ true}) \\ &= 1 - pr(T_{AWCT} \leq C | H_0 \text{ true}) \\ &= 1 - pr\left(\max_{0 \leq i \leq d} T_i \leq C | H_0 \text{ true} \right) \\ &= 1 - pr(T_1 \leq C, T_2 \leq C, \ldots, T_d \leq C | H_0 \text{ true}) \\ &= 1 - \int_{(-\infty, C)^d} \phi_d(\mathbf{0}, \boldsymbol{\Omega}) dT_1 \ldots dT_d, \end{aligned}$$

where $\phi_d(\mathbf{0}, \boldsymbol{\Omega})$ denotes the probability distribution function of a d-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and covariance $\boldsymbol{\Omega}$. Here, $\boldsymbol{\Omega}$ is equal to the correlation matrix corresponding to the covariance matrix estimated using Proposition 3. For a given critical value C, the value of p can be calculated using the R package `mvtnorm`.

The test power of T_{AWCT} under H_A satisfies $pr(\min_{0 \leq \kappa \leq 1} P(\kappa) < \alpha) \geq pr(P(\kappa) < \alpha)$, for any $0 \leq \kappa \leq 1$, where α is the significance level. Therefore, the asymptotic power of the proposed test is one if there exists $0 \leq \kappa \leq 1$ such that $pr(P(\kappa) < \alpha) \to 1$; that is, $T_{WCT}(\kappa)$ has asymptotic power equal to one. Hence, to study the asymptotic power of the adaptive test, we need only focus on the power of $T_{WCT}(\kappa)$, for $0 \leq \kappa \leq 1$. In the following, we write $T_{WCT}(\kappa)$ as T_{WCT} for conciseness. Denote $\Phi(x)$ as the cumulative distribution function of the standard normal, and z_α as the corresponding $(1-\alpha)$th quantile.

Denote $\iota_j = \boldsymbol{\mu}_1^j - \boldsymbol{\mu}_2^j$, for $j = 1, 2, \ldots, p$. Then, the alternative hypothesis $H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ means that an unknown proportion q $(0 < q \leq 1)$ of ι_j's is not equal to zero. Denote $\nu_A = E(T_{WCT} | H_A \text{ true})$. Then, the power of the WCT, that is, $P(\sqrt{p}(T_{WCT} - \nu_A)/\hat{\zeta} > z_\alpha | H_A \text{ true})$, is equal to

$$1 - P\left(\frac{\sqrt{p}\,(T_{WCT} - \nu_A)}{\hat{\zeta}} < z_\alpha - \frac{\sqrt{p}(\nu_A - \nu)}{\hat{\zeta}} \Big| H_A \text{ true} \right).$$

The asymptotic normality of $\sqrt{p}\,(T_{WCT} - \nu_A)/\hat{\zeta}$ and the consistency of $\hat{\zeta}$ for ζ can be invoked under conditions C.1–C.3. We then approximate the power of the WCT using

$$1 - \Phi\left(z_\alpha - \frac{\sqrt{p}(\nu_A - \nu)}{\zeta} \right),$$

which is a function of $\sqrt{p}(\nu_A - \nu)/\zeta$. Define $G_{j,\iota_j}(x)$ and $g_{j,\iota_j}(x)$ as the cumulative distribution function and probability density function, respectively, of t_j^2 under ι_j. Under the alternative hypothesis, when $\iota_j \neq 0$, as $n_1, n_2 \to \infty$, the distribution

of t_j^2 converges to a noncentral chi-squared distribution, with degree of freedom one and noncentrality parameter ι_j^2, denoted as $\chi_1^2(\iota_j^2)$. From Proposition 1,

$$\begin{aligned}
\nu_A - \nu &= E(T_{WCT}|H_A \text{ true}) - E(T_{WCT}|H_0 \text{ true}) \\
&= p^{-1}(1-\kappa)\left\{\sum_{j=1}^{p}\left[\int_0^R G_{j,\iota_j}(x)dx - \int_0^R G_{j,0}(x)dx\right]\right\} \\
&\quad + p^{-1}\sum_{j=1}^{p}\left\{\int_0^\infty x g_{j,\iota_j}(x)dx - \int_0^\infty x g_{j,0}(x)dx\right\} \\
&= p^{-1}(1-\kappa)\sum_{j=1}^{p}\{H_{R,j}(\iota_j) - H_{R,j}(0)\} + p^{-1}\sum_{j=1}^{p}\left\{\iota_j^2 + O(n^{-1})\right\} \\
&\approx p^{-1}\sum_{j=1}^{p}\left\{(1-\kappa)\left[h_{R,j}(0)\iota_j + \frac{h'_{R,j}(\tau_j)\iota_j^2}{2}\right] + \left[\iota_j^2 + O(n^{-1})\right]\right\} \\
&= p^{-1}\sum_{j=1}^{p}\left\{a_{\kappa,R}(\tau_j)\iota_j^2 + O(n^{-1})\right\},
\end{aligned}$$

where $H_{r,j}(x) = \int_0^r G_{j,x}(y)dy$, $h_{r,j}(x) = \partial H_{r,j}(x)/\partial x$, $h'_{r,j}(x) = \partial^2 H_{r,j}(x)/\partial x^2$, and $a_{\kappa,R}(\tau_j) = 1 + (1-\kappa)h'_{R,j}(\tau_j)/2$, with $\tau_j \in (0, \iota_j)$. Now, the power can be expressed as

$$1 - \Phi\left(z_\alpha - p^{-1/2}\frac{\sum_{j=1}^{p}\left\{a_{\kappa,R}(\tau_j)\iota_j^2 + O(n^{-1})\right\}}{\zeta}\right).$$

4. Simulation Studies

In this section, we illustrate the performance of the proposed test, the AWCT, by comparing it with that of existing methods in simulations. The other tests included in the comparison are the BS, CQ, GCT, and ASPU tests, all of which are sum-of-squares-based tests. We also include the CLX test for testing sparse alternatives. The test performance is compared in terms of size control and power under various settings.

Without loss of generality, with $\boldsymbol{\mu}_1 = 0$, let $\boldsymbol{\mu}_2 = 0$ under the null hypothesis, and set the first $[p^{1-\beta}]$ elements of $\boldsymbol{\mu}_2$ unequal to zero under the alternative hypothesis, where $\beta \in [0, 1]$ controls the signal sparsity. Three values of $\beta = 0.3$, 0.5, 0.7 are considered, corresponding to the cases with dense, medium, and sparse differences in the two population means, respectively. The magnitudes of $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ measure the signal strength. Two settings of magnitudes are considered: (i) the case with equal magnitude of $\mu_2^i = \{2r\,(1/n_1 + 1/n_2)\log p\}^{1/2}$, for $i = 1, 2, \ldots, m$, where r is a constant controlling the signal strength, and (ii) μ_2^i increases linearly over the range $[\{1.5r\,(1/n_1 + 1/n_2)\log p\}^{1/2},\ \{2.5r\,(1/n_1 + 1/n_2)\log p\}^{1/2}\ ]$, for $i = 1, 2, \ldots, m$.

We choose three specific models for the covariance structure from the work of Cai, Liu and Xia (2014), given as follows:

(a) $\mathbf{\Sigma} = (\sigma_{i,j})$, where $\sigma_{i,j} = 0.6^{|i-j|}$, for $1 \leq i, j \leq p$.

(b) $\mathbf{\Sigma} = (\sigma_{i,j})$, where $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.8$, for $2(k-1)+1 \leq i \neq j \leq 2k$, where $k = 1, 2, \ldots, [p/2]$, and $\sigma_{i,j} = 0$ otherwise.

(c) $\mathbf{\Sigma} = (\sigma_{i,j})$, where $\sigma_{i,i} = 1$ and $\sigma_{i,j} = |i-j|^{-5}/2$, for $i \neq j$.

In Model (a), the covariance matrix has a bandable structure, but has a sparse structure in Model (b). The entries of the covariance structure in Model (c) decay as a function of the lag $|i-j|$, which arises naturally in time series analysis. In this case, neither the covariance matrix nor its inverse is sparse.

Under each model, two independent random samples $\{\boldsymbol{X}_{1i}\}_{i=1}^{n_1}$ and $\{\boldsymbol{X}_{2j}\}_{j=1}^{n_2}$ are generated from a multivariate distribution with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and a common covariance matrix $\mathbf{\Sigma}$. The dimension p takes $p = 400$ and the sample sizes take $n_1 = n_2 = 200$. To illustrate the effects of the distributions, we examine three types of distributions: (i) the multivariate normal, (ii) the multivariate t-distribution with degrees of freedom $v = 3$, and (iii) a multivariate gamma distribution. The functions `rmvnorm` and `rmvt` from the R package `mvtnorm` and the function `rmvgamma` from the package `lcmix`, respectively, are used to generate the three types of distributions. Note that the parameter `sigma` in `rmvt` denotes the scale matrix, which is equal to $(v-2)\mathbf{\Sigma}/v$. To generate the third distribution, we generate a gamma(4,2) distribution with a shape parameter of four and a scale parameter of two for each dimension. To center its mean to zero, one can subtract the random samples from the mean of $4/2 = 2$.

The nominal significance level is set to $\alpha = 0.05$ and κ is adaptively selected from $\Gamma = \{0.05, 0.5, 0.95\}$. For the choice of L and R in our proposed test, the results are qualitatively the same for $L = 10, 20$, and 30 and for $R = 2.5$, 3, and 3.5. The results are also similar under different covariance matrix structures. For the sake of simplicity, we present only the results based on $L = 10$ and $R = 3$ under covariance Model (a). The power and empirical type-I error rate are calculated from 1,000 replications.

4.1. Empirical type-I error rate

Table 1 summarizes the empirical type-I error rates of the above tests under the multivariate normal distributions based on Model (a). Denote c as the ratio of p to n, that is, $c = p/n$. The results based on 1,000 and 2,000 replicates are presented, showing that the difference in the type-I error rate based on 1,000 and 2,000 replicates is negligible. For simplicity, we obtain the simulation results based on 1,000 replicates throughout the remainder of the paper.

In addition, we compare the computation times of among the AWCT, ASPU, and GCT tests. Consider $p = 400$ and $n_1 = n_2 = 200$ as an example. On

Table 1. The empirical type-I error rates of various tests under a multivariate normal distribution based on Model (a).

Number of replicates = 1,000												
	c=1						c=2					
n	AWCT	ASPU	GCT	CQ	BS	CLX	AWCT	ASPU	GCT	CQ	BS	CLX
200	0.06	0.05	0.10	0.05	0.04	0.04	0.06	0.05	0.08	0.06	0.05	0.04
250	0.05	0.04	0.09	0.05	0.04	0.04	0.05	0.05	0.07	0.05	0.04	0.04
300	0.06	0.06	0.09	0.05	0.04	0.05	0.06	0.05	0.07	0.05	0.04	0.05
Number of replicates = 2,000												
	c=1						c=2					
n	AWCT	ASPU	GCT	CQ	BS	CLX	AWCT	ASPU	GCT	CQ	BS	CLX
200	0.06	0.05	0.10	0.06	0.04	0.04	0.06	0.06	0.07	0.06	0.05	0.05
250	0.05	0.04	0.09	0.05	0.04	0.04	0.05	0.05	0.07	0.05	0.04	0.04
300	0.06	0.06	0.09	0.05	0.04	0.05	0.06	0.05	0.07	0.05	0.04	0.05

Table 2. The empirical type-I error rates of various tests under a multivariate gamma distribution based on Model (a).

	c=1						c=2					
n	AWCT	ASPU	GCT	CQ	BS	CLX	AWCT	ASPU	GCT	CQ	BS	CLX
200	0.06	0.04	0.11	0.05	0.04	0.04	0.06	0.06	0.08	0.06	0.05	0.05
250	0.05	0.06	0.09	0.06	0.05	0.04	0.06	0.04	0.08	0.05	0.05	0.05
300	0.05	0.05	0.10	0.05	0.04	0.05	0.05	0.06	0.07	0.06	0.05	0.05

Table 3. The empirical type-I error rates of various tests under a multivariate t_3 distribution based on Model (a).

	c=1						c=2					
n	AWCT	ASPU	GCT	CQ	BS	CLX	AWCT	ASPU	GCT	CQ	BS	CLX
200	0.05	0.04	0.09	0.05	0.01	0.03	0.05	0.05	0.07	0.05	0.00	0.04
250	0.06	0.04	0.08	0.06	0.01	0.04	0.06	0.04	0.07	0.06	0.00	0.05
300	0.05	0.03	0.08	0.05	0.01	0.03	0.04	0.04	0.06	0.04	0.00	0.04

a personal computer (MacBook Air with a 1.6 GHz Dual-Core Intel Core i5 processor and 8 GB memory), it takes around 6.78 seconds for the ASPU test to approximate the type-I error rate, 0.37 seconds for the AWCT test, and 0.05 seconds for the GCT test. Thus, the GCT and AWCT tests are clealy more computationally efficient than the ASPU test.

Table 1 shows that under the multivariate normal distribution, nearly all tests maintain close-to-nominal type-I error rates. Only the GCT exhibits inflated type-I error rates, perhaps because of its low convergence rate to the asymptotic null distribution. Tables 2 and 3 present the empirical type-I error rates of the above tests under the multivariate gamma and t_3 distributions, respectively. Table 2 show that, under the multivariate gamma distribution, the results are similar to those under the multivariate normal distribution. From Table 3, under the multivariate t_3 distribution, in addition to the GCT method, the BS method

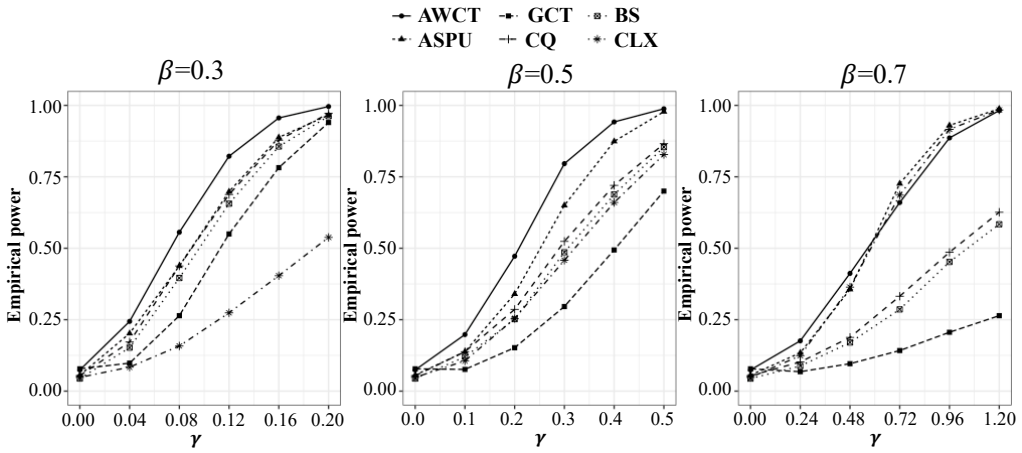

Figure 2. Power curves of the various tests against r under different sparsity levels of β, based on Model (a), with normal innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

also fails to maintain the nominal type-I error rate, whereas the other tests maintain close-to-nominal type-I error rates.

4.2. Power comparisons

Figure 2 compares the power curves of the above tests against r under different sparsity levels of β based on Model (a) with normal innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. For the case of dense signals ($\beta = 0.3$), the AWCT has the highest power, and the CLX has the lowest power. This is not surprising, the CLX is a supremum-based test, which is less efficient in terms of detecting dense signals. When β increases to $\beta = 0.5$, the AWCT has higher power than the ASPU, CQ, and BS, followed by the CLX and GCT, which has the lowest power. This illustrates that the power of the GCT decreases substantially as the sparsity level of the signals increases. When β further increases to $\beta = 0.7$, the AWCT, ASPU, and CLX methods exhibit competitive power, and outperform the CQ, BS, and GCT methods. To compare the power performance under skewed innovations, Figure 3 compares the power curves of the above tests against r under different sparsity levels of β based on Model (a), with centered gamma(4, 2) innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. The results are similar to those with normal innovations.

To illustrate the effect of heavy-tailedness on the performance of the proposed test, Figure 4 shows the power curves of the various tests against r under different sparsity levels of β, based on Model (a) with multivariate t_3 innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. The results do not differ greatly from those of the normal and skewed innovations.

In summary, Figures 2 to 4 indicate a good property of the proposed test. In particular, the AWCT always has the highest power, or power close to the highest. This indicates the capability of the AWCT to provide overall good power in a

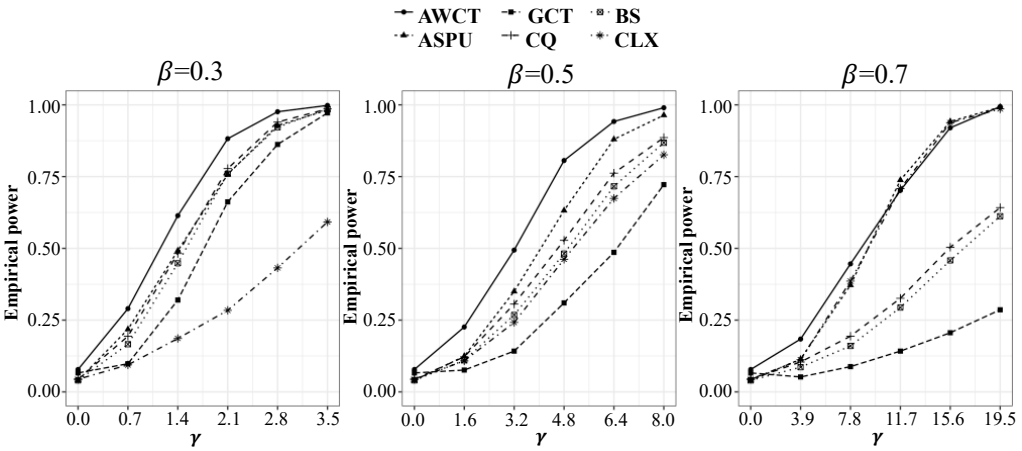

Figure 3. Power curves of the various tests against r under different sparsity levels of β, based on Model (a) with centered gamma(4, 2) innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

wide variety of situations. The simulation results under Models (b) and (c) are provided in the Supplementary Material, because they are similar to those under Model (a).

4.3. Effect of heteroscedasticity

Extreme values of t_j^2 tend to occur if $s_{1,jj}^2$ and $s_{2,jj}^2$ are very small under the alternative hypothesis. On the other hand, large values of $s_{1,jj}^2$ and $s_{2,jj}^2$ tend to reduce t_j^2, and thus extreme values do not occur. The size of a test is expected to be robust to any scaling of the variances. To investigate the effect of heteroscedasticity on the performance of the above tests, following the method of Gregory et al. (2015), we scale the standard deviation of each component by the square root of a realization from the exponential distribution with mean 1/2, shifted to the right by 1/2. Thus, the average scaling is one and the scaled variances are bounded away from zero.

We repeat the power simulation using the centered gamma(4,2) under Model (a) under the heteroscedastic condition; the results are shown in the Supplementary Material for simplicity. Our results show that the AWCT method maintains overall good power under the heteroscedastic condition in comparison with other tests.

4.4. Performance under unequal magnitudes of mean differences

The above analysis focuses mainly on the case with equal magnitude for the nonzero-mean differences. Here, we investigate the performance under unequal magnitudes for the nonzero-mean differences, which is more general, and natural in practice. A potential benefit of the AWCT is that it allocates different weights to the components with varying magnitudes, in contrast to the GCT.

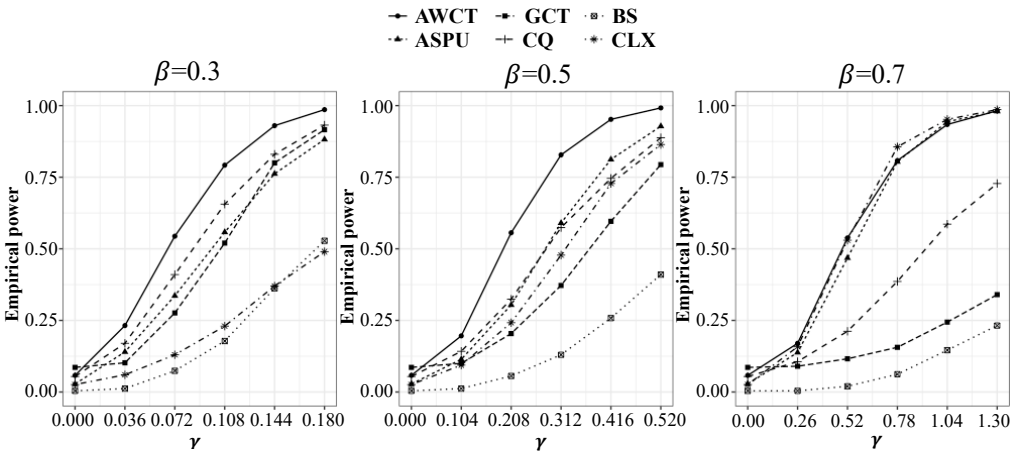

Figure 4. Power curves of the various tests against r under different sparsity levels of β, based on Model (a) with multivariate t_3 innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

Therefore, when the true mean differences between the two populations have unequal magnitudes, we expect the AWCT method to outperform the GCT significantly.

Figure 5 shows the power curves of the various tests against r under unequal magnitudes of mean differences, based on Model (a) with multivariate normal innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. For the components with nonzero means, the magnitudes are set to be linearly increasing over the range from $\{1.5r\,(1/n_1 + 1/n_2)\log p\}^{1/2}$ to $\{2.5r\,(1/n_1 + 1/n_2)\log p\}^{1/2}$, following the setting of Benjamini and Hochberg (1995). As shown in Figure 5, the AWCT outperforms the GCT, regardless of the value of β.

5. Real-Data Analysis

In this section, we apply the aforementioned methods to two real data sets: a DNA methylation data set and a data set from a semiconductor manufacturing process. Both data sets are publicly available. The first can be downloaded from the NCBI GEO website with GEO number GSE19711, and the second is available from the UC Irvine Machine Learning Repository `https://archive.ics.uci.edu/ml/datasets/SECOM`. Here, we present the application to DNA methylation data; the application to a semiconductor manufacturing process is given in the Supplementary Material. Death from ovarian cancer among women ranks fifth in the United States (Jemal et al. (2006)), and has been found to be associated with aberrant DNA methylation. A genome-wide DNA methylation profiling of the United Kingdom Ovarian Cancer Population Study (UKOPS) was conducted to identify methylation signatures associated with carcinogenesis (Teschendorff et al. (2010)). The data originate from the Illumina Infinium 27k Human DNA

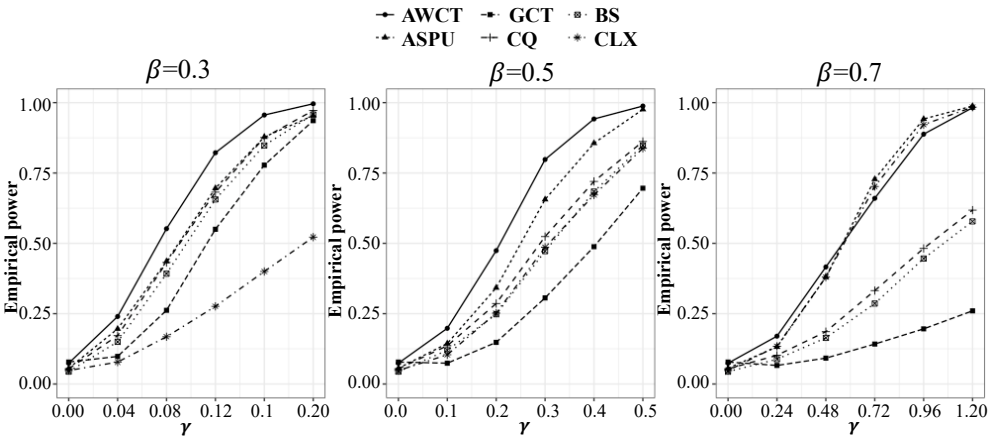

Figure 5. Power curves of the various tests against r under unequal magnitudes of mean differences, based on Model (a) with multivariate normal innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ when $\beta = 0.3, 0.5, 0.7$.

methylation Beadchip v1.2 with 27,578 CpGs, from 540 whole blood samples, including 266 samples from post-menopausal ovarian cancer patients, and 274 samples from age-matched normal controls.

In genomic data analysis, β-values and M-values are commonly used to quantify the level of DNA methylation (Bibikova et al. (2011)). The β-value is calculated from the intensity of the methylated allele ($\mathrm{Max}(M, 0)$) and the unmethylated allele ($\mathrm{Max}(U, 0)$), as follows:

$$\beta = \frac{\mathrm{Max}(M, 0)}{[\mathrm{Max}(M, 0) + \mathrm{Max}(U, 0) + 100]^{-1}}.$$

The β-values are usually preprocessed for the downstream statistical analysis, including quality control, background correction, and normalization. For differential DNA methylation analysis, the average β-value denotes the methylation level, or the percentage for an interrogated locus. The average β-value varies between zero and one. In an ideal situation, zero indicates that no copy of the CpG site in the sample is methylated. The value one indicates that every copy of the site is methylated. The average β-value approximates the methylation percentage for the population of a sampled CpG site. As an alternative, some investigators use the M-value, considering it to be statistically more valid (Du et al. (2010)). However, the interpretation of M-values is not as intuitive as it is for β-values. For this reason, we restrict our discussion to β-values.

We apply the AWCT, ASPU, GCT, CQ, BS, and CLX tests to test whether there is a significant difference in the DNA methylation levels between the cancer group and the normal group. The 27,578 CpGs of the ovarian cancer data are from all 23 pairs of chromosomes, including the sex chromosomes, namely,

Table 4. The p-values of the tests for the equality of the DNA methylation levels, measured using β-values on each chromosome (Chr).

Chr No.	1	2	3	4	5	6	7	8
AWCT	0	0	0	0	0	0	0	0
ASPU	0	0	0	0	0	0	0	0
GCT	0	0	0	0	0	0	0	0
CQ	0	0	0	0	0	0	0	0
BS	0.03	0.03	0.02	2.03×10^{-3}	6.15×10^{-3}	3.72×10^{-3}	0.01	6.51×10^{-3}
CLX	3.34×10^{-14}	7.44×10^{-13}	1.04×10^{-12}	5.87×10^{-13}	7.17×10^{-12}	1.47×10^{-13}	7.77×10^{-16}	0
Chr No.	9	10	11	12	13	14	15	16
AWCT	0	0	0	0	0	0	0	0
ASPU	0	0	0	0	0	0	0	0
GCT	0	0	0	0	0	0	0	0
CQ	0	1.11×10^{-16}	0	0	0	0	0	2.11×10^{-15}
BS	0.01	0.03	0.02	0.01	5.33×10^{-4}	0.02	0.02	0.09
CLX	9.75×10^{-11}	5.80×10^{-14}	1.11×10^{-16}	4.88×10^{-15}	0	1.87×10^{-14}	1.05×10^{-14}	6.66×10^{-16}
Chr No.	17	18	19	20	21	22	X	
AWCT	0	0	0	0	0	0	0	
ASPU	0	0	0	0	0	0	0	
GCT	0	0	0	0	0	0	0	
CQ	0	8.62×10^{-14}	0	0	0	1.83×10^{-13}	0	
BS	0.05	0.02	0.06	3.72×10^{-3}	4.28×10^{-4}	0.04	0	
CLX	1.35×10^{-14}	2.35×10^{-6}	1.55×10^{-13}	2.55×10^{-15}	4.69×10^{-13}	1.20×10^{-10}	2.72×10^{-12}	

chromosomes X and Y. We exclude chromosome Y from our analysis, because there are only seven CpGs from this chromosome, in which the sample size is larger than the dimension of the data. Prior to analysis, each missing value is replaced with the mean of the nonmissing values for the same CpGs in the same group.

Table 4 shows the p-values produced by the six tests for the equality of the methylation levels measured using the β-values on each chromosome. The R value is set to three for the AWCT. Nearly all the tests reject the null hypothesis at the 5% significance level. The only exception is the BS test on chromosomes 16 and 19. The p-values of the AWCT, ASPU, and GCT methods are nearly zero for all chromosomes.

The small p-values in Table 4 indicate that the differences in the DNA methylation levels on each CpGs between the cancer and the normal group are dense, and that some are large in magnitude. Thus, after identifying the CpGs with significant differences, the remaining CpGs are still likely to yield additional signals, which need more further investigation. For this purpose, we first exclude those CpGs with significant differences in the following analysis. In particular, we exclude those CpGs with p-values less than 0.05, based on the univariate t-test, with a Bonferroni correction within each chromosome. The differences in the remaining CpGs are of the "dense, but weak" pattern.

6. Conclusion

The classical two-sample tests for high-dimensional mean vectors are often designed to focus on sparse or dense mean differences. However, the sparsity level of mean differences is often unknown. In addition, the mean differences can have varying magnitudes, but are often assumed to be equal in existing methods. Here, we propose a robust test, capable of performing relatively well without the assumptions on the mean differences or the magnitude of each component. The proposed test comprises two steps: dynamically allocating weights to components with varying magnitudes, and then combining multiple WCTs to be adaptive to different sparsity levels of the mean differences.

The proposed test, the AWCT, can be viewed as a generalization of the GCT, which places equal weight on each component. Furthermore, the AWCT shares the idea of the ASPU by optimizing the power among a class of tests. Our simulation studies and real examples both demonstrate that the proposed test achieves good overall performance with a wide variety of signal sparsity, especially for the medium case, as opposed to existing approaches that focus on either sparse or dense signals.

Supplementary Material

The online Supplementary Material includes the Appendix (Proofs of Main Theorems), related proofs and additional numerical results.

Acknowledgments

The authors thank the co-editor Professor Su-Yun Huang, the associate editor, and the two anonymous reviewers for their helpful and constructive comments. The work of Dr. Shu was funded by the Science and Technology Development Fund of Macau SAR (FDCT/0033/2020/A1), the Department of Science and Technology of Guangdong Province (EF020/FBA-SLJ/2022/GDSTC), and the University of Macau Research Committee (MYRG2022-00017-FBA). Jinfeng Xu's research was supported by the General Research Fund (17308820) of Hong Kong, a start-up grant for a new faculty at the City University of Hong Kong (7200742), and the National Natural Science Foundation of China (72033002).

References

Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica* **28**, 43–62.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M. et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295.

Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis* **123**, 160–171.

Brockwell, P. J. and Davis, R. A. (2013). *Time Series: Theory and Methods*. Springer Science & Business Media.

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

Capizzi, G. and Masarotto, G. (2003). An adaptive exponentially weighted moving average control chart. *Technometrics* **45**, 199–207.

Chakraborty, A. and Chaudhuri, P. (2017). Tests for high-dimensional data based on means, spatial signs and spatial ranks. *The Annals of Statistics* **45**, 771–799.

Chang, J., Zheng, C., Zhou, W.-X. and Zhou, W. (2017). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics* **73**, 1300–1310.

Chen, S. X. and Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L. et al. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587.

Dutter, R. and Huber, P. J. (1981). Numerical methods for the nonlinear robust regression problem. *Journal of Statistical Computation and Simulation* **13**, 79–113.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association* **91**, 674–688.

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V. and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association* **110**, 837–849.

Holland, P.W. and Welsch, R.E. (1977). Robust regression using iteratively reweighted least-squares. *Communi-Cations in Statistics-Theory and Methods* **6**, 813–827.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics* **2**, 360–378.

Ishii, A., Yata, K. and Aoshima, M. (2019). Inference on high-dimensional mean vectors under the strongly spiked eigenvalue model. *Japanese Journal of Statistics and Data Science* **2**, 105–128.

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C. et al. (2006). Cancer statistics, 2006. *CA: A Cancer Journal for Clinicians* **56**, 106–130.

Kumar, A., Zhang, X., Zhang, Q. X., Jong, M. C., Huang, G., Vincent, L. W. S. et al. (2011). Residual stress analysis in thin device wafer using piezoresistive stress sensor. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **1**, 841–851.

Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* **100**, 518–532.

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

Srivastava, M. S. and Kubokawa, T. (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis* **115**, 204–216.

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

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H. et al. (2010). Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research* **20**, 440–446.

Wang, L., Peng, B. and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association* **110**, 1658–1669.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika* **103**, 609–624.

Yu, Y., Zhu, H., Frantz, J., Reding, M., Chan, K. and Ozkan, H. (2009). Evaporation and coverage area of pesticide droplets on hairy and waxy leaves. *Biosystems Engineering* **104**, 324–334.

Zhang, J.-T., Guo, J., Zhou, B. and Cheng, M.-Y. (2020). A simple two-sample test in high dimensions based on L_2-norm. *Journal of the American Statistical Association* **115**, 1011–1027.

Yidi Qu

Department of Statistics & Actuarial Science, The University of Hong Kong, Hong Kong, China.

E-mail: u3533935@connect.hku.hk

Lianjie Shu

Faculty of Business Administration, University of Macau, Macau, China.

E-mail: ljshu@um.edu.mo

Jinfeng Xu

Department of Biostatistics, City University of Hong Kong, Hong Kong, China.

E-mail: jinfenxu@cityu.edu.hk

(Received April 2022; accepted January 2023)