Statistica Sinica **35** (2025), 293-319
doi:https://doi.org/10.5705/ss.202022.0348

POWER BOOSTING: FUSION OF MULTIPLE TEST STATISTICS VIA RESAMPLING

Efang Kong[#*1], Yu Liu[#2] and Yingcun Xia[1,3]

[1] *University of Electronic Science and Technology of China,*
[2] *Sichuan Normal University and* [3] *National University of Singapore*

Abstract: For the same null hypothesis, there usually exist multiple valid test statistics. In nearly all cases, any individual statistic is only powerful against specific types of alternatives, and could be rather weak in picking up signals of other types. It is thus crucial, especially in high-dimensional settings, to combine the information contained in different test statistics in order to maintain robust power against a wide range of alternatives, thus avoiding the worst-case scenario. Methods have been proposed for similar purposes, but they are either computationally expensive or lack theoretical justification. In this paper, we present a general and easy-to-implement procedure for fusing multiple valid statistics using resampling methods, such as bootstrap or permutation. The consistency of this procedure is proved for three popular high-dimensional hypothesis testing problems. The results of numerical studies show that this fusion procedure maintains robust performance against a wide range of alternatives, whereas individual test statistics often suffer from extremely low power.

Key words and phrases: Consistency of test, high-dimensional data, independence test, permutation, two-sample mean comparison.

1. Introduction

Testing high-dimensional null hypotheses has been the subject of intensive studies. One popular approach, which includes the works of Kosorok and Ma (2007), Bancroft, Du and Nettleton (2013), and Liang (2016), breaks the null hypothesis into multiple univariate tests, and focuses on the false discovery rate. For studies on power, the family-wise error, Kim and Akritas (2010) note that for any given null hypothesis, there usually exist multiple valid statistics, each of which may detect certain types of signals, but suffer from very low power against others. Thus the test statistic and types of alternatives are connected in terms of power enhancement or boosting. For example, with the alternative restricted to be sparse, Fan et al. (2015) shows how a given test statistic can be made consistent and more powerful for cross-sectional data. This idea of possible power enhancement against specific alternatives is later examined in a more general framework by Kock and Preinerstorfer (2019). We study a similar problem of

[#]Contributed equally to this work.
*Corresponding author. E-mail: kongefang@uestc.edu.cn

power boosting from a different, yet more practical angle. We propose an efficient procedure for fusing statistics that could ensure robust power performance against arbitrary alternatives, thus avoiding the worst-case scenario. In this sense, fusing test statistics is particularly useful in practice when choosing between opposing recommendations made based on different test statistics.

To formulate the setup, suppose $\{T_{n,k}, k = 1, \ldots, K\}$ is a collection of statistics, where K is a fixed integer, such that for any given k, H_0 is rejected for large $T_{n,k}$. Note that a naive form of combination, such as a weighted average $\sum_{k=1}^{K} a_k T_{n,k}$, with $a_k \geq 0$, is not a good choice, because it is difficult to specify appropriate values for the coefficients a_k so that the statistical significance of one $T_{n,k}$ is not obscured by trivial variations in other $T_{n,k}$ of a larger scale. This is one of the motivating factors behind the monotone transformation of individual statistics to make them relatively comparable before being combined. One example is Fisher's combined p-value

$$U_n := -2 \sum_{k=1}^{K} \log\{1 - F_{n,k}(T_{n,k})\}, \tag{1.1}$$

where $F_{n,k}(.)$ is the null distribution function of $T_{n,k}$. Its relative popularity is largely because it follows a $\chi^2(.)$ distribution if $T_{n,k}$, for $k = 1, \ldots, K$, are independent. Another related example is an equivalence of the smallest p-value:

$$U_n := \max_{k=1,\ldots,K} F_{n,k}(T_{n,k}), \tag{1.2}$$

and H_0 is rejected whenever the p-value associated with some $T_{n,k}$ is too small. Examples of fusion statistics like (1.1) and (1.2) both suggest that transforming $T_{n,k}$ using its distribution function into a uniform $(0, 1)$ is a reasonable choice. However, be it (1.1) or (1.2), in practice, the unknown $F_{n,k}(.)$ has to be replaced with their respective estimates first in order to obtain an empirical version $\hat{U}_n$. The biggest challenge in their use is to obtain an efficient approximation of the null joint distribution of $\{T_{n,k}, k = 1, \ldots, K\}$. Using (1.2) in a high-dimensional setting is discussed in Xu et al. (2016) for the two-sample mean comparison problem, where the approximation of the null distribution is obtained using the standard two-step procedure: first, derive the (asymptotic) form of $F_{n,k}(.)$ and $F_n(.)$, the latter being the (null) joint distribution of $\{T_{n,k}, k = 1, \ldots, K\}$; second, find the tail probabilities associated with these asymptotic (null) distributions using numerical approximations (with plugged-in estimates of the parameters). This classical two-step approach is not only computationally intensive, but also suffers from low numerical efficiency.

In this study, we investigate how to use resampling methods, either bootstrap or permutation, depending on the specific testing problem, to directly approximate the null distributions of U_n, or rather $\hat{U}_n$, for the purpose of fusing test

statistics in high-dimensional hypothesis testing, where the dimension of the data is not negligible relative to the sample size. A streamlined setup is as follows, with (1.2) as the fusion statistic. Let $\mathbf{X}_1^n = \{X_1, \ldots, X_n\}$ denote the original sample. With a sufficiently large number B, $\mathbf{X}_1^{n,(b)}$, for $b = 1, \ldots, B$, denotes B new samples generated using either bootstrap or permutation, for which H_0 holds true. For $b = 1, \ldots, B$ and $k = 1, \ldots, K$, let $T_{n,k}^{(b)}$ denote the values of the test statistic $T_{n,k}$ calculated from the sample $\mathbf{X}_1^{n,(b)}$. For any $k = 1, \ldots, K$, we estimate $F_{n,k}(.)$ by $\hat{F}_{n,k}(.)$, the empirical distribution function based on $\{T_{n,k}^{(1)}, \ldots, T_{n,k}^{(B)}\}$. An empirical version of (1.2) is then defined as

$$\hat{U}_n := \max_{k=1,\ldots,K} \hat{F}_{n,k}(T_{n,k}). \tag{1.3}$$

Next, we compare this with the empirical distribution function of its resampling counterpart:

$$\hat{U}_n^{(b)} := \max_{k=1,\ldots,K} \hat{F}_{n,k}(T_{n,k}^{(b)}), \quad b = 1, \ldots, B. \tag{1.4}$$

Lastly, at significance level α, we reject H_0 if

$$B^{-1} \sum_{b=1}^{B} I(\hat{U}_n^{(b)} \geq \hat{U}_n) \leq \alpha, \tag{1.5}$$

where $I(\cdot)$ denotes the indicator function. We say a statistical test is consistent if its type-I error is identical to the nominal significance level α, at least asymptotically. In this paper, we prove the consistency of the above fusion procedure, namely, (1.3)–(1.5), in the context of three popular high-dimensional hypothesis testing problems, discussed in, among others Chung and Romano (2016), Cai, Liu and Xia (2014), and Heller, Heller and Gorfine (2013), for a selection of test statistics. Our main results are summarized as follows:

(i) we show the consistency of the empirical bootstrap-based fusion procedure for the one-sample mean test, where K is the number of statistics to be fused, and can increase with n;

(ii) we show the consistency of the permutation-based fusion procedure for the two-sample mean comparison, where K can also increase with n;

(iii) we show the consistency of the permutation-based fusion procedure for the test of independence between two random vectors; as a byproduct, we provide a theoretical justification for the practice in Heller, Heller and Gorfine (2013), where the permutation distribution of the HHG statistic is used to approximate its null distribution.

The rest of the paper is organized as follows. Section 2 and Section 3 present the one-sample mean test and the two-sample mean comparison, respectively.

Section 4 discusses testing the independence between two (high-dimensional) random vectors. A brief discussion on possible extensions is given in Section 5. Numerical results are given in Section 6. Regulation conditions and proofs are gathered in the Appendix.

2. Test of One-Sample Mean

Suppose $X_i \in R^p$, for $i = 1, \ldots, n$, are independent copies of $X = (X^1, \ldots, X^p)^\top$, with mean μ and covariance matrix Σ^X. Without loss of generality, suppose the diagonal elements of Σ^X are all ones. Testing $H_0 : \mu = 0$, referred to as the one-sample location model in Kock and Preinerstorfer (2019), is based on the sample mean $\bar{X}_n$, usually standardized by the sample covariance matrix. When p is large, so that the inversion of a $p \times p$ matrix is much less feasible, if at all possible, a more popular replacement is given by

$$\delta_n = (\delta_{n,1}, \ldots, \delta_{n,p})^\top = n^{1/2} \hat{D}_n^{-1/2} \bar{X}_n,$$

where $\hat{D}_n = \mathrm{diag}(\hat{\sigma}_{nj}^2, j = 1, \ldots, p)$ is a diagonal matrix of the sample variances. The use of $\hat{D}_n$ instead of the sample covariance matrix is to avoid having to compute the inverse of a high-dimensional matrix; see, for example, Bai and Saranadasa (1996), Srivastava and Du (2008), and Kong et al. (2022). For any integer $k \geq 1$, let $A_k(\cdot)$ be a function so that for any vector $\nu \in R^p$, $A_k(\nu)$ returns the average of its largest (in absolute value) k elements. Apparently, for any $k \geq 1$, $A_k(\delta_n)$ is a pivotal statistic, so that we reject H_0 if $A_k(\delta_n)$ is too large. However, as noted in Cai, Liu and Xia (2014), Kim and Akritas (2010), and Gregory et al. (2015), no statistic is uniformly more powerful than others (against all possible alternatives). For example, when the signals are sparse, but strong, $A_1(\delta_n)$, namely, the supremum statistic considered in Chernozhukov, Chetverikov and Kat (2019) and Cai, Liu and Xia (2014), has greater power than $A_k(\delta_n)$ with a large k, because the latter is not greatly influenced by a small number of large differences. Similarly, in the case of dense, but weak alternatives, $A_k(\delta_n)$ with a small k is not likely to be extreme enough to serve as evidence to reject H_0. Furthermore, as demonstrated in Kong et al. (2022), in the latter case, it is also beneficial to consider $A_k(\delta_n)$ with $k = s_n$, where s_n is some positive integer that can increase with n.

Without loss of generality, suppose $1 \leq l_1 \leq l_2 \leq \cdots \leq l_K \leq s_n$ is a sequence of positive integers. For $k = 1, \ldots, K$, let

$$T_{n,k} = T_{n,k}(\delta_n) = A_{l_k}(\delta_n), \tag{2.1}$$

be the corresponding sequence of statistics. We now show that they can be combined using the empirical bootstrap-based fusion procedure (1.3)–(1.5). For $b = 1, \ldots, B$, let $\mathbf{X}_1^{n,(b)} = \{X_1^{n,(b)}, \ldots, X_n^{n,(b)}\}$ be an empirical bootstrapped

sample, that is $X_i^{n,(b)}$, for $i = 1, \ldots, n$, are independent and identically distributed (i.i.d) draws (with replacement) from $\mathbf{X}_1^n = \{X_i, i = 1, \ldots, n\}$. Let $\bar{X}_n^{(b)} = n^{-1} \sum_i X_i^{n,(b)}$ denote the bootstrapped sample mean, and $\hat{D}_n^{(b)}$ the bootstrap version of $\hat{D}_n$. Write $\delta_n^{(b)} = n^{1/2}(\hat{D}_n^{(b)})^{-1/2}(\bar{X}_n^{(b)} - \bar{X}_n)$,

$$T_{n,k}^{(b)} = A_{l_k}(\delta_n^{(b)}), \quad k = 1, \ldots, K, \quad b = 1, \ldots, B,$$

and carry out steps (1.3)–(1.5). For any nondecreasing function $G_{n,k}(.)$, for $k = 1, \ldots, K$,

$$I\left[\bigcap_{k=1}^{K}\{G_{n,k}(T_{n,k}(\delta_n)) \le u\}\right] = I\left[\bigcap_{k=1}^{K}\{T_{n,k}(\delta_n) \le G_{n,k}^{-1}(u)\}\right]. \tag{2.2}$$

Thus, the consistency of this bootstrap-based fusion procedure is a direct consequence of the theorem below. Let $F_n(.)$ denote the joint distribution of $\{T_{n,k}, k = 1, \ldots, K\}$ under H_0, and $F_n^*(.|\mathbf{X}_1^n)$ denote their joint bootstrap distribution, namely, the joint distribution of $\{T_{n,k}, k = 1, \ldots, K\}$, calculated using the bootstrap samples derived from $\mathbf{X}_1^n$, as described above.

Theorem 1. *Suppose* Conditions (C1)–(C3) *in the Appendix hold. Then,*

$$\sup_{t_1,\ldots,t_k \in R} \left| F_n(t_1, \ldots, t_K) - \mathrm{P}\left[\bigcap_{k=1}^{K}\{T_{n,k}(\underline{Z}) \le t_k\}\right]\right| = o(1),$$

$$\sup_{t_1,\ldots,t_k \in R} \left| F_n^*(t_1, \ldots, t_K|\mathbf{X}_1^n) - \mathrm{P}\left[\bigcap_{k=1}^{K}\{T_{n,k}(\underline{Z}) \le t_k\}\right]\right| = o_p(1),$$

where $\underline{Z} \sim N(0, \Sigma^X)$ *denotes the multivariate normal distribution with mean zero and covariance matrix* Σ^X*, and* $T_{n,k}(\underline{Z})$ *is as defined in* (2.1)*, with* δ_n *replaced with* $\underline{Z}$*.*

Remark 1. Chernozhukov, Chetverikov and Kat (2019) discuss testing H_0 based on the supremum statistic, where its null distribution is also approximated using an empirical bootstrap, with the only difference being that the same sample $\hat{D}_n$, instead of its bootstrapped version, is used to standardize the bootstrapped sample mean, that is, $\delta_n^{(b)}$ is defined as $n^{1/2}(\hat{D}_n)^{-1/2}(\bar{X}_n^{(b)} - \bar{X}_n)$. The second identity in Theorem 1 about the bootstrap distribution still holds in this case; nevertheless, a simulation study indicates that doing so tends to incur larger type-I errors; see Kong et al. (2022).

3. Two-Sample Mean Comparison

Suppose p-dimensional random vectors $X_1, \ldots, X_m$ are independent copies of $X \sim P_1(.)$, with mean μ^X and variance Σ^X, and $Y_1, \ldots, Y_n$ are independent copies of $Y \overset{i.i.d.}{\sim} P_2(.)$, with mean μ^Y and variance Σ^Y. The null hypothesis of

interest is $H_0 : \mu^X = \mu^Y$, which is referred to as the two-sample location model in Kock and Preinerstorfer (2019). The procedure and the main results in this section are stated for equal sample sizes, that is, $m = n$. A brief discussion is given at the end of this section on how the method can be adapted to the samples of unequal sizes.

As in the one-sample case, nearly all existing statistics for testing H_0 are based on the sample-mean difference $\delta_n = \bar{X}_m - \bar{Y}_n$; see, for example, Xue and Yao (2020), Cai, Liu and Xia (2014), and Zhang, Guo and Cheng (2020). For any $k = 1, \ldots, K$, let $T_{n,k}(\delta_n)$ be as defined in (2.1), and reject H_0 if $T_{n,k}(\delta_n)$ is too large. For any of these tests to be consistent, valid approximations to its null distribution are essential. Xue and Yao (2020) use an empirical bootstrap to determine the critical values for the supremum statistic. A different option is to use the permutation method. Chung and Romano (2016) prove that for a multivariate two-sample mean comparison, certain statistics are proper, in the sense that its permutation distribution function converges (uniformly) to its null distribution. The permutation method is also popular in practice; see, for example, Nettleton, Recknor and Reecy (2008), Chang and Tian (2016), and Efron and Tibshirani (2007). Its theoretical properties are examined in Kong et al. (2022) for the problem of a high-dimensional two-sample mean comparison, and it is shown to outperform the bootstrap method by a significant margin.

In the present context, the permutation procedure for fusing the sequence of statistics $\{T_{n,k}(\delta_n), k = 1, \ldots, K\}$ goes as follows. Following the notation used in Chung and Romano (2016), write $N = 2n$ and the pooled-sample $Z^N = \{Z_1, \ldots, Z_N\}$, where $Z_i = X_i$, for $i = 1, \ldots, n$, and $Z_{n+j} = Y_j$, for $j = 1, \ldots, n$. Thus, $\bar{X}_n$ can be interpreted as the average of the first half of the sample, $\{Z_1, \ldots, Z_n\}$, and $\bar{Y}_n$ is the average of the second half of the sample, $\{Z_{n+1}, \ldots, Z_N\}$.

Let G_N denote the set of all permutations of $\{1, \ldots, N\}$. For any $\pi = (\pi(1), \ldots, \pi(N)) \in G_N$, let Z^N_π denote the rearranged Z^N through permutation π, and $Z^N_{\pi(i)}$, for $i = 1, \ldots, N$, be the ith entry of Z^N_π. Recompute $\bar{X}_n$ and $\bar{Y}_n$ for Z^N_π, and denote the difference between them as $\delta_n(Z^N_\pi)$. Note that we use the notation $\delta_n(Z^N_\pi)$ to highlight its dependence on the permutated sample Z^N_π, whereas the simple δ_n is reserved for the sample mean difference calculated for the original (unpermutated) sample. For any $k = 1, \ldots, K$, let $T_{n,k}(Z^N_\pi)$ denote the value of $T_{n,k}(.)$, as in (2.1), when evaluated for $\delta_n(Z^N_\pi)$; its marginal (permutation) distribution of $T_{n,k}(Z^N_\pi)$ conditional on Z^N is thus

$$\hat{F}_{n,k}(t|Z^N) = \frac{1}{N!} \sum_{\pi \in G_N} I\Big\{T_{n,k}(Z^N_\pi) \le t\Big\}, \quad t \in R. \tag{3.1}$$

In this case, $\hat{U}_n$ of (1.3) is given by $\hat{U}_n = \max_{k=1,\ldots,K} \hat{F}_{n,k}(T_{n,k}(\delta_n)|Z^N)$. We

reject H_0 if

$$\frac{1}{N!}\sum_{\pi\in G_N} I\Big\{\max_{k=1,\ldots,K}\hat{F}_{n,k}(T_{n,k}(Z_\pi^N)) < \hat{U}_n\Big\} > 1-\alpha. \tag{3.2}$$

As a result of (2.2), the consistency of the above procedure (3.1) and (3.2) is a direct consequence of the next theorem. Let $F_n(.)$ denote the joint distribution of $\{T_{n,k}(\delta_n), k=1,\ldots,K\}$ under H_0, and $F_n^*(.|Z^N)$ denote their joint permutation distribution, that is,

$$F_n^*(t_1,\ldots,t_K|Z^N) := \frac{1}{N!}\sum_{\pi\in G_N} I\left[\bigcap_{k=1}^{K}\{T_{n,k}(Z_\pi^N)\le t_k\}\right], \quad t_1,\ldots,t_K\in R,$$

the joint distribution of $\{T_{n,k}(Z_\pi^N), k=1,\ldots,K\}$ calculated for the randomized sample derived from Z^N(via permutation π uniformly distributed on G_N).

Theorem 2. *Suppose Conditions* (C1)–(C3) *in the Appendix hold and that the same set of conditions also hold when* (Y,Σ^Y) *replaces* (X,Σ^X). *Then,*

$$\sup_{t_1,\ldots,t_K\in R}\Big|F_n^*(t_1,\ldots,t_K|Z^N) - F_n(t_1,\ldots,t_K)\Big| \to 0, \text{ in probability.} \tag{3.3}$$

Remark 2. Similarly to Section 2, we can also consider cases where the test statistics $\{T_{n,k}(\delta_n), k=1,\ldots,K\}$ are evaluated for marginal-standardized δ_n, that is, $\delta_n = n^{1/2}(\hat{D}_n)^{-1/2}(\bar{X}_n - \bar{Y}_n)$, where $\hat{D}_n = \mathrm{diag}(\hat{\Sigma}_n)$, the diagonal matrix consisting of the diagonal elements of

$$\hat{\Sigma}_n = \frac{1}{2n}\sum_{i=1}^{n}(X_i-\bar{X}_n)(X_i-\bar{X}_n)^\top + \frac{1}{2n}\sum_{i=1}^{n}(Y_i-\bar{Y}_n)(Y_i-\bar{Y}_n)^\top.$$

In this case, $\hat{D}_n$ is recomputed for each permutated sample, and Theorem 2 continues to hold if $\hat{D}_n$ is accurate enough, as per Assumption (A6) of Kong et al. (2022).

Remark 3. When the two samples are of unequal sizes ($m\neq n$), Kong et al. (2022) prove that the limit of the permutation distribution of the statistics, be it $T_{n,k}(\delta_n)$ or its marginally standardized version, does not coincide with their respective (null) distributions, unless $\Sigma^X = \Sigma^Y$. One solution is to apply the binning procedure in Kong et al. (2022) to obtain pseudo samples of equal sizes, and then proceed as before. If $m/(m+n) = c + O(N^{-1/2})$, for some $c\in(0,1)$, then similarly to Theorem 2, we can prove the consistency of the fusion procedure (3.1) and (3.2) based on these pseudo samples.

4. Test of Vector Independence

Let X and Y stand for random vectors of dimension p and q, respectively, with $\mathcal{D}_X$ and $\mathcal{D}_Y$ as their respective domains. Suppose we have n independent copies $\{(X_i, Y_i)\}_{i=1}^n$ of (X, Y), and we are interested in testing the null hypothesis H_0: X and Y are independent. Write $\mathbf{X}_1^n = \{X_1, \ldots, X_n\}$ and $\mathbf{Y}_1^n = \{Y_1, \ldots, Y_n\}$. In the univariate case, DiCiccio and Romano (2017) consider the test of H_0 based on the sample correlation $\rho_n(.)$, and prove that its null distribution can be approximated by random permutations of $\mathbf{Y}_1^n$ or $\mathbf{X}_1^n$.

Compared with $\rho_n(.)$, the HHG statistic of Heller, Heller and Gorfine (2013) is able to identify nonlinear association. The notion behind it is simple: suppose $d_X(.)$ and $d_Y(.)$ are two distance metrics, such as the Euclidean distance; if H_0 is false, then there must exist two distinct points $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$, so that the two binary random variables $I\{d_X(X, \mathbf{x}_1) \le d_X(\mathbf{x}_1, \mathbf{x}_2)\}$ and $I\{d_Y(Y, \mathbf{y}_1) \le d_Y(\mathbf{y}_1, \mathbf{y}_2)\}$ are correlated. The HHG statistic is then based on the Pearson's correlation for the corresponding 2×2 contingency table:

$$T_n(\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2; d_X(.), d_Y(.)) = n^{1/2} \frac{A_{1,1} - A_{1.}A_{.1}}{(A_{1.}A_{.1})^{1/2}}, \tag{4.1}$$

where

$$\begin{aligned} A_{1,1} &:= A_{1,1}(\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2; d_X(.), d_Y(.)) \\ &= \frac{1}{n} \sum_{i=1}^n I\big\{d_X(X_i, \mathbf{x}_1) \le d_X(\mathbf{x}_1, \mathbf{x}_2)\big\} I\big\{d_Y(Y_i, \mathbf{y}_1) \le d_Y(\mathbf{y}_1, \mathbf{y}_2)\big\}, \\ A_{1.} &:= A_{1.}(\mathbf{x}_1, \mathbf{x}_2; d_X(.)) = \frac{1}{n} \sum_{i=1}^n I\big\{d_X(X_i, \mathbf{x}_1) \le d_X(\mathbf{x}_1, \mathbf{x}_2)\big\}, \\ A_{.1} &= A_{.1}(\mathbf{y}_1, \mathbf{y}_2; d_Y(.)) = \frac{1}{n} \sum_{i=1}^n I\big\{d_Y(Y_i, \mathbf{y}_1) \le d_Y(\mathbf{y}_1, \mathbf{y}_2)\big\}. \end{aligned} \tag{4.2}$$

In Heller, Heller and Gorfine (2013), the null distribution of the statistic (4.1) is approximated by random permutations of $\mathbf{Y}_1^n$. This practice is intuitively correct, but no theoretical justification has been provided yet. Because $A_{1.}$ and $A_{.1}$, the two marginal terms in (4.1), are both invariant to permutations (of $\mathbf{Y}_1^n$), it is the numerator, $A_{1,1} - A_{1.}A_{.1}$, that determines the permutation distribution of (4.1). Thus, henceforth, we do not discriminate between (4.1) and its numerator. Variations of (4.1), while retaining its contingency-table-derived form, can be constructed by altering choices for the following two factors:

(i) **values specified for $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$.** Apparently, the statistic (4.1) associated with any specific values of $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ is more sensitive to dependency that occurs close to the specified locations. Violations of H_0 in locations further away might not be strong enough to yield significant

changes. By combining statistics associated with varied choices of $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ scattered in $\mathcal{D}$, we can gather evidence (of dependence) from different locations.

(ii) **types of distance metrics for $d_X(\cdot)$ and $d_Y(\cdot)$.** This factor, as noted in Heller, Heller and Gorfine (2013), could be designed to capture the localized dependency between X and Y. For example, we could consider distance metrics $d_X(\cdot)$ that depend only on a certain sub-vector $X_{\mathcal{S}}$ of X, so that the resulting statistic is more powerful against alternatives when the association between (X, Y) is largely due to that between the sub-vector $X_{\mathcal{S}}$ and Y.

These variations of (4.1), notwithstanding belong to a general class of statistics of the following form:

$$n^{-1} \sum_{i=1}^{n} \{a(X_i) - \bar{a}_n\}\{d(Y_i) - \bar{d}_n\}, \tag{4.3}$$

where $a(\cdot)$ and $d(\cdot)$ are both square integrable functions, with $d(\cdot)$ being categorical (i.e., taking only a finite number of possible values), and $\bar{a}_n = n^{-1} \sum a(X_i)$ and $\bar{d}_n = n^{-1} \sum d(Y_i)$ are their respective sample averages. To see this is the case, set

$$a(X) = I\big\{d_X(X, \mathbf{x}_1) \le d_X(\mathbf{x}_1, \mathbf{x}_2)\big\}, \quad d(Y) = I\big\{d_Y(Y, \mathbf{y}_1) \le d_Y(\mathbf{y}_1, \mathbf{y}_2)\big\}.$$

Then, (4.3) reduces to the numerator in (4.1).

Without loss of gerality, suppose for $k = 1, \ldots, K$, $a_k(\cdot)$ and $d_k(\cdot)$ are functions satisfying the requirements above specified for (4.3). Write

$$T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^n) = n^{-1/2} \sum_{i=1}^{n} \{a_k(X_i) - \bar{a}_n^{(k)}\}\{d_k(Y_i) - \bar{d}_n^{(k)}\}, \quad k = 1, \ldots, K, \tag{4.4}$$

where, $\bar{a}_n^{(k)}$ and $\bar{d}_n^{(k)}$, for $k = 1, \ldots, K$, are the sample averages of $a_k(X_i)$ and $d_k(Y_i)$, respectively. In the language of Hajek, Sidak and Sen (1999), $a_k(X_i)$ is referred to as the coefficient, and $d_k(Y_i)$ are the scores. We focus on the combination of statistics of this general form using the fusion procedure, where the resampling is done via random permutations of $\mathbf{Y}_1^n$.

For any $\pi \in G_n$, let $\{\pi(1), \ldots, \pi(n)\}$ denote the rearranged $\{1, \ldots, n\}$ through permutation π, and $\mathbf{Y}_1^{n,\pi} = \{Y_{\pi(1)}, \ldots, Y_{\pi(n)}\}$. For $k = 1, \ldots, K$, evaluate $T_{n,k}(\cdot)$ for the permuted sample as

$$T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^{n,\pi}) = n^{-1/2} \sum_{i=1}^{n} \{a_k(X_i) - \bar{a}_n^{(k)}\}\{d_k(Y_{\pi(i)}) - \bar{d}_n^{(k)}\}, \tag{4.5}$$

with their marginal and joint permutation distributions given by

$$\hat{F}_{n,k}(t|\mathbf{X}_1^n, \mathbf{Y}_1^n) = \frac{1}{n!}\sum_{\pi\in G_n} I\Big\{T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^{\pi(n)}) \le t\Big\}, \quad t \in R, \tag{4.6}$$

$$\hat{F}_n(t_1, \ldots, t_K|\mathbf{X}_1^n, \mathbf{Y}_1^n) := \frac{1}{n!}\sum_{\pi\in G_n} I\left[\bigcap_{k=1}^{K}\{T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^{\pi(n)}) \le t_k\}\right], \tag{4.7}$$

respectively. Let $\hat{U}_n$ be as defined in (1.3), with $\hat{F}_{n,k}(.|\mathbf{X}_1^n, \mathbf{Y}_1^n)$ replacing $\hat{F}_{n,k}(\cdot)$, for $k = 1, \ldots, K$. Similarly to (3.2), we reject H_0 if $\hat{R}_n^U(\hat{U}_n|\mathbf{X}_1^n, \mathbf{Y}_1^n) \ge 1 - \alpha$, where

$$R_n^{\hat{U}}(u|\mathbf{X}_1^n, \mathbf{Y}_1^n) := \frac{1}{n!}\sum_{\pi\in G_n} I\left[\bigcap_{k=1}^{K}\{F_{n,k}(T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^{\pi(n)})|\mathbf{X}_1^n, \mathbf{Y}_1^n) \le u\}\right]. \tag{4.8}$$

Let $F_n(\cdot)$ denote the joint distribution of $\{T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^n), k = 1, \ldots, K\}$ under H_0.

Theorem 3. *Under H_0, with probability one,*

$$\sup_{t_1,\ldots,t_K\in R} \big|\hat{F}_n(t_1, \ldots, t_K|\mathbf{X}_1^n, \mathbf{Y}_1^n) - F_n(t_1, \ldots, t_K)\big| = o(1). \tag{4.9}$$

Based on Theorem 3, the consistency of the fusion procedure (4.6)–(4.8) is a straightforward result.

Corollary 1. *Under H_0, with probability one,*

$$\sup_{u\in(0,1)} |\hat{R}_n^U(u|\mathbf{X}_1^n, \mathbf{Y}_1^n) - \mathrm{P}(\hat{U}_n \le u)| = o(1).$$

Remark 4. Based on Theorem 3 and the continuous mapping theorem, it is straightforward to see that the fusion procedure (4.6)–(4.8) is also consistent if the fusion statistic U_n of (1.2) is replaced with any continuous function of $\{T_{n,k}(\cdot), k = 1, \ldots, K\}$. For example, suppose $\{(\mathbf{x}_k, \mathbf{y}_k) : k = 1, \ldots, K\}$ is a collection of (fixed) grid points in $\mathcal{D}$. We could then consider the summation, or the maximum, of the squared (4.1) taken over these grid points; that is,

$$\tilde{U}_n = \sum_{k,l=1}^{K} T_n^2(\mathbf{x}_k, \mathbf{y}_k, \mathbf{x}_l, \mathbf{y}_l; d_X(\cdot), d_Y(\cdot)), \tag{4.10}$$

$$\tilde{U}_n = \max_{k,l} T_n^2(\mathbf{x}_k, \mathbf{y}_k, \mathbf{x}_l, \mathbf{y}_l; d_X(\cdot), d_Y(\cdot)). \tag{4.11}$$

Note that (4.10) is the Cramér–von-Mises-type of statistic studied in Heller, Heller and Gorfine (2013); Heller et al. (2016). Thus, as a byproduct, Theorem 3 also provides theoretical justifications for the practice in Heller, Heller and Gorfine (2013); Heller et al. (2016) of approximating the null distributions of these aggregations numerically by using their permutation distributions.

For the same reason, the consistency of the fusion procedure (4.6)–(4.8) also holds for the Kolmogorov–Smirnov-type statistic (4.11), or when $T_n(\cdot)$ in (4.10)

or (4.11) is replaced by the G likelihood-ratio,

$$A_{1,1}\log\left(\frac{A_{1,1}}{A_{1.}A_{.1}}\right)+A_{1,2}\log\left(\frac{A_{1,2}}{A_{1.}A_{.2}}\right)+A_{2,1}\log\left(\frac{A_{2,1}}{A_{2.}A_{.1}}\right)+A_{2,2}\log\left(\frac{A_{2,2}}{A_{2.}A_{.2}}\right), \tag{4.12}$$

where $A_{i,j}, A_{i,.}, A_{.,j}$, for $i,j=1,2$, are as given in (4.2). These four fused statistics can go through one more round of the fusion procedure, and the resulting test procedure would still be consistent.

Remark 5. For the proof of Theorem 3, the permutation distribution is derived based on the notion that when π is uniformly distributed on G_n, $\pi(i)$ can be interpreted as the rank of U_i, for $i=1,\ldots,n$, where $U_1,\ldots,U_n$ are i.i.d. $U(0,1)$. In this sense, $T_{n,k}(\mathbf{X}_1^n,\mathbf{Y}_1^{n,\pi})$ of (4.5) falls into the category of simple linear rank statistics (Hajek, Sidak and Sen, 1999). The theoretical tools currently available are enough to derive the limiting distribution of individual rank statistics, but not for their joint limiting distributions, as required in our case. It is for this extension to the multivariate case that we require the function $d_k(\cdot)$ to be categorical. Removing of such restrictions is left to future research.

5. Extensions

Engaging fusion statistics other than (1.2) is perfectly possible. Indeed, the results in Theorems 1–3 continue to hold if $F_{n,k}(\cdot)$ in the definition of (1.3) is replaced with any monotone function.

As observed in Sections 2 to 4, the consistency of the fusion procedure (1.3)–(1.5), depends on both the sequence of the test statistics $\{T_{n,k}, k=1,\ldots,K\}$ to be fused and the fusion statistic, U_n, itself. For the fusion statistic (1.2), the fusion procedure is consistent as long as the joint bootstrap (or permutation) distribution function of $\{T_{n,k}, k=1,\ldots,K\}$ is a valid approximation of their joint null distribution. Were we to consider a sequence of test statistics other than those studied here, then the consistency of the fusion procedure needs to be re-evaluated, because the bootstrap (or permutation) distribution is not necessarily always a valid approximation of the null, even in the non-high-dimensional (fixed-dimensional) setting; see, for example, Chung and Romano (2013, 2016).

Having said that, certain variations (or extensions) of the proposed procedure can be verified in a relatively straightforward manner. For example, $F_{n,k}(\cdot)$ in (1.2) or $\hat{F}_{n,k}(\cdot)$ in (1.3) can be replaced with an arbitrary monotone function, and the results in Theorem 1, Theorem 2, and Theorem 3 will continue to hold. Another possibility is to allow K, the number of statistics to be fused, to also increase with n. For example, in the two-sample mean comparison problem of Section 3, we do not known a priori the number of coordinates where μ^X and μ^Y differ from each other. Thus, $T_{n,k}(\delta_n)$ is calculated for as many k as possible, hoping that one of these k-values is close to the true count. Without loss of

generality, for $k = 1, \ldots, s_n (\leq p)$, define

$$T_{n,k} = A_k(\delta_n);$$

we can then repeat the fusion procedure (3.1) and (3.2) with K replaced by s_n. The proof of the consistency of the procedure is similar to when K is fixed, if the rate at which $s_n \to \infty$ is slow enough. Specifically, if s_n is allowed to be as large as p, then p is at most of order $o(n^{1/7})$, rather than the exponential rate implied by Condition (C3) in the Appendix.

we cannot make general recommendations for choosing between different fusion statistics, because the existence of an optimal fusion statistic is, to the best of our knowledge, still an open question. For the sequence of test statistics of (2.1), a general form of the type of fusion statistic for which the consistency of the corresponding fusion procedure still holds is

$$U_n = F(f_k(T_{n,k}), k = 1, \ldots, K), \tag{5.1}$$

where $F(\cdot) : R^K \to R$ and $f_k(\cdot) : R \to R$, for $k = 1, \ldots, K$. For the overall function to be convex, it is sufficient that either

- $f_k(\cdot)$ are all convex; $F(\cdot)$ is convex and nondecreasing in each argument, or
- $f_k(\cdot)$ are all concave; $F(\cdot)$ is convex and nonincreasing in each argument.

As a result, we have for any $u \in R$, there exists some s_n-sparsely convex set $A \subset \mathbb{R}^p$ (Definition 3.1 of Chernozhukov, Chetverikov and Kato (2017)), such that

$$I(U_n \leq u) = I(\delta_n \in A).$$

Write $\underline{T}_n = (T_{n,k}(\delta_n), k = 1, \ldots, K)$, $\underline{T}_n^{(b)} = (T_{n,k}(\delta_n^{(b)}), k = 1, \ldots, K)$. Under certain regularity conditions, we can apply Proposition 3.2 of Chernozhukov, Chetverikov and Kato (2017) and prove, similarly to Theorem 1, that

$$\sup_{A \in \mathcal{A}^{sp}(s_n)} \left| \mathrm{P}_n^*(\underline{T}_n^{(b)} \in A | \mathbf{X}_1^n) - \mathrm{P}(\underline{T}_n \in A) \right| = o_p(1),$$

where $\mathcal{A}^{sp}(s_n)$ denotes the class of all s_n sparsely convex sets in $\mathbb{R}^p$, and $\mathrm{P}_n^*(\cdot|\cdot)$ denotes the bootstrap distribution conditional on $\mathbf{X}_1^n$. An analogue of Theorem 2, and consequently the consistency of the corresponding fusion procedure, can then be proved similarly.

Asymptotically the empirical version of the aforementioned Fisher's combined p-value of (1.1) can be written in the form of (5.1). To see this, first note that, based on Theorem 1, we have for any given k, $T_{n,k}(\cdot)$ converges in distribution to $T_{n,k}(\underline{Z})$, which, as shown in the proof of Theorem 1, is equivalent to the maximum of a Gaussian vector of dimension $2^k \binom{p}{k}$. Second, Theorem 1 of Cai, Liu and Xia (2014) states that under certain regularity conditions,

the maximum of a p-dim Gaussian vector with unit variances has its limiting distribution as the type-I extreme value distribution, that is,

$$\exp\Big(-\pi^{-1/2}\exp\Big\{-\frac{t^2-2\log p+\log\log p}{2}\Big\}\Big).$$

Finally, it is easy to check that

$$-\log\Big(1-\exp\Big(-\pi^{-1/2}\exp\Big\{-\frac{t^2-2\log p+\log\log p}{2}\Big\}\Big)\Big)$$

is indeed convex in t.

6. Simulation Studies

In this section, we examine the performance of the fusion procedure (1.3)–(1.5), denoted by $fused$, when it is applied to the three testing problems discussed in Sections 2–4.

We use the following notation: I_p, the $p \times p$ identity matrix; $N_p(\mu, \Sigma)$, the p-dim normal with mean μ and covariance matrix Σ; and $T_p(k, \mu, \Sigma)$, the p-dim t-distribution with k degrees of freedom, mean μ, and covariance Σ. The significance level α is fixed as 5%. Empirical sizes are calculated based on 5,000 repetitions, and the empirical powers are based on 1,000 repetitions. Within each repetition, the bootstrap (or permutation) distributions are calculated based on $B(= 10000)$ resampling via bootstrap (or permutation).

6.1. One-sample mean test

In testing $H_0 : \mu = 0$, the performance of the fusion procedure is compared with that of the individual test statistics, namely, $T_{n,1}$ and $T_{n,p}$ of (2.1). Also included in the comparison are the statistics of Bai and Saranadasa (1996), denoted by T_{BS}, and Srivastava and Du (2008), denoted by T_{SD}. These two summation-type statistics were originally proposed for a two-sample mean comparison, and are now adopted for the current purpose. Their p-values should be decided based on their asymptotic distributions, but because these tend to be over-inflated, we use the bootstrap.

The sample size is fixed at 100. We consider two designs for X: $N_p := N_p(\mu, \Sigma)$ and $T_p := T_p(5, \mu, \Sigma)$ with $\Sigma = D^{1/2}RD^{1/2}$, where D and R are generated as follows:

Σ_1: $D = I_p$, $R = (\rho_{i,j})$, where $\rho_{i,i} = 1$, and $\rho_{i,j} = 0.25$, if $i \neq j$.

Σ_2: $D = \mathrm{diag}(\sigma^2)$, where $\sigma_j, j = 1, \ldots, p, \overset{i.i.d.}{\sim} U(2, 3)$; $R = (\rho_{i,j})$ with $\rho_{i,j} = 0.25^{|i-j|}$.

Σ_3: D is the same as in Σ_2; R is the same as in Σ_1.

Table 1. Empirical sizes(%) of different tests.

$(Dist., \Sigma)$	p	$T_{n,1}$	$T_{n,p}$	$fused$	T_{BS}	T_{SD}
(N_p, Σ_1)	100	4.88	4.88	5.08	5.12	5.60
	200	5.20	4.70	5.14	4.92	5.46
	500	5.40	5.50	5.24	5.50	6.06
	1,000	4.98	4.90	4.86	5.06	5.60
(T_p, Σ_1)	100	3.48	4.16	3.66	4.22	4.78
	200	3.36	4.26	3.66	4.34	4.76
	500	3.36	4.32	3.90	4.38	5.00
	1,000	3.08	4.30	3.58	4.50	4.92
(N_p, Σ_2)	100	4.88	4.88	4.96	5.30	5.30
	200	5.34	5.34	5.44	5.76	5.76
	500	5.44	5.44	5.56	6.06	6.06
	1,000	4.84	4.84	4.88	5.14	5.14
(N_p, Σ_3)	100	5.22	4.94	5.32	5.16	5.58
	200	5.28	4.88	5.12	5.02	5.64
	500	5.62	5.44	5.54	5.50	6.02
	1,000	5.28	4.68	4.88	4.86	5.42

The results for the empirical sizes, for different combinations of distributions, Σ and dimension p, are given in Table 1. The size of $fused$ is fairly close to the nominal size in nearly all settings, and is relatively more stable than its competitors.

Examples of alternatives are generated by specifying nonzero values for some entries of μ in the above examples. Specifically, for $d = 0.1, 0.5, 0.9$, $\lfloor dp \rfloor$ components of μ are randomly selected and are independently assigned values drawn from $U(-s, s)$, for some $s > 0$, and the other entries of μ remain zero. Here, d controls the sparsity of the signal, and s determines the signal strength. Table 2 reports the empirical power for $p = 1000$, different combinations of distributions $(Dist)$, Σ and (d, s), for the four competing methods. What is immediately obvious is that the $fused$ statistics enjoy universally higher power than when using $T_{n,1}$ or $T_{n,p}$ alone. It also significantly outperforms both T_{BS} and T_{SD}.

6.2. Two-sample mean comparison

In this section, in addition to $T_{n,1}$ and $T_{n,p}$, we compare the proposed fusion procedure with the statistics considered in Xu et al. (2016) and Chen, Li and Zhon (2019), referred to as T_{XLPW} and T_{CLZ}, respectively. The statistics studied in Aoshima and Yata (2018), Chen and Qin (2010), and Zhang, Guo and Cheng (2020) are similar to T_{CLZ} in definition, require similar assumptions, and show similar performance, and thus are excluded from the comparison. Simulation examples are taken from Xu et al. (2016), where the two p-dim random vectors

Table 2. Empirical powers(%) of different tests.

$(Dist., \Sigma)$	(d, s)	$T_{n,1}$	$T_{n,p}$	$fused$	T_{BS}	T_{SD}
(N_p, Σ_1)	(0.1, 0.31)	89.7	9.9	92.1	10.2	11.7
	(0.5, 0.22)	82.7	42.7	92.9	45.1	53.0
	(0.9, 0.19)	77.2	85.2	94.3	87.3	91.0
(T_p, Σ_1)	(0.1, 0.41)	89.2	10.1	92.7	9.9	11.8
	(0.5, 0.28)	78.0	43.4	87.8	39.8	49.4
	(0.9, 0.25)	76.2	84.7	91.4	81.7	88.3
(N_p, Σ_2)	(0.1, 7.20)	92.8	4.4	93.6	5.0	5.0
	(0.5, 6.50)	94.8	5.4	94.9	5.9	6.2
	(0.9, 6.20)	93.5	7.0	93.5	7.1	7.3
(N_p, Σ_3)	(0.1, 0.75)	90.0	9.5	92.1	9.7	11.0
	(0.5, 0.55)	87.8	50.1	96.8	45.6	59.4
	(0.9, 0.45)	77.3	75.2	91.6	70.5	83.0

X and Y are generated according to

$$X = (\xi^1, \xi^2), \quad Y = (\eta^1, \eta^2) + \mu^Y, \quad \mu^Y = (\mu_1^Y, 0);$$

here, ξ^1, η^1 are both of length $p/2$, both with entries being independent $U(-1, 1)$, and ξ^2 and η^2 are independent $T_{p/2}(3, 0, \Sigma)$, with $\Sigma = (0.6^{|i-j|})$. When evaluating empirical sizes, $\mu_1^Y = 0$; for the empirical power comparison, with any given $\beta \in (0, 1)$ and $s \in (0, 1)$, $p_0 = \min(\lfloor p^\beta \rfloor, p/2)$ components of μ_1^Y are randomly selected and set to equal s, so that s is an indicator of the signal strength, and the sparsity of the signals is controlled by β.

With $n = 100$, Table 3 shows the results for the empirical sizes of the various methods for different p. The two columns labelled T^a_{XLPW} and T^a_{CLZ} are the empirical sizes of T_{XLPW} and T_{CLZ}, respectively, when the critical value is obtained based on the theoretical asymptotic null distributions, with plugged-in parameter estimates, as given in Xu et al. (2016) and Chen, Li and Zhon (2019), respectively. Obviously, empirical sizes obtained in this manner are unduly high, but if the critical values are approximated using permutations, then the results for T_{XLPW} and T_{CLZ} and the other three statistics are all fairly close to the nominal 5%. Note that the computation times required by the first three methods are much shorter than those of T_{XLPW} and T_{CLZ}, especially when p gets larger.

With $p = 500$ and $\beta \in \{0.9, 0.8, 0.6, 0.5, 0.4, 0.2\}$, the empirical power of each method versus the signal strength s is as depicted in Figure 1. For T_{XLPW} and T_{CLZ}, because of their aforementioned unduly high type-I errors induced by the asymptotic distributions, we only report their power when the critical value is obtained using the permutation distribution. The general pattern is that as the degree of sparsity increases, the best method switches from $T_{n,p}$ to $T_{n,1}$. This is in line with the observation we made at the beginning of Section 2. In comparison,

Table 3. Size(%) (computation time in seconds) of different tests.

p	$T_{n,1}$	$T_{n,p}$	$fused$	T_{XLPW}	T_{CLZ}	T^a_{XLPW}	T^a_{CLZ}
100	5.26	5.26	5.78	5.22	5.32	7.40	7.02
	(0.04)	(0.03)	(0.29)	(1.57)	(1.91)	(0.66)	(0.001)
200	5.28	4.96	5.18	5.22	4.72	7.07	6.94
	(0.05)	(0.05)	(0.43)	(5.33)	(6.36)	(2.56)	(0.01)
500	4.70	5.06	4.92	4.50	4.66	9.93	7.65
	(0.10)	(0.09)	(0.93)	(31.70)	(39.68)	(15.55)	(0.04)
1,000	5.16	4.78	5.18	5.24	4.62	14.81	7.42
	(0.19)	(0.16)	(1.79)	(127.73)	(164.40)	(91.34)	(0.24)

$fused$ is always among the top two best methods, regardless of the sparsity of the signals.

6.3. Independence test of random vectors

The code developed by Heller, Heller and Gorfine (2013) calculates four HHG-type statistics: *hhg.sc* of (4.10), *hhg.mc* of (4.11), *hhg.sl*, and *hhg.ml*. The first two are defined as in (4.10) and (4.11), respectively, and the last two are also defined according to (4.10) and (4.11), but with T_n^2 replaced with the G likelihood-ratio of (4.12). Also included in the comparison is $fused$ of these four statistics, the corresponding testing procedure based on which, as noted in Section 4, continues to be consistent. Among the existing tests of independence, we select the two popular methods, namely, the Hilbert–Schmidt independence criterion (HSIC) of Pfister et al. (2018) and the distance correlation (DC) of Huo and Székely (2016), for comparison.

Observations of X and Y are generated according to the following models, some taken from Zeng, Xia and Tong (2018). M1–M4 are univariate, and M5 and M6 are multidimensional.

M0 (Independent). $X \sim N_p(0, I)$ and $Y \sim N_p(0, I)$ are independent.

M1 (Linear with additive noise). $Y = X + 2.6\epsilon$, where $X, \epsilon \overset{i.i.d}{\sim} N(0, 1)$.

M2 (Circle with additive noise). $X = \sin(2\pi\theta) + 0.35\epsilon$, $Y = \cos(2\pi\theta) + 0.35\varepsilon$, where $\epsilon, \varepsilon \overset{i.i.d}{\sim} N(0, 1)$, $\theta \sim U(0, 1)$.

M3 (Quadratic with additive noise). $Y = (X - 0.5)^2 + 0.76\epsilon$, $X, \epsilon \overset{i.i.d}{\sim} U(0, 1)$.

M4 (Cloud with contaminated noise). $(X, Y) = Z \times \{0.2\mu + 0.2(\epsilon_1, \varepsilon_1 + 0.5)\} + (1 - Z)(\epsilon_2, \varepsilon_2)$, where $Z = 1$ or 0 with probability 0.82 and 0.18, respectively, μ is evenly selected from $\{\mu_1 = (0, 0), \mu_2 = (2, 0), \mu_3 = (4, 0), \mu_4 = (1, 1), \mu_5 = (3, 1), \mu_6 = (0, 2), \mu_7 = (2, 2), \mu_8 = (4, 2), \mu_9 = (1, 3), \mu_{10} = (3, 3)\}$, $\epsilon_1, \epsilon_2, \varepsilon_1, \varepsilon_2 \overset{i.i.d}{\sim} U(0, 1)$, and are independent of μ and Z.

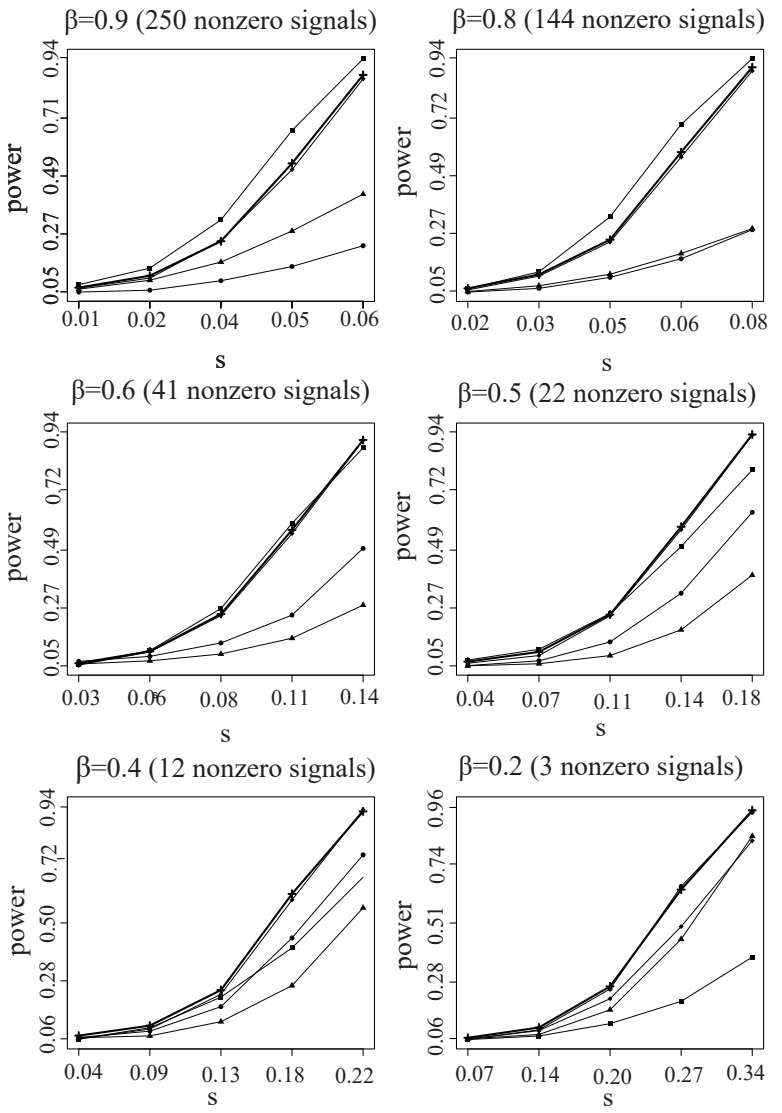

Figure 1. Power against signal strength s with different sparsity d and dimension $p = 500$: —•— for $T_{n,1}$, —■— for $T_{n,p}$, —+— for $fused$, —▲— for T_{XLPW}, —♦— for T_{CLZ}.

M5 (Multivariate conditional variance). $X = (X_1, \ldots, X_p)$ and $\phi = (\phi_1, \ldots, \phi_p)$ are independent $N_p(0, I_p)$; with $p_1 = \lfloor 0.7p \rfloor$, $Y_j = \phi_j(X_j + 0.6), j = 1, \ldots, p_1$, $Y_j = \phi_j, j = p_1, \ldots, p$.

M6 (Multivariate cloud with additive noise). $\phi = (\phi_1, \ldots, \phi_p)$ and $\psi = (\psi_1, \ldots, \psi_p)$ are independent $N_p(0, I_p)$. With $p_1 = \lfloor 0.8p \rfloor$ and μ as specified in (M4), $(X_j, Y_j) = \mu + 0.2(\phi_j, \psi_j)$, $j = 1, \ldots, p_1$, $(X_j, Y_j) = (\phi_j, \psi_j)$, $j = p_1 + 1, \ldots, p$.

Table 4. Empirical sizes (%).

(n, p)	①	②	③	④	①∼④	$HSIC$	DC
(100, 1)	5.08	4.48	4.92	4.68	4.78	5.06	5.54
(200, 1)	3.64	3.80	4.26	4.02	4.30	3.36	4.72
(100, 4)	5.32	4.92	3.72	4.58	4.10	5.80	5.42
(100, 12)	5.02	4.76	4.92	5.26	5.52	5.12	5.22
(100, 20)	3.58	3.32	5.12	5.72	4.24	5.14	4.92

①, . . . , ④ represent the statistics *hhg.sc*, *hhg.sl*, *hhg.mc*, and *hhg.ml*, respectively. ①∼④ represents the $fusion$ of statistic ① to statistic ④.

Table 5. Empirical powers (%).

(n, p)	M	①	②	③	④	①∼④	$HSIC$	DC
(100, 1)	1	70.7	69.5	38.2	36.0	62.2	94.6	65.9
	2	49.9	54.5	53.3	51.6	53.6	6.4	53.6
	3	51.7	49.6	43.1	51.6	52.4	24.7	33.3
	4	4.2	3.8	29.4	28.2	24.0	0.0	0.0
(200, 1)	1	96.6	96.3	73.1	75.0	94.3	100.0	94.4
	2	93.6	94.6	87.6	89.7	91.7	16.4	93.8
	3	90.8	90.5	88.4	95.7	94.7	62.6	72.9
	4	60.6	59.4	93.7	95.9	93.8	0.6	0.6
(100, 4)	5	89.7	87.8	53.1	46.2	85.7	48.4	25.4
	6	52.2	41.3	88.1	87.9	85.2	0.0	0.0
(100, 12)	5	92.1	90.7	41.0	42.8	87.8	29.4	20.1
	6	18.3	11.0	60.6	87.6	74.7	0.0	0.0
(100, 20)	5	91.9	90.8	40.5	41.8	86.5	24.5	20.5
	6	20.4	10.0	63.7	96.3	91.5	0.0	0.0

①, . . . , ④ represent the statistics *hhg.sc*, *hhg.sl*, *hhg.mc*, *hhg.ml*, respectively. ①∼④ stands for the $fusion$ of statistics ① to statistic ④.

For Model M0, where X and Y are independent, Table 4 contains the empirical sizes of all test statistics. All methods maintain reasonable control over the type-I error. For Models M1–M6, their power is given in Table 5. In the univariate case, the $fused$ statistic based on ①∼④ consistently delivers high power across all four models, whereas each of its six competitors has strengths and weakness; for example, both HSIC and DC are powerless in detecting the dependency in M4. As for the multivariate case, HSIC and DC become unreliable for M6. As for the four HHG-type statistics, the two maximum-type statistics, ③ and ④, perform better with M6 than with M5, and vice versa for the two summation-type statistics ① and ②. Again, our $fused$ of the four HHG statistics, that is, ①∼④, maintains satisfactory power for both models, supporting our claim that when testing against an unknown alternative, the $fused$ statistic is, in general, a better choice than any individual statistic.

7. Real-Data Examples

Genome-wide association studies (GWAS) identify risk genetic variants for major human diseases by genotyping millions of single nucleotide polymorphisms (SNPs) in large cohorts. With data collected by the Wellcome Trust Case Control Consortium (WTCCC), we apply the two-sample mean comparison procedures of Section 3 to analyze the association between the SNPs and two diseases: type-2 diabetes (T2D), and rheumatoid arthritis (RA). In the case of T2D, there are 1,952 observations with 307,089 SNPs, and for RA, there are 1,969 observations with 305,394 SNPs. For either disease, the data are split into two groups: individuals with the disease (X), and individuals without the disease (Y). If the means of these two groups are different, then this indicates an association between the said disease and the SNPs. The p-values of the existing methods mentioned in Section 3 are all highly significant, suggesting an overwhelmingly strong association that it could be picked up by any valid tests, regardless of whether the test is sparse or dense sensitive.

In order for the data to be suitable for assessing the competitiveness of different tests, we need to first reduce the strength of the signals by thinning out the SNPs. This is realized through the following steps. First, calculate the p-value of each SNP, as in the case of a univariate mean-comparison problem, and rank the SNPs according to their p-values in ascending order. The now ordered SNPs are then divided into 1,000 roughly equal-sized groups, with about 300 SNPs in each group. Randomly select one SNP from each group to obtain a total of 1,000 SNPs. Finally, calculate the p-values of all competing methods using data on these 1,000 randomly selected SNPs. Repeat this procedure 200 times. The boxplot of the 200 p-values for each method is depicted in Figure 2, with the left panel occupied by those related to T2D, and the right panel by RA.

The first thing revealed by these plots is that the pattern related to the power of the various methods is largely in line with what we have seen in the simulation studies. These plots also highlight the potential use of the fusion statistic to choose between recommendations made by different testing methods. Specifically, for example, in the case of T2D, with a significance level set anywhere between 5% and 1%, the two statistics $T_{n,1}$ and $T_{n,p}$ make opposite recommendations for a majority of of the 200 occasions, with $T_{n,p}$ recommending rejection, and $T_{n,1}$ suggesting otherwise. In these occasions, the $fused$ statistic can then be used to decide which recommendation is more likely to be correct.

Appendix: Assumptions and Proofs

A.1. Further notations and regularity conditions for Sections 2 and 3

For any $\nu = (v_1, \ldots, v_p) \in R^p$, let $|\nu|_\infty$ denote the supremum norm and $|\nu|_1 = (|v_1| + \cdots + |v_p|)/p$, the L_1 norm.

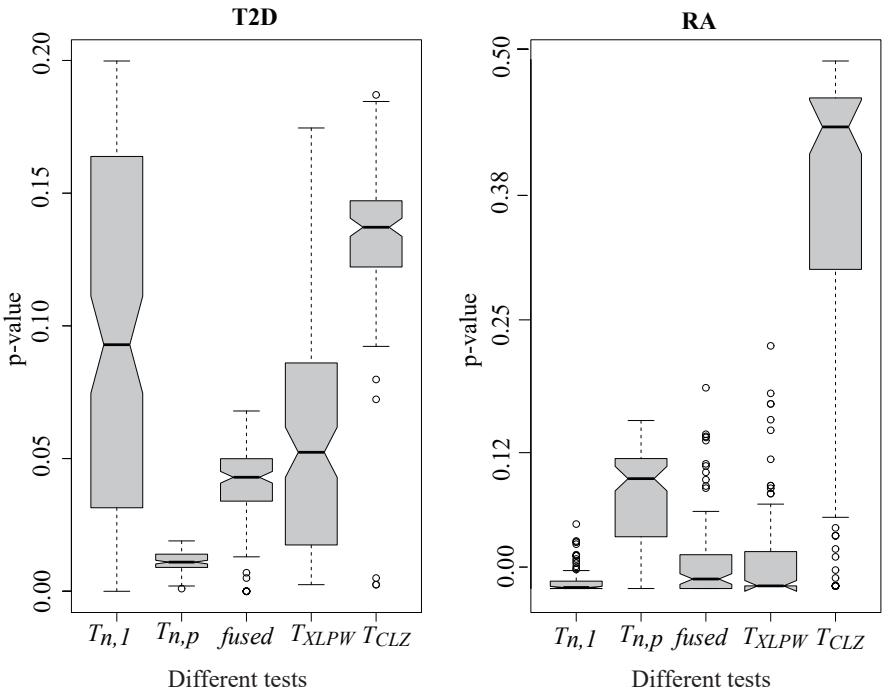

Figure 2. Boxplots of p-values for different statistics; the horizontal gray dashed line is the significance level.

For ease of exposition, suppose there exists some sequence of constants $B_n \geq 1$, such that $|X|_\infty \leq B_n$, $i = 1, \ldots, n$, $k = 1, \ldots, p$, with probability one. Moreover, for any $p-$dim vector with at most s_n nonzero elements being either 1 or -1, standardize it so that it has unit L_1 norm; let $\mathcal{C}(p, s_n)$ denote the collection of all such $p-$dim vectors, obviously with a cardinality no more than $(2p)^{s_n}$.

(C1) The diagonal elements of Σ^X are bounded both from below and above. The minimum eigenvalue of Σ^X is bounded from below by some constant $c_3 > 0$.

(C2) There exist finite constants $c_{n,1} > 0$ such that for any $\nu \in \mathcal{C}(p, s_n)$,

$$E\left\{ \exp\left(\frac{|\nu^\top X|}{c_{n,1}} \right) \right\} \leq 2, \quad E\{|\nu^\top X|^{2+k}\} \leq c_{n,1}^k, \ k = 1, 2. \tag{A.1}$$

(C3) $s_n = o(p)$, and $B_n s_n \log p = o(n^{1/7})$.

Conditions (C1)–(C3) could be found in Chernozhukov, Chetverikov and Kato (2017) so that the high-dimensional central limit theorem holds for simple convex

sets; see also Kong et al. (2022). Among them, (C3) dictates how large s_n and P could get relative to n.

A.2. Proof of Theorem 1

Let $D = \mathrm{diag}(\Sigma) = (\sigma_j^2)_{j=1,\ldots,p}$, where $\sigma_j^2 = Var(X^j)$. For a random variable Z, and any $r > 0$, let $|Z|_r = \{E(|Z|^r)\}^{1/r}$, and its Orlicz norm be defined as

$$|Z|_\psi = \inf\left\{C > 0 : E\psi\left(\left|\frac{Z}{C}\right|\right) \le 1\right\}, \quad \text{where } \psi(t) = e^t - 1.$$

A useful inequality is that $|Z|_r \le r!|Z|_\psi$. Condition (C2) implies that $|X^j|_\psi \le c_{n1}$, for all $j = 1, \ldots, p$. Then by Lemma 2.2.2 of van der Vaart and Wellner (1996), $|\max_{j=1,\ldots,p} X^j|_\psi \le c_{n1} \log p$. Consequently, $|\max_{j=1,\ldots,p} X^j|_4 \le c_{n1} \log p$, and based on Lemma D.3 of Chernozhukov, Chetverikov and Kat (2019), we have for any $c \in (0, 1)$,

$$\mathrm{P}\left(\max_{j=1,\ldots,p}\left|\frac{\hat\sigma_{nj}}{\sigma_j} - 1\right| \ge n^{-(1-c)/2} c_{n1}^2 \log^3 p\right) \le n^{-c},$$

whence $a_n = \sup_j |\hat\sigma_{nj} - \sigma_j| = o_p\{(\log p)^{-1}\}$. Let $\tilde\delta_n = D^{-1/2}\bar X_n$. Then for any $\epsilon > 0$, and $k = 1, \ldots, K$,

$$\mathrm{P}(T_{n,k}(\delta_n) \le t) \le \mathrm{P}(T_{n,k}(\bar X_n) \le t + \epsilon) + \mathrm{P}\left(|\bar X_n|_\infty \ge \frac{\epsilon}{a_n}\right), \tag{A.2}$$

$$\mathrm{P}(T_{n,k}(\delta_n) \le t) \ge \mathrm{P}(T_{n,k}(\bar X_n) \le t - \epsilon) + \mathrm{P}\left(|\bar X_n|_\infty \ge \frac{\epsilon}{a_n}\right). \tag{A.3}$$

Note that $I(T_{n,k}(\bar X_n)) \le t) \Leftrightarrow I(n^{1/2}\bar X_n \in A)$, for some $m - generated$ set A, namely a set generated by the intersection of $m-$half spaces (Chernozhukov, Chetverikov and Kato, 2017), where the half spaces are defined via vectors belonging to $\mathcal{C}(p, s_n)$, whence $m \le (2p)^{s_n}$. To see this is case, note that for any $p-$dim vector μ, $T_{n,k}(\mu) \le t$ is equivalent to: for any $\nu \in \mathcal{C}(p, s_n)$, $\mu^\top \nu \le t$, i.e. the intersection of half-spaces defined via vectors in $\mathcal{C}(p, s_n)$.

The terms on the RHS of (A.2) and (A.3) concerning $T_{n,k}(\bar X_n)$ could then be dealt with through similar arguments used in proving Theorem 3 of Kong et al. (2022), mostly involving high-dimensional Gaussian approximation, i.e., Proposition 2.2 of Chernozhukov, Chetverikov and Kato (2017), followed by anti-concentration inequalities. The probability to the RHS of (A.2) or (A.3) concerning $|\bar X_n|_\infty$ is $o_p(1)$, and could be similarly proved by making use of the fact that $a_n = o_p\{(\log p)^{-1}\}$.

The bootstrapped sample has mean $\bar X_n$ and variance matrix $\hat\Sigma_n = n^{-1}\sum_i (X_i - \bar X_n)(X_i - \bar X_n)^\top$. Thus the convergence of the bootstrap distribution could be proved through similar arguments in conjunction with Proposition 4.3 of Chernozhukov, Chetverikov and Kato (2017).

A.3. Proof of Theorem 2

Similar to the arguments below (A.3), with $T_{n,k}(\cdot)$ as defined in (2.1), we have

$$I\left[\bigcap_{k=1}^{K}\{T_{n,k}(Z_\pi^N) \le t_k\}\right] = I\{n^{1/2}\delta_n(Z_\pi^N) \in A\},$$

where $A \subset R^p$ is some $m-$*generated* set with $m \le (2p)^{s_n}$. Through arguments similar to those used in proving Theorem 3 of Kong et al. (2022), we have

$$\sup_A |\mathrm{P}^*\{n^{1/2}\delta_n(Z_\pi^N) \in\} - \mathrm{P}\{N(0, \Sigma^X + \Sigma^Y) \in A\}| \to 0, \text{ in probability}$$
$$\sup_A |\mathrm{P}\{n^{1/2}\delta_n(Z^N) \in A\} - \mathrm{P}\{N(0, \Sigma^X + \Sigma^Y) \in A\}| \to 0,$$

where the supremum is taken over all $m-$*generated* set with $m \le (2p)^{s_n}$, while the probability P^* is taken conditional on Z^N, with respect to π uniformly distributed on G_N.

A.4. Proof of Theorem 3

Let $a = (a_k, k = 1, \ldots, K)^\top$, $d = (d_k, k = 1, \ldots, K)^\top$, where $a_k = E\{a_k(X)\}$, $d_k = E\{d_k(Y)\}$. $\bar{a}_n = (\bar{a}_n^{(k)}, k = 1, \ldots, K)^\top$; and $\bar{d}_n = (\bar{d}_n^{(k)}, k = 1, \ldots, K)^\top$. Let $(Z_1, \ldots, Z_k) \sim N(0, \Sigma)$, the $K-$dim Gaussian with covariance matrix $\Sigma = [\sigma_{k,l}s_{k,l}]$, where $\sigma_{k,l} = \mathrm{Cov}(a_k(X), a_l(X))$, $s_{k,l} = \mathrm{Cov}(d_k(Y), d_l(Y))$.

The proof of Theorem 3 is broken down into the following two lemmas, which deals with the (joint) null distribution and the joint permutation distribution, respectively.

Lemma 1.

$$\max_{k=1,\ldots,K} \sup_{t_k \in R} |F_{n,k}(t_k) - \mathrm{P}(Z_k \le t_k)| = o(1),$$
$$\sup_{t_1,\ldots,t_K \in R} \left| F_n(t_1, \ldots, t_{s_n}) - \mathrm{P}\left(\bigcap_{k=1}^{K}\{Z_k \le t_k\}\right)\right| = o(1),$$

Proof of Lemma 1. For fixed K, the assertion is simply the multivariate CLT. Here, we prove these two statements under a more general set-up where $K = s_n$, the number of statistics, is allowed to grow as n increases, while the function $a_k(X)$ and $d_k(Y), k = 1, \ldots, s_n$, could be any measurable functions satisfying the moment conditions (A1)-(A3) below.

(A1) $\inf_{k=1,\ldots,s_n} E[\{a_k(X)d_k(Y)\}^2] > 0$.

(A2) There exists some sequence of constants $B_n \ge 1$, possibly growing to infinity as $n \to \infty$, such that for all $i = 1, \ldots, n$, and $k = 1, \ldots, s_n$,

$$E\left[\exp\left\{\frac{|a_k(X)d_k(Y)|}{B_n}\right\}\right] \le 2, \quad E\{|a_k(X)d_k(Y)|^{2+l}\} \le B_n^l, \quad l = 1, 2. \quad \text{(A.4)}$$

(A3) $B_n^{1/3}\{\log(ns_n)\}^{7/6} = o(1)$.

For ease of exposition, write $T_{n,k} := T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^n)$, and

$$T_{n,k} = S_{n,k} - n^{1/2}(\bar{a}_n^{(k)} - a_k)(\bar{b}_n^{(k)} - d_k),\ S_{n,k} = n^{-1/2}\sum_i (a_{n,i}^{(k)} - a_k)(d_{n,i}^{(k)} - d_k).$$

For any $t_k \in R, k = 1, \ldots, s_n$, and $\epsilon > 0$

$$\begin{aligned}
&\mathrm{P}(T_{n,k} \le t) \le \mathrm{P}(S_{n,k} \le t + \epsilon^2) \\
&\quad + \mathrm{P}\{n^{1/2}|(\bar{a}_n - a)|_\infty \ge n^{1/4}\epsilon\} + \mathrm{P}\{n^{1/2}|(\bar{d}_n - d)|_\infty \ge n^{1/4}\epsilon\}, \qquad \text{(A.5)} \\
&\mathrm{P}\{n^{1/2}|(\bar{a}_n - a)|_\infty \ge n^{1/4}\epsilon\} \\
&= \mathrm{P}(|W_1|_\infty \ge n^{1/4}\epsilon) + o(1) = O\left\{\frac{n^{-1/4}(\log s_n)^{1/2}}{\epsilon}\right\} + o(1), \qquad \text{(A.6)} \\
&\mathrm{P}(n^{1/2}|(\bar{d}_n - d)|_\infty \ge n^{1/4}\epsilon) \\
&= \mathrm{P}(|W_2|_\infty \ge n^{1/4}\epsilon) + o(1) = O\left\{\frac{n^{-1/4}(\log s_n)^{1/2}}{\epsilon}\right\} + o(1), \qquad \text{(A.7)} \\
&\sup_{t\in R} |\mathrm{P}(S_{n,k} \le t + \epsilon^2) - \mathrm{P}(Z_k \le t + \epsilon^2)| = O(n^{-1/2}B_n^{3/2}) \qquad \text{(A.8)}
\end{aligned}$$

where W_1 (or W_2) is s_n-dim zero-mean Gaussian vector with covariance matrix identical to that of $n^{1/2}\bar{a}_n$ (or $n^{1/2}\bar{d}_n$), while Z_k is $N(0, \sigma_{k,k}s_{k,k})$; here (A.6) and (A.7) follow from Proposition 2.1 of Chernozhukov, Chetverikov and Kato (2017), Lemma D.3 of Chernozhukov, Chetverikov and Kato (2015) and the Chebyshev inequality, while (A.8) is a result of the Berry-Esseen Bounds and (A.4).

Reverse the direction of the inequality in (A.5), we have

$$\begin{aligned}
\mathrm{P}(T_{n,k} \le t) \ge \mathrm{P}(S_{n,k} \le t - \epsilon^2) &- \mathrm{P}\{n^{1/2}|(\bar{a}_n - a)|_\infty \ge n^{1/4}\epsilon\} \\
&- \mathrm{P}\{n^{1/2}|(\bar{d}_n - d)|_\infty \ge n^{1/4}\epsilon\};
\end{aligned}$$

for the three terms to the RHS, results parallel to (A.6)–(A.8) could be similarly proved. Thus

$$\begin{aligned}
&\max_{k=1,\ldots,s_n} \sup_{t\in R} |\mathrm{P}(T_{n,k} \le t) - \mathrm{P}(Z_k \le t)| \\
&= O\left\{\epsilon^2 + n^{-1/2}B_n^{3/2} + \frac{n^{-1/4}(\log s_n)^{1/2}}{\epsilon)}\right\} + o(1),
\end{aligned}$$

where the right hand side is $o(1)$, if $n^{-1/4}(\log s_n)^{1/2} = o(1)$. This proves the first assertion on the (null) marginal distribution.

As for the joint (null) distribution, first note that similar to (A.6),

$$\begin{aligned}
\mathrm{P}\left(\bigcap_{k=1}^{s_n}\{T_{n,k} \le t_k\}\right) \le \mathrm{P}\left(\bigcap_{k=1}^{s_n}\{S_{n,k} \le t + \epsilon^2\}\right) &+ \mathrm{P}\{n^{1/2}|(\bar{a}_n - a)|_\infty \ge n^{1/4}\epsilon\} \\
&+ \mathrm{P}\{n^{1/2}|(\bar{d}_n - d)|_\infty \ge n^{1/4}\epsilon\}.
\end{aligned}$$

We could then again apply Proposition 2.1 of Chernozhukov, Chetverikov and Kato (2017) and Nazarov's inequality (Nazarov, 2003) to see that

$$\sup_{t_1,\ldots,t_{s_n}\in R}\left|\mathrm{P}\left(\bigcap_{k=1}^{s_n}\{S_{n,k}\le t_k\}\right)-\mathrm{P}\left(\bigcap_{k=1}^{s_n}\{Z_k\le t_k\}\right)\right|$$
$$=O\left\{\frac{B_n^2\log^7(ns_n)}{n}\right\}^{1/6}=o(1),$$
$$\sup_{t_1,\ldots,t_{s_n}\in R}\left|\mathrm{P}\left(\bigcap_{k=1}^{s_n}\{Z_k\le t_k\}\right)-\mathrm{P}\left(\bigcap_{k=1}^{s_n}\{Z_k\le t_k+\epsilon^2\}\right)\right|\le C\epsilon^2(\log s_n)^{1/2}.$$

The proof is thus complete if ϵ could be chosen such that $\epsilon=o\{n^{-1/4}(\log s_n)^{1/2}\}$ and $\epsilon=o\{(\log s_n)^{-1/4}\}$.

Lemma 2. *With probability one,*

$$\sup_{t_1,\ldots,t_K\in R}\left|R_n(t_1,\ldots,t_K|\mathbf{X}_1^n,\mathbf{Y}_1^n)-\mathrm{P}\left(\bigcap_{k=1}^{K}\{Z_k\le t_k\}\right)\right|=o(1).$$

Proof of Lemma 2. For ease of exposition, the conclusion will be proved for the case where $\{d^{(k)}(\cdot),k=1,\ldots,K\}$ are all binary. The proof could be trivially adapted for the more general cases, where $k=1,\ldots,K$, $d^{(k)}(\cdot)$ is categorical taking a finite number of values.

For $k=1,\ldots,K$, and $i=1,\ldots,n$, write

$$a_{n,i}^{(k)}=a_k(X_i),\ c_{n,i}^{(k)}=n^{-1/2}(a_{n,i}^{(k)}-\bar{a}_n^{(k)}),\ d_{n,i}^{(k)}=d^k(Y_i).$$

Re-arrange the n observations $\{(X_i,Y_i):i=1,\ldots,n\}$ via the following steps: firstly, move the rows with $d_{n,i}^{(1)}=0$ ahead of those rows with $d_{n,i}^{(1)}=1$; secondly, for rows with the same value of $d_{n,i}^{(1)}$, sort them according to the value of $\{d_{n,i}^{(2)},i=1,\ldots,n\}$ (again in ascending order); repeat this process until the last step where the rows with the same value of $d_{n,i}^{(k)}$, for all $k=1,\ldots,K-1$, are sorted according to their respective value of $d_{n,i}^{(K)}$. To illustrate, in the case of $K=3$, there are eight possibilities for the rows of $n\times 3$ matrix, and they are arranged in the following order:

(1) 0 0 0	(5) 1 0 0
(2) 0 0 1	(6) 1 0 1
(3) 0 1 0	(7) 1 1 0
(4) 0 1 1	(8) 1 1 1

For the re-arranged observations, without loss of generality, we still use $c_{n,i}^{(k)}$ and $d_{n,i}^{(k)}$ $i=1,\ldots,n,k=1,\ldots,K$, to denote the corresponding 'coefficients' and 'scores'. As a result of the strong law of large numbers (SLLN), there exist square

integrable functions $\phi_k(\cdot), k = 1, \ldots, K$, on (0,1), such that

$$\sup_{n\to\infty} \int_0^1 \left\{\phi_k(u) - d^{(k)}_{n,1+[nu]}\right\}^2 du = 0, \quad k = 1, \ldots, K. \tag{A.9}$$

For illustration purposes, here we only give the specific forms for $\psi_1(\cdot)$ and $\psi_2(\cdot)$; the explicit form of other $\phi_k(\cdot)$ could be derived through similar arguments. Let $B_1 = \{y \in R^q : d^1(Y) = 0\}$, $B_2 = \{y \in R^q : d^2(Y) = 0\}$, $q_1 = \Pr(Y \in B_1)$, $q_2 = \Pr(Y \in B_2)$, $q_{1,2} = \Pr(Y \in B_1 \cap B_2)$. Then $\psi_1(\cdot)$ and $\psi_2(\cdot)$ could be defined as:

$$\phi_1(u) = \begin{cases} 0, u \in (0, q_1) \\ 1, o.w. \end{cases}, \qquad \phi_2(u) = \begin{cases} 0, u \in (0, q_{1,2}) \\ 1, u \in [q_{1,2}, q_1) \\ 0, u \in [q_1, q_1 + q_2 - q_{1,2}) \\ 1, u \in [q_1 + q_2 - q_{1,2}, 1). \end{cases}$$

For these 'score' functions, it holds that for any $k, l = 1, \ldots, K$,

$$\bar{\psi}_k := \int_0^1 \psi_k(u)du = 1 - q_k, \quad \int_0^1 \phi_k(u)\psi_l(u)du = 1 - q_k - q_l + q_{k,l}, \quad k \neq l.$$

In view of (A.9), when π is uniformly distributed on S_n, so that $\pi(i)$ is the rank of U_i, with $U_1, \ldots, U_n$ i.i.d. $U(0,1)$, we could apply Theorem 6.1 of Hajek, Sidak and Sen (1999) and claim that for all $k = 1, \ldots, K$,

$$T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^{n,\pi}) = \sum_i c^{(k)}_{n,i} d^{(k)}_{n,\pi(i)} = \sum_i c^{(k)}_{n,i}\{\psi_k(U_i) - \bar{\psi}_k\} + o_p(1). \tag{A.10}$$

Thus conditional on $(\mathbf{X}_1^n, \mathbf{Y}_1^n)$, $\{T_{n,k}(\mathbf{X}_1^n, \mathbf{Y}_1^{n,\pi}), k = 1, \ldots, K\}$ are jointly normal with covariance matrix given by

$$\mathrm{Cov}\left[\sum_i c^{(k)}_{n,i}\psi_k(U_i), \sum_i c^{(l)}_{n,i}\psi_l(U_i)\right] = \sum_i c^{(k)}_{n,i} c^{(l)}_{n,i}(q_{k,l} - q_k q_l).$$

Moreover, by SLLN, with probability one,

$$\sum_i c^{(k)}_{n,i} c^{(l)}_{n,i} = \frac{1}{n}\sum_i (a^{(k)}_{n,i} - \bar{a}^{(k)}_n)(a^{(l)}_{n,i} - \bar{a}^{(l)}_n) \to \sigma_{k,l},$$

and the proof is thus complete.

Acknowledgments

We thank the associate editor and two anonymous referees for their constructive comments and suggestions. Kong and Liu are joint first authors. Kong was supported by the National Natural Science Foundation of China (12271081 and 11931014). Liu is supported by the Natural Science Foundation of

Sichuan Province (2023NSFSC1357). Xia was supported by the National Natural Science Foundation of China (72033002 and 11931014) and the MOE's Academic Research Fund (A-8000021-00-00) of Singapore.

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.

Bancroft, T., Du, C. and Nettleton, D. (2013). Estimation of false discovery rate using sequential permutation p-values. *Biometrics* **69**, 1–7.

Cai, 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.

Chang, W. and Tian, W. (2016). GSA-Lightning: Ultra-fast permutation-based gene set analysis. *Bioinformatics* **32**, 3029–3031.

Chen, S., Li, J. and Zhong, P. (2019). Two-sample and ANOVA tests for high dimensional means. *The Annals of Statistics* **47**, 1443–1474.

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

Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields* **162**, 47–70.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* **45**, 2309–2352.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *Review of Economic Studies* **86**, 1867–1900.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics* **41**, 484–507.

Chung, E. and Romano, J. P. (2016). Multivariate and multiple permutation tests. *Journal of Econometrics* **193**, 76–91.

DiCiccio, C. and Romano, J. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association* **112**, 1211–1220.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics* **1**, 107–129.

Fan, J., Liao, Y. and Yao, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83**, 1497–1541.

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

Hajek, J., Sidak, Z. and Sen, P. K. (1999). *Theory of Rank Tests*. 2nd Edition. Academic, New York.

Heller, R., Heller, Y. and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100**, 503–510.

Heller, R., Heller, Y., Kaufman, S., Brill, B. and Gorfine, M. (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *Journal of Machine Learning Research* **17**, 978–1031.

Huo, X. and Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics* **58**, 435-447.

Kim, M. H. and Akritas, M. G. (2010). Order thresholding. *The Annals of Statistics* **38**, 2314–2350.

Kock, A. B. and Preinerstorfer, D. (2019). Power in high-dimensional testing problems. *Econometrica* **87**, 1055–1069.

Kong, E., Wang, L., Xia, Y. and Liu, J. (2020). Permutation test for two-sample means and signal identification of high-dimensional data. *Statistica Sinica* **32**, 89–108.

Kosorok, M. R. and Ma. S. (2007). Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *The Annals of Statistics* **35**, 1456–1486.

Liang, K. (2016). False discovery rate estimation for large-scale homogeneous discrete p-values. *Biometrics* **72**, 639–648.

Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis* **LNM 1807**, 169–187. Springer, Berlin & Heidelberg.

Nettleton, D., Recknor, J. and Reecy, J. M. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics* **24**, 192–201.

Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **80**, 5–31.

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.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York.

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

Xue, K. and Yao, F. (2020). Distribution and correlation free two-sample test high dimensional means. *The Annals of Statistics* **48**, 1304–1328.

Zeng, X., Xia, Y. and Tong, H. (2018). Jackknife approach to the estimation of mutual information. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 9956–9961.

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

(Received October 2022; accepted April 2023)