AN ADAPTIVE WEIGHTED COMPONENT TEST FOR HIGH-DIMENSIONAL MEANS

Qu, Yidi; Shu, Lianjie; Xu, Jinfeng

Published in:
Statistica Sinica

Published: 01/10/2024

Document Version:
Post-print, also known as Accepted Author Manuscript, Peer-reviewed or Author Final version

Published version (DOI):
10.5705/ss.202022.0143

Publication details:
Qu, Y., Shu, L., & Xu, J. (2024). AN ADAPTIVE WEIGHTED COMPONENT TEST FOR HIGH-DIMENSIONAL MEANS. *Statistica Sinica*, *34*(4), 1951-1971. https://doi.org/10.5705/ss.202022.0143

Qu, Y., Shu, L., & Xu, J. (Accepted/In press). AN ADAPTIVE WEIGHTED COMPONENT TEST FOR HIGH-DIMENSIONAL MEANS. Statistica Sinica. https://doi.org/10.5705/ss.202022.0143.

Statistica Sinica

AN ADAPTIVE WEIGHTED COMPONENT TEST FOR HIGH-DIMENSIONAL MEANS

Yidi Qu, Lianjie Shu, and Jinfeng Xu

Abstract:

Two streams of two-sample tests for high-dimensional data have been recently studied, including the sum-of-squares-based and supremum-based tests. The former stream of tests is more powerful against dense differences in two population means, while the latter is more powerful against sparse differences. However, the level of sparsity and signal strength are often unknown in practice. It is unclear which type of tests should be applied. This paper develops an adaptive weighted component test (AWCT) to provide an overall good power against a wide variety of alternative hypotheses with unknown sparsity level and varying signal strengths. The basic idea of it is to first allocate different weights onto components with varying magnitudes in a sum-of-squares-based test, and then to combine multiple weighted component tests (WCTs) to make the underlying test adaptive to different sparsity levels of the mean differences. The asymptotic properties of the proposed test are studied. Numerical comparisons demonstrate the superior performance of the proposed test across a wide 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 are the same in two populations. 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 samples from two populations having mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ 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, i.e., $n = n_1 + n_2$.

In low dimensional cases, i.e., $p \ll n$, there are some methods developed to test the difference in mean vectors between two populations. The classical T^2 test of Hotelling (1931), which has desirable properties and satisfactory power in conventional low-dimensional cases. However, with rapid advances in sensing technology and data acquisition systems, high-dimensional data appear in many settings. In high-dimensional cases, the dimension of data can exceed the number of sampled observations, i.e., $p > n$, leading to the so-called "large-p-small-n" problem. For example, the genetic data can

contain thousands of DNA segments from only about one or two hundred patients in the sample (Chen and Qin (2010)). In a 200-mm 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, the traditional multivariate two-sample tests such as the T^2 test, either cannot be directly applied or have too low power. It is straightforward to see that 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 or likelihood ratio tests may have power decreasing to the Type I error rate as p increases, even for the simple one-sample test on the mean of a normal distribution with known covariance matrix.

Various two-sample tests for high-dimensional data have been proposed in the literature, which can be grouped into two categories: the sum-of-squares-based tests and the supremum-based tests. The first category is motivated by the L_2-type distance between two mean vectors where all entries are considered. A sample of research in the first category is listed

below. Several researchers have attempted to extend the T^2 statistic to the case with $p > n$ by replacing the sample covariance matrix with a nonsingular matrix. For example, Bai and Saranadasa (1996) proposed a straightforward procedure (referred to here as the BS test) by replacing the sample covariance matrix with an identity matrix. In order to simplify the theoretical derivation, Chen and Qin (2010) suggested a test (referred to here as the CQ test) to remove the cross-product terms from the BS test. To account for possibly varying variances of the components of the data, one may replace the sample covariance matrix by a diagonal version; see, for example, Srivastava and Du (2008), Srivastava (2009), and Srivastava and Kubokawa (2013). In order to avoid full estimation of the covariance matrix, Gregory et al. (2015) proposed a generalized component test (GCT) by assuming that the p components admit a logical ordering such that the dependence between components is related to their displacement. Moreover, to accommodate the strongly spiked eigenvalues (SSE) in high-dimensional data, Aoshima and Yata (2018) and Ishii et al. (2019) proposed distance-based tests which utilize the estimated eigen-structures and obtained their limiting distributions. Zhang et al. (2020) proposed a Welch-Satterthwaite χ^2 type test to further relax the restrictive assumptions on the covariance structure. Other approaches involve the use of the

random projection method Srivastava et al. (2016)), the interpoint distance (Biswas and Ghosh (2014)), and the spatial sign ranks (Wang et al. (2015), Chakraborty et al. (2017)). The second category is motivated by the L_∞-type distance between two mean vectors where only the largest deviation is utilized. A sample of research in this category includes Chang et al. (2017) and the CLX test proposed by Cai et al. (2014).

However, these two streams of tests are designed for two extreme situations, respectively. In particular, the first category is particularly efficient in the dense situation when almost all the components in the two mean vectors exhibit some differences. In contrast, the second category is particularly efficient in the sparse situation when a few leading components in the two mean vectors suffer from substantial changes. As a result, no one test can perform relatively well in both dense and sparse mean differences in $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

In reality, the sparsity level of mean differences, i.e., the number of zero elements in $\boldsymbol{\mu_1} - \boldsymbol{\mu_2}$, is often unknown. Also, the sparsity level could be in-between the two extreme situations, which is not that dense or not that sparse. Therefore, it is unclear how to choose a powerful test between the above two categories when the sparsity level of mean differences is unknown. Moreover, most of the above tests often focus discussions on the case with

equal signal strength (or the same magnitude) for each component of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. To get away from the assumptions of known sparsity level of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and equal shift magnitude in each component, it is more reasonable to develop 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 level and varying magnitudes of mean differences.

The proposed testing procedure involves two steps. The first step is to introduce a robust weighting function capable of allocating different weights onto components of varying magnitudes in a sum-of-squares-based test. This naturally generalizes the GCT with equal weight onto each component as a special case. Intuitively, this can improve 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 is to combine multiple weighted component tests (WCTs) obtained in the first step to select the most powerful test among the candidate tests. This step can make the proposed test adaptive to different sparsity levels of mean differences. This step is similar to the idea of the adaptive sum-of-powers test (referred to here as the ASPU test) of Xu et al. (2016). For simplicity, we denote the adaptive WCT as AWCT

throughout the remaining of the paper.

Note that our approach differs from the ASPU test in two important aspects. First, the proposed approach can dynamically allocate weights onto components according to 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 as it is more reasonable to assume the components to suffer from shifts of different magnitudes in practice. Second, although both the ASPU and AWCT tests make a combined use of multiple individual tests to improve the test power when the sparsity level of signal is unknown, the individual tests are working in a different way. 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 distances, while the individual WCT test do it by tuning the weighting parameter of a robust weight function such as Huber's function. Therefore, the proposed approach can also be expected to provide an overall good test power when the components have varying magnitudes of mean shifts, in addition to its robustness to the sparsity level of 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, compar-

ing its performance with that of the BS, CQ, GCT, CLX, and ASPU in terms of power and maintenance of nominal size. Section 5 presents two real examples. Concluding remarks are presented in Section 6. Proofs of asymptotic theories are provided in the supplementary materials.

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 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

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

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 can test the mean difference in the jth component. To take all the signal information, one can compute the sum of t_j^2 over all components like the GCT statistic, $j = 1, \ldots, p$. However, the components often have varying magnitudes. It is reasonable to put larger weights onto large components for improving the power of the test statistic. For this

purpose, we establish the WCT statistics as follows:

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

where ω_j is weight allocated onto t_j. Clearly, the WCT statistic is a natural generalization of the GCT statistic as it allows for different weights onto different components t_j. When ω_j is fixed as a constant, equal weight is put on all the components. In this case, the WCT performs essentially like the GCT.

Different weighting functions can be used. Here we consider some weight motivated by robust procedures such as Huber's function (Dutter and Huber (1981)) and Welsch's function (Holland and Welsch 1977). In this paper, for the sake of simplicity, we restrict our discussion to Huber's weight function as follows:

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

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 $R \to \infty$, $\omega_j = \kappa$. The same weight is allocated on t_j^2 along each component. Therefore, one cannot choose too large value of R in practice in order to adaptively allocate weights onto 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 to exceed R. Notices 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 11.38% probability for it to exceed $R = 2.5$ and 6.13% probability to exceed $R = 3.5$, respectively. In this paper, we follow this practice to choose $R \in [2.5, 3.5]$, with primary focus on $R = 3$ for simplicity.

The parameter κ controls the relative weight allocated onto 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 $R = 3$. As can be seen from Figure 1, a smaller κ value allocates relatively small weights onto smaller components t_j^2 but relatively large weights onto 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 $t_j^2 \leq R$ and $\omega_j = 1 - R/t_j^2$ for $t_j^2 > R$. This implies we only consider the extremely large components t_j^2 in the WCT statistic but 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 with 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

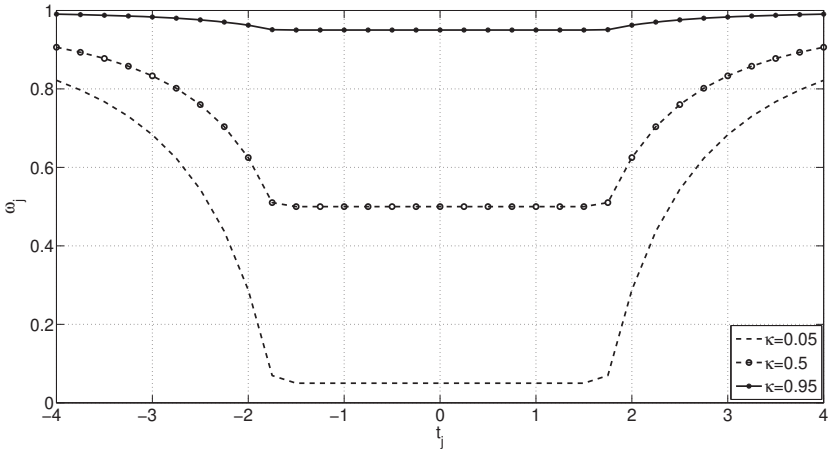

Figure 1: Plot of ω_j under different values of κ when $R = 3$.

for all the components. Therefore, one can expect the WCT to perform essentially like the GCT, which has good test power in the case with dense signals.

The above discussion indicates that the parameter κ has an important impact on the power of WCT. The WCT statistic in Equation (2 .1) can be rewritten as

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

Whether $T_{WCT}(\kappa)$ is powerful depends on the unknown sparsity level, i.e., pattern of nonzero signals. To provide an overall good power, one can incorporate multiple testing in the procedure so that at least one of them would yield a high power for a particular application with unknown truth.

This can actually be achieved by combining multiple WCTs as follows:

$$T_{AWCT} = T_{WCT}(\arg\min_{0\leq\kappa\leq 1} P(\kappa)),$$

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

In practice, one has to choose some candidate values of κ for the proposed test to improve the test performance when the sparsity level of signal is unknown. In principle, one can choose many candidate values of κ. However, this would greatly complicate the underlying test but may only improve the test power marginally. To make a trade-off between simplicity and test power, we just 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 choice of candidate values of κ can be similarly analyzed. As will be shown later in this paper, such choice of $\kappa \in \Gamma = \{0.05, 0.5, 0.95\}$ can provide 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 $\{Z^j : j \in [a, b]\}$, i.e., $\mathcal{F}_a^b = \sigma\left\{Z^a, Z^{a+1}, \ldots, Z^b\right\}$, where Z^j denotes the jth element of $\boldsymbol{Z}$. For all positive integers $s < p$, the strong mixing coefficients are defined as

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

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

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

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

$$\max_{k=1,2; 1 \le i \ne j \le p} \left| \sigma_{k,ij} \right| / \left(\sigma_{k,ii} \sigma_{k,jj} \right)^{1/2} < 1 - \eta$$

for some $\eta > 0$.

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

C.3 $n_1/n_2 \to c \in (0, \infty)$ and $p = o(n^2)$; $\max_{1 \leqslant j \leqslant p} E\left[\exp\left\{h\left(X_{k1}^j - \mu_k^j\right)^2\right\}\right] < \infty$ 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 eigenvalues and covariance 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 datasets whose components admit an ordering in time, space or some other index such that their dependence diminishes as components are further apart. Taking the methylation values as an example, the measurements are taken along a chromosome. The location of each measurement is recorded, and this provides an index over which dependence could be modeled. Under C.1-C.3, the asymptotic normality of the test statistic T_{WCT} and its asymptotic joint distribution are derived in the following theorems, respectively.

Theorem 1. *Assume that the conditions C.1-C.3 hold. Under H_0, we have*

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

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

Proof. See Appendix.

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

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

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

Proof. See Appendix.

Denote $I_j = I(t_j^2 \leq R)$, and rewrite the mean of 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*

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

where $F(x)$ *and* $f(x)$ *denote the cumulative distribution function and probability density function of the* χ_1^2 *distribution, respectively. Thus, the term* $\int_0^\infty x f(x)dx$ *equals 1 and would be replaced by 1 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 Materials. Then denote $K_i = (\kappa-1)I_i t_i^2 + t_i^2 + (1-\kappa)RI_i$, then, $\zeta^2 = p^{-1}Var(\sum_{j=1}^p \omega_j t_j^2) = p^{-1}\sum_{j=1}^p Var\{K_j\} + p^{-1}\sum_{i\neq j} Cov\{K_i, K_j\}$.

Proposition 2. *Assume that the conditions C.1-C.3 hold. Under* H_0*, we have*

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

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

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

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

We estimate ζ^2 by

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

where $\mathfrak{p}(x)$ is a piecewise function of x such that $\mathfrak{p}(0) = 1$, $|\mathfrak{p}(x)| \leq 1$ for all x, and $\mathfrak{p}(x) = 0$ for $|x| > 1$, and L is a user-selected lag window size. Here we utilize the Parzen window (Brockwell and Davis (2013)), i.e.,

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

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

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

Proposition 3. *Assume that the conditions C.1-C.3 hold. Under H_0, for $0 \leq \kappa_s, \kappa_t \leq 1$, we have*

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

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

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

$$- (2 - \kappa_s - \kappa_t) \int_0^R F(x)dx - 1 + O(1/n).$$

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*

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

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

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

3 .2 Asymptotic type I error and power analysis

Denote $T = \sqrt{p}(T_{WCT} - \nu)/\zeta$. Assuming that the 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

$$\begin{aligned}p =& pr(T_{AWCT} > C | H_0 \text{ true}) \\ =& 1 - pr(T_{AWCT} \leq C | H_0 \text{ true}) \\ =& 1 - pr(\max_{0 \leq i \leq d} T_i \leq C | H_0 \text{ true})\end{aligned}$$

$$
\begin{aligned}
&=1-pr(T_1 \le C, T_2 \le C, \ldots, T_d \le C | H_0 \text{ true}) \\
&=1-\int_{(-\infty, C)^d} \phi_d(\mathbf{0}, \boldsymbol{\Omega}) dT_1 \ldots dT_d,
\end{aligned}
$$

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

The test power of T_{AWCT} under H_A satisfies $pr(\min_{0 \le \kappa \le 1} P(\kappa) < \alpha) \ge pr(P(\kappa) < \alpha)$ for any $0 \le \kappa \le 1$, where α is the significance level. Therefore, the asymptotic power of the proposed test is one if there exists $0 \le \kappa \le 1$ such that $pr(P(\kappa) < \alpha) \to 1$, that is, $T_{WCT}(\kappa)$ has the asymptotic power equal to one. Hence, to study the asymptotic power of the adaptive test, we only need to focus on the power of $T_{WCT}(\kappa)$ for $0 \le \kappa \le 1$. In the following, we write $T_{WCT}(\kappa)$ as T_{WCT} for conciseness. Also, 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 $H_A : \boldsymbol{\mu}_1 \ne \boldsymbol{\mu}_2$ means that an unknown proportion q $(0 < q \le 1)$ of $\iota_j's$ is not equal to zero. Denote $\nu_A = E(T_{WCT} | H_A \text{ true})$, then the power of the

WCT, i.e., $P\left(\sqrt{p}\left(T_{WCT}-\nu_A\right)/\hat{\zeta} > z_\alpha | H_A \text{ true}\right)$, is equal to

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

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

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

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 of t_j^2 under ι_j, respectively. 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 non-central chi-square distribution with degree of freedom 1 and noncentrality parameter of ι_j^2, denoted as $\chi_1^2(\iota_j^2)$. According to Proposition 1,

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

$$=p^{-1}\sum_{j=1}^{p}\left\{a_{\kappa,R}(\tau_j)\iota_j^2+O(n^{-1})\right\},$$

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

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

4 . Simulation Studies

In this section, we illustrate the performance of the proposed test, AWCT, by comparing it with some existing methods through simulations. The tests in comparison include the BS, CQ, GCT, and ASPU tests, all of which belong to the sum-of-squares-based tests. Also, the CLX test aimed at testing sparse alternatives is included for comparison. The performances were 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 0 under the alternative hypothesis, where $\beta\in[0,1]$ controls the signal sparsity. Three values of $\beta=0.3,\ 0.5,\ 0.7$ were considered, which correspond to the cases with dense, medium, and sparse differences in 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 to control the signal strength; and (ii) μ_2^i linearly increases 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 $i = 1, 2, \ldots, m$.

Three specific models for the covariance structure chosen from Cai et al. (2014) were considered, which are given as follows:

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

(b) $\boldsymbol{\Sigma} = (\sigma_{i,j})$, 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) $\boldsymbol{\Sigma} = (\sigma_{i,j})$ 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, while it 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 $\{\boldsymbol{X}_{2j}\}_{j=1}^{n_2}$ were generated from a multivariate distribution with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively and a common covariance matrix $\boldsymbol{\Sigma}$. The dimension p

takes $p = 400$ and the sample sizes take $n_1 = n_2 = 200$. To illustrate the effects of distributions, three different types of distributions are studied, including (i) the multivariate normal, (ii) the multivariate t distribution with degrees of freedom of $v = 3$, and (iii) a multivariate gamma distribution. The functions `rmvnorm` and `rmvt` from R package `mvtnorm` as well as the function `rmvgamma` from package `lcmix` were used to generate the three types of distributions, respectively. 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 the shape parameter of 4 and scale parameter of 2 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 choice of L and R in our proposed test, the results are found to be qualitatively the same for $L = 10, 20$ and 30 as well as for $R = 2.5$, 3 and 3.5. Also, the results are similar under different covariance matrix structures. For the sake of simplicity, we only present the results based on $L = 10$ and $R = 3$ under covariance Model (a). The power and empirical Type I error rate are calculated from 1000 replications, respectively.

Table 1: The empirical type I error rates of various tests under 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 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 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

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, i.e., $c = p/n$. Both the results based on 1,000 and 2,000 replicates are presented. It is clear 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 remaining of the paper.

In addition, we also compare the computation time among the AWCT, ASPU, and GCT tests, as suggested by a referee. Take the case with $p = 400$ and $n_1 = n_2 = 200$ as an example. Under 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. Clearly, both the GCT and AWCT tests are more computationally efficient than the ASPU test.

Table 1 shows that under the multivariate normal distribution, nearly all the tests can maintain very close-to-nominal Type I error rates. Only the GCT exhibits inflated Type I error rates, which is perhaps due to its low convergence rate to asymptotic null distribution. Tables 2 - 3 further present the empirical type I error rates of the above tests under the multivariate gamma and t_3 distributions, respectively. As can be seen from Table 2, 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 also fails to maintain the nominal type I error rate, while the others can 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 $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Under the case with dense signals ($\beta = 0.3$), the AWCT has the highest power and the CLX has the lowest power. This is not surprising

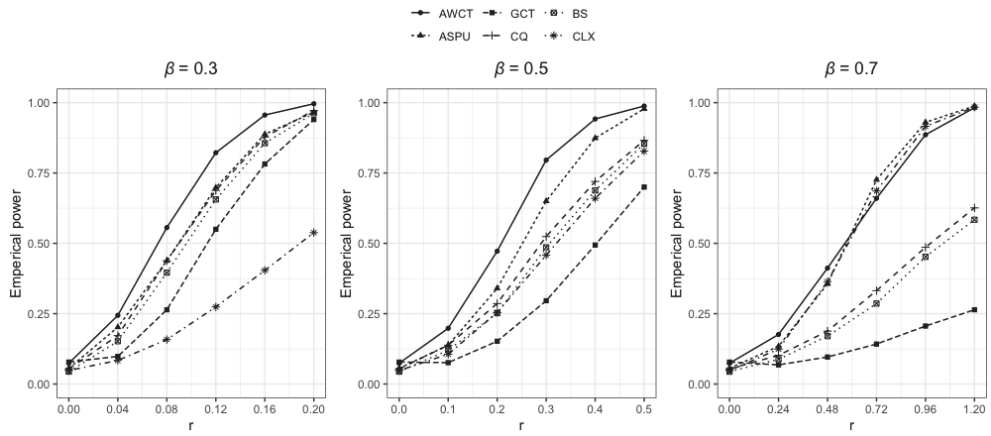

Figure 2: Power curves of the various tests against r under different sparsity levels of β based on Model (a) with normal innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

as the CLX is a supremum-based test, which is less efficient in detecting dense signals. When β increases to $\beta = 0.5$, the AWCT has higher power than ASPU, and than CQ and BS, followed by CLX and GCT. In this case, the GCT has the lowest power. This illustrates that the power of the GCT decreases substantially as the sparsity level of signals increases. When β further increases to $\beta = 0.7$, the AWCT, ASPU and CLX methods have very competitive power, and perform far better than the CQ and BS methods, and than the GCT method. 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

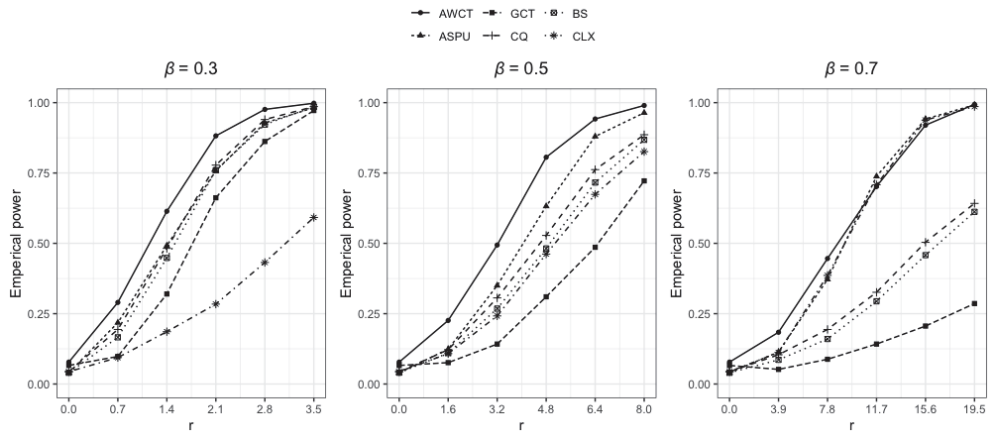

Figure 3: Power curves of the various tests against r under different sparsity levels of β based on Model (a) with centered gamma(4, 2) innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

the case with normal innovations.

To illustrate the effect of heavy-tailedness on the performance of the proposed test, Figure 4 further displays 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 did not differ greatly from those of the normal- and skewed-innovations.

To sum up, Figures 2 – 4 indicates 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 an overall good power in a wide variety of situations. The simulation results under Models (b) and (c) are placed in the supplementary materials as they

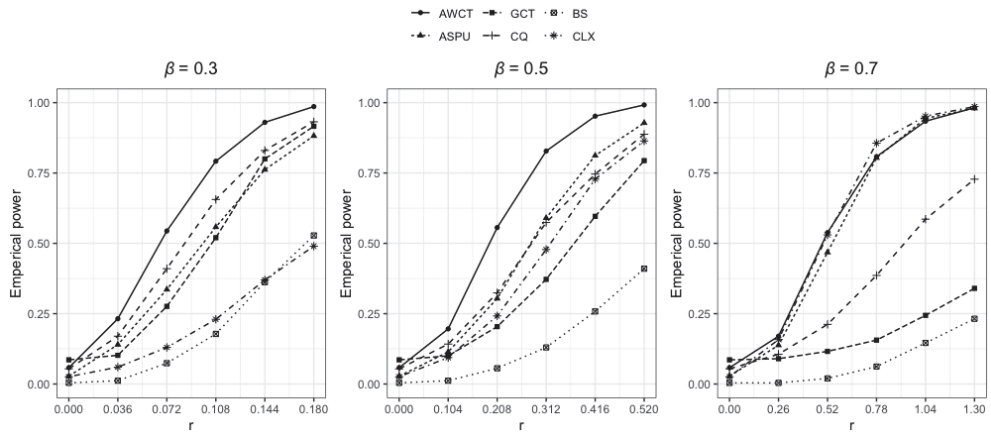

Figure 4: Power curves of the various tests against r under different sparsity levels of β based on Model (a) with multivariate t_3 innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

are quite 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 will not occur. The size of a test is expected to be robust to any scaling of the variances. To investigate the impact of heteroscedasticity on the performance of the above tests, following the method of Gregory et al. (2015), the standard deviation of each component was scaled by the square root of a realization from the exponential distribution with mean 1/2 shifted to the right by 1/2. Thus,

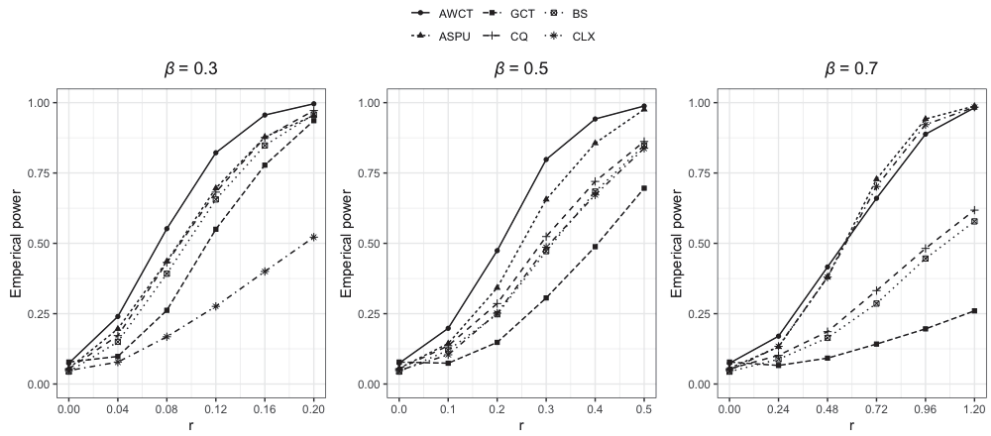

Figure 5: Power curves of the various tests against r under unequal magnitudes of mean differences based on Model (a) with multivariate normal innovations and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ when $\beta = 0.3, 0.5, 0.7$.

the average scaling is 1 and the scaled variances are bounded away from 0.

The power simulation utilizing the centered gamma(4,2) under Model (a) was repeated under the heteroscedastic condition, which is shown in the supplementary materials for simplicity. As can be seen from the results, the AWCT method can still maintain an overall good power under the heteroscedastic condition when comparing with other tests.

4 .4 Performance under unequal magnitudes of mean differences

The above analysis mainly focuses on the case with equal magnitude for the nonzero-mean differences. It is also of interest to investigate the per-

formance under the case with unequal magnitudes for the nonzero-mean differences, which is more general and natural in practice. A potential benefit of the AWCT is capable of allocating different weights to components with varying magnitudes compared to the GCT. Therefore, when the true mean differences between the two populations have unequal magnitudes, the AWCT method is expected to perform far better than the GCT.

Figure 5 displays 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 can be seen from Figure 5, the AWCT outperforms the GCT, regardless of the value of β.

5 . Real Data Analysis

In this section, we apply the above to two real datasets: DNA methylation dataset and the dataset from a semi-conductor manufacturing process. Both datasets are publicly available. The first can be downloaded from the NCBI GEO website with GEO number GSE19711, the second dataset is available in the UC Irvine Machine Learning Repository `https:`

//archive.ics.uci.edu/ml/datasets/SECOM. We present the application to DNA methylation data and the application to a semi-conductor manufacturing process is given in the Supplementary Materials. Death from ovarian cancer among women ranks fifth in the United States (Jemal et al. (2006)) and was found to be associated with aberrant DNA methylation. A genome wide DNA methylation profiling of 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 methylation Beadchip v1.2 with 27578 CpGs from 540 whole blood samples, which include 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 methylated allele ($\text{Max}(M, 0)$) and unmethylated allele ($\text{Max}(U, 0)$), given by

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

The β-values are usually preprocessed for the downstream statistical analysis, including 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 0 and 1. In an ideal situation, "0" indicates that no copy of the CpG site in the sample is methylated. The value "1" 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. Alternatively, some investigators use the M-value, considering the M-value alternative statistically more valid (Du et al. (2010)). However, the interpretation of M-values is not as intuitive as for β-values. For this reason, we restrict discussion to β-values.

The above six tests, including AWCT, ASPU, GCT, CQ, BS and CLX , were each applied to test whether there is significant difference in DNA methylation levels between the cancer group and the normal group. The 27578 CpGs of the ovarian cancer data are from all the 23 pairs of chromosomes including the sex chromosomes, chromosome X and Y. We excluded the chromosome Y in our analysis since there are only 7 CpGs from the chromosome Y, where the sample size is larger than the dimension of data. Prior to analysis, each missing value was 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 in testing the equality of the methylation levels measured by the β-values on each chro-

mosome. The R value is set to be 3 for the AWCT. Nearly all the tests can reject the null hypothesis at the 5% significance level. The only exception includes the BS test on Chromosome 16 and 19. The p-values of the AWCT, ASPU, and GCT methods are nearly 0 for all chromosomes.

The small p-values in Table 4 indicate that the differences of DNA methylation level on each CpGs between the cancer and the normal group are dense and that some of them are large in magnitude. So after identifying those CpGs with significant differences, the remaining CpGs are still likely to yield additional signals, which also need more detailed investigations. 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 Bonferroni correction within each chromosome. The differences in the remaining CpGs are of the "dense but weak " pattern.

6 . Conclusions

The classical two-sample tests for high-dimensional mean vectors are often particularly designed to test sparse or dense mean differences. However, the sparsity level of mean differences is often unknown. Also, the mean differences can have varying magnitudes, while they are often assumed to

Table 4: The p-values of the various tests for testing equality of the DNA methylation levels measured by 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}	

be equal in the existing literature. It is reasonable to develop a robust test capable of performing relatively well without the assumption of dense or sparse mean differences and the assumption of equal magnitude for each component. For this purpose, this paper develops a new test consisting of two steps: dynamically allocating weights onto components with varying magnitudes and combining multiple weighted component tests (WCTs) to be adaptive to different sparsity levels of mean differences.

The proposed test, AWCT, can be viewed as a generalization of the GCT that puts equal weight on each component. Also, the AWCT shares the idea similar to the ASPU by optimizing the power among a class of tests. Both the simulation studies and real examples demonstrate that the proposed test can achieve an overall good performance with a wide variety of signal sparsity especially for the medium case. As a comparison, existing approaches often cater for a particular situation where signals are either sparse or dense.

Supplementary Materials

The online supplementary materials include the Appendix (Proofs of Main Theorems), related proofs and additional numerical results.

Acknowledgments

The authors sincerely thank the Co-editor Professor Su-Yun Huang, associate editor, and two anonymous reviewers for their very 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 General Research Fund (17308820) of Hong Kong, Start-up grant for new faculty at City University of Hong Kong (7200742), and the National Natural Science Foundation of China (72033002).

References

Aoshima, M. and K. Yata (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica 28*(1), 43–62.

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

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Bibikova, M., B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, et al. (2011). High density dna methylation array with single cpg site resolution. *Genomics 98*(4), 288–295.

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

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

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

Capizzi, G. and G. Masarotto (2003). An adaptive exponentially weighted moving average control chart. *Technometrics 45*(3), 199–207.

Chakraborty, A., P. Chaudhuri, et al. (2017). Tests for high-dimensional data based on means, spatial signs and spatial ranks. *The Annals of Statistics 45*(2), 771–799.

Chang, J., C. Zheng, W.-X. Zhou, and W. Zhou (2017). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics 73*(4), 1300–1310.

Chen, S. X. and Y. L. Qin (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics 38*(2), 808–835.

Du, P., X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin (2010). Compar-

ison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics 11*(1), 587.

Dutter, R. and P. J. Huber (1981). Numerical methods for the nonlinear robust regression problem. *Journal of Statistical Computation and Simulation 13*(2), 79–113.

Fan, J. (1996). Test of significance based on wavelet thresholding and neyman's truncation. *Journal of the American Statistical Association 91*(434), 674–688.

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

Hotelling, H. (1931). The generalization of student's ratio. *Ann. Math. Statistic. 2*, 360–378.

Ishii, A., K. Yata, and M. Aoshima (2019). Inference on high-dimensional mean vectors under the strongly spiked eigenvalue model. *Japanese Journal of Statistics and Data Science 2*(1), 105–128.

Jemal, A., R. Siegel, E. Ward, T. Murray, J. Xu, C. Smigal, and M. J. Thun (2006). Cancer statistics, 2006. *CA: a cancer journal for clinicians 56*(2), 106–130.

Kumar, A., X. Zhang, Q. X. Zhang, M. C. Jong, G. Huang, L. W. S. Vincent, V. Kripesh, C. Lee, J. H. Lau, D. L. Kwong, et al. (2011). Residual stress analysis in thin device wafer using piezoresistive stress sensor. *IEEE Transactions on Components, Packaging and Manufacturing Technology 1*(6), 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*(3), 518–532.

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

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

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

Teschendorff, A. E., U. Menon, A. Gentry-Maharaj, S. J. Ramus, D. J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C. G. Bell, A. P. Maxwell, et al. (2010). Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research 20*(4), 440–446.

Wang, L., B. Peng, and R. Li (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association 110*(512), 1658–1669.

Xu, G., L. Lin, P. Wei, and W. Pan (2016). An adaptive two-sample test for high-dimensional means. *Biometrika 103*(3), 609–624.

Yu, Y., H. Zhu, J. Frantz, M. Reding, K. Chan, and H. Ozkan (2009). Evaporation and coverage area of pesticide droplets on hairy and waxy leaves. *biosystems engineering 104*(3), 324–334.

Zhang, J.-T., J. Guo, B. Zhou, and M.-Y. Cheng (2020). A simple two-sample test in high dimensions based on l 2-norm. *Journal of the American Statistical Association 115*(530), 1011–1027.

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

E-mail: (u3533935@connect.hku.hk)

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

E-mail: (ljshu@um.edu.mo)

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

E-mail: (jinfenxu@cityu.edu.hk)