Electronic Journal of Statistics
Vol. 16 (2022) 4090–4132
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2033

Two-sample test for equal distributions in separate metric space: New maximum mean discrepancy based approaches*

Jin-Ting Zhang

Department of Statistics and Applied Probability,
National University of Singapore,
3 Science Drive 2, Singapore 117546, Singapore
e-mail: stazjt2020@nus.edu.sg

and

Łukasz Smaga

Faculty of Mathematics and Computer Science,
Adam Mickiewicz University,
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
e-mail: ls@amu.edu.pl

Abstract: This article develops statistical methods for testing the equality of two distributions based on two independent samples generated in some separable metric space. Such methods are broadly applicable in identifying similarity or distinction of two complicated data sets (e.g., high-dimensional data or functional data) collected in a wide range of research or industry areas, including biology, bioinformatics, medicine, material science, among others. Recently a so-called maximum mean discrepancy (MMD) based approach for the above two-sample problem has been proposed, resulting in several interesting tests. However, the main theoretical and numerical results of these MMD based tests depend on the very restricted assumption that the two samples have equal sample sizes. In addition, these tests are generally implemented via permutation when the equal sample size assumption is violated. In real data analysis, this equal sample size assumption is hardly satisfied, and dropping away some of the observations often means the loss of priceless information. It is also of interest to know if an MMD-based test can be conducted generally without using permutation. In this paper, we further study this MMD based approach with the equal sample size assumption removed. We establish the asymptotic null and alternative distributions of the MMD test statistic and its root-n consistency. We propose methods for approximating the null distribution, resulting in easy and quick implementation. Numerical experiments based on artificial data and two real data sets from two different areas of applications demonstrate that in terms of control of the type I error level and power, the resulting tests perform better or no worse than several existing competitors.

MSC2020 subject classifications: Primary 62H15; secondary 62G10.
Keywords and phrases: Equality of distribution, hypothesis testing, maximum mean discrepancy, three-cumulant matched chi-square approximation, two-sample problem.

Received February 2021.

*Zhang's work was supported by the National University of Singapore Academic Research grant R-155-000-212-114.

1. Introduction

1.1. Two-sample problem and some applications

With the development of data collection techniques, complicated multivariate data objects (e.g., high-dimensional data or functional data) in some separable metric space are frequently encountered in various areas including biology, bioinformatics, medicine, material science, among others. For such data, we focus on the problem of comparing two samples based on their distributions. We investigate a statistical test, which compares the null hypothesis of equal distributions against the alternative hypothesis, in which distributions are not equal. This is known as the two-sample problem.

Statistical tests for this problem are applicable in many areas. One of them is an important issue of data integration, i.e., we investigate if two samples are part of the same larger dataset, or if they should be treated as originating from two different sources. For example, in bioinformatics, the two samples of microarray data based on different experimental methods and lab facilities are often of interest. Specimens obtained from these different settings need to be compared to decide whether all specimens can be analyzed together. This is reasonable since when we do not reject the null hypothesis, the two samples can be combined into one larger sample, which can be used for further and more accurate analysis. On the other hand, rejecting the equality of distributions indicates a difference in the way samples are generated, which means observations should not be integrated directly. It is also of interest to differentiate between groups of people who are healthy or ill, or who suffer from different subtypes of a particular cancer. One can also think about integrating two sets of observations for different subtypes of cancer into one joint class, or treating them as distinct sets.

For the two-sample problem, there are many solutions in the literature. The first one is probably the multivariate version of the t-test by Hotelling [14], which however is constructed based on the assumption of multivariate normality with identical and unknown covariance structure, limiting its applicability. This drawback is overcome by non-parametric test such as the Kolmogorov-Smirnov test and the Wald-Wolfowitz runs test and their multivariate extensions [3, 9]. Nevertheless, there are many modern tests with better properties. Some of them are the further extensions of these classical tests. For example, the shrinkage-based diagonal Hotelling's tests [7]. In particular, the tests using characteristic functions deserve attention. They are constructed based on an estimator of the weighted distance between characteristic functions of two random vectors with a certain weight function. Perhaps the most famous one is the so-called energy test [22], while recently, Chen et al. [5] propose a modification of this test by using a density of some random distribution as the weight function. Several tests proposed in [17] are also interesting. The other class of tests is based on the maximum mean discrepancy (MMD), which is also considered in this paper. We review known results in more detail in the next section.

1.2. Existing maximum mean discrepancy based tests

The MMD based two-sample test for equal distributions is introduced by Borgwardt et al. [4]. Their results are expanded by Smola et al. [21], which use a Hilbert space embedding. In these tests, the test statistic is based on the maximum deviation of the expectation of a function evaluated on each of the random variables, taken over a sufficiently rich function class. The choice of this class is crucial. Fortunately, considering a reproducing kernel Hilbert space, the MMD statistic can be easily derived (see Section 2). For a given data type, the kernel can be appropriately chosen. For this reason, the MMD test is applicable in all data types, such as vectors, strings, and graphs [4, Section 2.3].

Unfortunately, in all papers about MMD tests cited here, the main results are obtained assuming the sample sizes are equal. In real data analysis, this equal sample size assumption is hardly satisfied. In fact, among the two data sets presented in Section 5.3, the two samples of any data set have different sample sizes. In these cases, to apply the tests mentioned above, one has to remove some of the observations so that the resulting two samples have the same sample size. This, however, often means the loss of priceless information. This equal sample size assumption is a technical condition, which allows good theoretical results and easy implementation of the MMD based tests, but it is not a necessary one. In fact, to construct a test, Borgwardt et al. [4] use the asymptotic normal distribution of test statistic, which asymptotically controls the type I error level and is consistent under fixed alternatives. However, the empirical sizes of this test are equal to zero in their experiments (see Tables 1-3 in [4]). Thus, this test is extremely conservative for small and moderate sample sizes, which may result in some loss of power.

Gretton et al. [10, 11, 13] derive a more accurate asymptotic null distribution, which is a χ^2-type mixture. Based on it, different tests are proposed in these papers, e.g., using the Gamma approximation, the null distribution estimate using the empirical Gram matrix spectrum, the Pearson curves approximation, and the resampling procedures. The two former methods are found to be less computationally intensive but perform less accurately in some cases, which was noticed in [13]. On the other hand, the reverse is true for the latter two procedures. Finally, Gretton et al. [12] derive the asymptotic null and alternative distributions allowing for different sample sizes, but the resulting null limiting distribution has a very complicated form, which motivates little applications; see Remark 3.1 for some discussion. They do not use it and consider mainly the case of equal sample sizes as well as slightly modified approximations to the null distribution from earlier papers (e.g., Gamma and Pearson's ones). One other thing is worth mentioning. In the numerical experiments of all papers cited in this paragraph, an unrealistically large sample size is usually used, which is needed for proper size control. Furthermore, these tests are generally implemented via permutation when the equal sample size assumption is violated.

1.3. Goals and results

As described above, the existing MMD based tests have some disadvantages. Thus, it is of interest to construct a test using MMD, which

- is available for equal as well as unequal sample sizes, which is more realistic in practice.
- maintains the type I error level and has reasonable power for all cases of a number of observations and their dimension. In particular, for high-dimensional small sample size setting.
- can be implemented easily without using permutation.

Moreover, such a test could have some good theoretical properties.

In this paper, we propose a two-sample test satisfying the above conditions. Namely, we consider the use of a MMD statistic for the two-sample problem without assuming the same sample sizes. We base our test on the unbiased estimator of squared MMD. Under the null hypothesis, we prove that the asymptotic distribution of test statistic is a χ^2-type mixture, which will be used to construct a test. Considering fixed and local alternative hypotheses, we show the consistency of new testing procedures, indicating good power behavior. Theoretical results are established under mild conditions. For implementation, we use the three-cumulant matched χ^2-approximation; see Zhang [24]. To apply it, the asymptotic null distribution of test statistic is first considered. Although the resulting test performs very well for large sample sizes, it is generally conservative for small and moderate ones. Thus we derive the first three cumulants of the test statistic and use them in the second method, which is also accurate for small and moderate number of observations. The resulting two new tests are easy to implement with no permutation involved. The finite sample behavior of the new tests and their comparison with several existing tests are studied in numerical experiments based on artificial and real data.

The remainder of this paper is organized as follows. Section 2 presents the hypothesis testing problem and its “kernel counterpart”. The construction of the test statistic is also stated there. In Section 3, the asymptotic null distribution and power of the new test statistic are proved under mild conditions. Section 4 contains the construction of two new testing procedures. The numerical experiments are presented in Section 5. Finally, Section 6 concludes the paper. The proofs of theoretical results and the detailed numerical results are presented in Appendices A and B respectively. In the Supplementary Materials, the codes to perform new methods and to conduct simulations are given.

2. Hypotheses and test statistic

Assume that we have the following two samples of observed random elements in $\mathcal{Y}$, a separable metric space:

$$y_{\alpha 1},\dots,y_{\alpha n_\alpha} \overset{\text{i.i.d.}}{\sim} P_\alpha \ (\alpha = 1, 2), \tag{1}$$

where P_1 and P_2 are unknown Borel probability measures on $\mathcal{Y}$ as defined in Borgwardt et al. [4]. Let $n = n_1 + n_2$ denote the total sample size. Of interest is to test if the two Borel probability measures are the same:

$$H_0 : P_1 = P_2, \text{ versus } H_1 : P_1 \neq P_2. \tag{2}$$

Let $K(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \to R$ be a characteristic reproducing kernel. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) generated by $K(\cdot,\cdot)$. For any $u, v \in \mathcal{H}$, the inner product and L^2-norm of $\mathcal{H}$ are defined as $\langle u, v\rangle$ and $\|u\| = \langle u, u\rangle^{1/2}$ respectively. Let $\phi(y) = K(\cdot, y) : \mathcal{Y} \to \mathcal{H}$ be the canonical feature mapping. It follows that $\phi(\mathcal{Y}) \subset \mathcal{H}$. Using this feature mapping, we then have the following two induced samples of random elements in the RKHS $\mathcal{H}$:

$$x_{\alpha 1} = \phi(y_{\alpha 1}), \ldots, x_{\alpha n_\alpha} = \phi(y_{\alpha n_\alpha}) \ (\alpha = 1, 2). \tag{3}$$

Set $\mu_\alpha = E(x_{\alpha 1})$ $(\alpha = 1, 2)$. According to Borgwardt et al. [4], $\mathrm{MMD}^2(P_1, P_2) = \|\mu_1 - \mu_2\|^2$, and hence testing (2) using the two samples (1) is equivalent to testing the following hypotheses using the two induced samples (3):

$$H_0 : \mu_1 = \mu_2, \text{ versus } H_1 : \mu_1 \neq \mu_2. \tag{4}$$

Notice that the dimension of μ_1 and μ_2 can be very large. To test (4), following [2, 6, 26], an L^2-norm based test statistic using (3) can be constructed by

$$T_n = \frac{n_1 n_2}{n}(S_{11} + S_{22} - 2S_{12}), \tag{5}$$

where for $\alpha = 1, 2$,

$$\begin{aligned} S_{\alpha\alpha} &= \tfrac{2}{n_\alpha(n_\alpha - 1)} \textstyle\sum_{1 \le i < j \le n_\alpha} \langle x_{\alpha i}, x_{\alpha j}\rangle, \\ S_{12} &= \tfrac{1}{n_1 n_2} \textstyle\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \langle x_{1i}, x_{2j}\rangle = \langle \bar{x}_1, \bar{x}_2\rangle \end{aligned} \tag{6}$$

($\bar{x}_\alpha = n_\alpha^{-1} \sum_{i=1}^{n_\alpha} x_{\alpha i}$) are unbiased estimators for $\|\mu_1\|^2$, $\|\mu_2\|^2$ and $\langle \mu_1, \mu_2\rangle$ respectively, so that $S_{11} + S_{22} - 2S_{12}$ estimates $\mathrm{MMD}^2(P_1, P_2)$ unbiasedly [4]. Notice that for (4), a linear-time statistic is proposed and studied in [15].

Notice that the two induced samples (3) are not directly computable, since the canonical feature mapping $\phi(y)$ is implicitly defined through the reproducing kernel. Fortunately, the reproducing kernel $K(\cdot,\cdot)$ and its canonical feature mapping $\phi(\cdot)$ have the useful kernel trick $K(y, y') = \langle \phi(y), \phi(y')\rangle$. Thus, using (3) and (6), we have

$$S_{\alpha\alpha} = \frac{2}{n_\alpha(n_\alpha - 1)} \sum_{1 \le i < j \le n_\alpha} K(y_{\alpha i}, y_{\alpha j}), \ S_{12} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(y_{1i}, y_{2j}),$$

where $\alpha = 1, 2$. Therefore, we can compute T_n in (5) using the original two samples (1) easily.

3. Asymptotic properties

3.1. Asymptotic null distribution

Let $\tilde{K}(y,y')$ denote the centered version of $K(y,y')$, i.e.,

$$\begin{aligned}\tilde{K}(y,y') = &\ \langle \phi(y)-\mu, \phi(y')-\mu'\rangle \\ = &\ K(y,y') - E_{z'}(K(y,z')) - E_z(K(z,y')) + E_{z,z'}(K(z,z')),\end{aligned} \tag{7}$$

where $\mu = E(\phi(y))$, $\mu' = E(\phi(y'))$, z and z' are independent and they are independent copies of y and y' respectively. Then by some algebra (see Appendix A), we have

$$\begin{aligned}S_{\alpha\alpha} = &\ \tilde{S}_{\alpha\alpha} + 2\langle \bar{x}_\alpha - \mu_\alpha, \mu_\alpha\rangle + \|\mu_\alpha\|^2, \\ S_{12} = &\ \tilde{S}_{12} + \langle \bar{x}_1 - \mu_1, \mu_2\rangle + \langle \bar{x}_2 - \mu_2, \mu_1\rangle + \langle \mu_1, \mu_2\rangle,\end{aligned} \tag{8}$$

where $\alpha = 1, 2$ and

$$\tilde{S}_{\alpha\alpha} = \frac{2}{n_\alpha(n_\alpha - 1)} \sum_{1\le i<j\le n_\alpha} \tilde{K}(y_{\alpha i}, y_{\alpha j}),\ \tilde{S}_{12} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \tilde{K}(y_{1i}, y_{2j}). \tag{9}$$

It is seen from (9) that in the expressions of $\tilde{S}_{11}$, $\tilde{S}_{22}$ and $\tilde{S}_{12}$, the mean embeddings of the two distributions have been subtracted. It follows that

$$T_n = \tilde{T}_n + 2Q_n + \frac{n_1 n_2}{n}\|\mu_1 - \mu_2\|^2, \tag{10}$$

where

$$\tilde{T}_n = \frac{n_1 n_2}{n}(\tilde{S}_{11} + \tilde{S}_{22} - 2\tilde{S}_{12}),\ Q_n = \frac{n_1 n_2}{n}\langle (\bar{x}_1 - \mu_1) - (\bar{x}_2 - \mu_2), \mu_1 - \mu_2\rangle. \tag{11}$$

It is seen that $\tilde{T}_n$ has the same distribution as that of T_n under the null hypothesis. Thus, studying the null distribution of T_n is equivalent to studying the distribution of $\tilde{T}_n$.

We impose the following assumptions:

Assumption 1. *We have* $y_{\alpha 1}, \ldots, y_{\alpha, n_\alpha} \overset{i.i.d.}{\sim} P$ $(\alpha = 1, 2)$ *where* P *is a probability measure on* $\mathcal{Y}$*.*

Assumption 2. *As* $n \to \infty$*, we have* $n_1/n \to \tau \in (0, 1)$*.*

Assumption 3. $K(y, y')$ *is a reproduced kernel such that* $E_y(\tilde{K}(y, y)) < \infty$*.*

Assumption 1 means that the null hypothesis is satisfied and the common probability measure of the two samples is P. Assumption 2 is a regularity condition for two-sample problems, and it requires that the group sample sizes tend to infinity proportionally. Assumption 3 ensures that $\tilde{K}(y, y')$ is square integrable, i.e., $E_{y,y'}(\tilde{K}^2(y, y')) < \infty$, and it has the following Mercer's expansion

$$\tilde{K}(y, y') = \sum_{r=1}^{\infty} \lambda_r \psi_r(y)\psi_r(y'), \tag{12}$$

where $\lambda_1, \lambda_2, \dots$ are the eigenvalues of $\tilde{K}(y, y')$ and $\psi_1(y), \psi_2(y), \dots$ are the associated orthonormal eigenelements in the sense that

$$\int_{\mathcal{Y}} \tilde{K}(y, y')\psi_r(y)P(dy) = \lambda_r \psi_r(y'), \ \int_{\mathcal{Y}} \psi_r(y)\psi_s(y)P(dy) = \delta_{rs}, \tag{13}$$

where $\delta_{rs} = 1$ when $r = s$ and 0 otherwise, $r, s = 1, 2, \dots$. In fact, under Assumption 3, by (12), we have

$$\begin{array}{rl} E_y(\tilde{K}(y, y)) = & \sum_{r=1}^{\infty} \lambda_r < \infty, \\ E_{y,y'}(\tilde{K}^2(y, y')) = & \sum_{r=1}^{\infty} \lambda_r^2 \le (\sum_{r=1}^{\infty} \lambda_r)^2 < \infty, \end{array} \tag{14}$$

where $y, y' \overset{\text{i.i.d.}}{\sim} P$. We have the following useful theorem, which is proved in Appendix A.

Theorem 3.1. *Under Assumptions 1–3, as $n \to \infty$, we have $\tilde{T}_n \xrightarrow{d} \tilde{T}$, where*

$$\tilde{T} \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r (A_r - 1), \ A_r \overset{i.i.d.}{\sim} \chi_1^2.$$

Remark 3.1. *Gretton et al. [12] used $S_{11} + S_{22} - 2S_{12}$ as their test statistic, which as mentioned earlier, estimates $\mathrm{MMD}^2(P_1, P_2)$ unbiasedly [4]. They obtained the asymptotic null distribution of $W_n = n(S_{11} + S_{22} - 2S_{12})$ as the distribution of the following random variable [12, Theorem 12]:*

$$\tilde{W} \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r \left(\left(\frac{z_{1r}}{\sqrt{\tau}} - \frac{z_{2r}}{\sqrt{1-\tau}} \right)^2 - \frac{1}{\tau(1-\tau)} \right), \tag{15}$$

where $\tau = \lim_{n\to\infty}(n_1/n)$ as defined in Assumption 2 and z_{1r}'s and z_{2r}'s are i.i.d. from $\mathcal{N}(0, 1)$. The random variable $\tilde{W}$ is rather complicated in form, motivating little applications. In fact, Gretton et al. [12] did not use the distribution of $\tilde{W}$ to approximate the null distribution of W_n.

Remark 3.2. *Although the expressions of $\tilde{T}$ and $\tilde{W}$ are very different in form, we can show that they are equal up to a constant factor. In fact, let $z_r = z_{1r}/\sqrt{\tau} - z_{2r}/\sqrt{1-\tau}$, $r = 1, 2, \dots$. We have $z_r \overset{i.i.d.}{\sim} N(0, 1/\tau + 1/(1-\tau))$, $r = 1, 2, \dots$, and hence $z_r^2 \overset{i.i.d.}{\sim} \chi_1^2/(\tau(1-\tau))$. It follows that $\tilde{T} \overset{d}{=} \tau(1-\tau)\tilde{W}$ as desired.*

Remark 3.3. *Although the expressions of $\tilde{T}$ and $\tilde{W}$ are equal up to a constant factor, the proof of Theorem 3.1, which we present in Appendix A, is different from that of Theorem 12 of [12].*

It is worthwhile to mention that the result of Theorem 3.1 can not be applied directly to conduct the proposed test T_n since the eigenvalues λ_r's are unknown and they depend on the null probability measure P as seen from (13). Nevertheless, we will use the result of Theorem 3.1 to approximate the null distribution of T_n in Section 4 for constructing new tests.

3.2. Asymptotic power

In this subsection, we investigate the asymptotic power of the proposed test under the following local alternative hypothesis:

$$H_{1n} : \mu_2 = \mu_1 + n^{-(1/2-\Delta)} h, \tag{16}$$

where $0 < \Delta \leq 1/2$ and h is a constant element in the RKHS $\mathcal{H}$ such that $0 < \|h\| < \infty$ and

$$\sigma_\alpha^2 = E(\langle x_{\alpha 1} - \mu_\alpha, h\rangle)^2 > 0 \ (\alpha = 1, 2). \tag{17}$$

Notice that when $\Delta = 1/2$, the local alternative hypothesis (16) reduces to a fixed alternative hypothesis $H_1 : \mu_2 = \mu_1 + h$ and when $0 < \Delta < 1/2$, (16) is a strict local alternative hypothesis. A strict local alternative hypothesis will tend to the null hypothesis as the total sample size n tends to infinity. A test is usually called to be root-n consistent if it can detect a strict local alternative hypothesis with probability tending to 1 as the total sample size tends to infinity. A root-n consistent test is often desired since the root-n rate is the best rate of a local alternative hypothesis, which can be detected by a test.

Under (16) and by (10) and (11), we can write T_n as

$$T_n = \tilde{T}_n + 2Q_n + \frac{n_1 n_2 \|h\|^2}{n^{2-2\Delta}}, \tag{18}$$

where

$$Q_n = \frac{n_1 n_2}{n^{3/2-\Delta}} \left(\langle \bar{x}_1 - \mu_1, h\rangle - \langle \bar{x}_2 - \mu_2, h\rangle\right). \tag{19}$$

Theorem 3.2. *Assume that* $|K(y, y')| \leq B_K$ *for all* $y, y' \in \mathcal{Y}$ *for some* $B_K < \infty$*. Then we have*

$$E(\tilde{T}_n) = 0, \ \operatorname{var}(\tilde{T}_n) \leq 64 B_K^2,$$

and

$$\operatorname{var}(Q_n) = n_1^2 n_2^2 n^{2\Delta-3} \left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right),$$

where $\sigma_1^2, \sigma_2^2 \leq 4\|h\|^2 B_K$*.*

Theorem 3.3. *Assume that* $|K(y, y')| \leq B_K$ *for all* $y, y' \in \mathcal{Y}$ *for some* $B_K < \infty$*. Then under Assumption 2 and the local alternative hypothesis (16), as* $n \to \infty$*, we have*

(a) $\tilde{T}_n/(\operatorname{var}(Q_n))^{1/2} \xrightarrow{p} 0$,

(b) $Q_n/(\operatorname{var}(Q_n))^{1/2} \xrightarrow{d} \mathcal{N}(0, 1)$,

(c) $[T_n - n_1 n_2 \|h\|^2/(n^{2-2\Delta})]/\sqrt{\operatorname{var}(T_n)} \xrightarrow{d} \mathcal{N}(0, 1)$, *and hence*

$$P(T_n \geq \hat{C}_\epsilon) = \Phi\left(\frac{n^\Delta \|h\|^2}{2(\sigma_1^2/\tau + \sigma_2^2/(1-\tau))^{1/2}}\right)(1 + o(1)) \to 1,$$

where $\hat{C}_\epsilon$ *denotes a consistent estimator of* C_ϵ*, the upper* 100ϵ *percentile of* $\tilde{T}_n$ *with* ϵ *being the given significance level,* τ *is defined in Assumption 2, and* $\Phi(\cdot)$ *denotes the cumulative distribution of* $\mathcal{N}(0, 1)$*.*

The proofs of Theorems 3.2 and 3.3 are given in Appendix A. Theorem 3.3(c) shows that the T_n test is root-n consistent, indicating the optimal power behavior under large sample sizes. In addition, Theorem 3.3(c) shows that the asymptotic power of T_n does not depend on the value of $\hat{C}_\epsilon$ as long as it is consistent for C_ϵ. In Section 4 below, we shall discuss how to obtain a consistent estimator of C_ϵ. For a finite number of observations, this test can also successfully detect the difference in distributions of samples, which will be established in numerical experiments of Section 5.

4. Implementation

In this section, we consider two T_n tests based on the three-cumulant (3-c) matched χ^2-approximation [24, 25]. First, the asymptotic null distribution of T_n given in Theorem 3.1 is used. Next, we propose an approximation based on the first three cumulants of T_n under the null hypothesis.

Since (by Theorem 3.1) $\tilde{T}$ is a χ^2-type mixture with unknown coefficients being the eigenvalues of $\tilde{K}(y, y'), y, y' \overset{\text{i.i.d.}}{\sim} P$, where P is the common probability measure of the two samples under the null hypothesis, it is very reasonable to approximate its distribution using the 3-c matched χ^2-approximation. The key idea is to approximate the distribution of $\tilde{T}$ using that of the random variable $R \overset{d}{=} \beta_0 + \beta_1\chi^2_d$. The parameters β_0, β_1 and d are determined via matching the first three cumulants of $\tilde{T}$ and R. The first three cumulants of R are given by $\beta_0 + \beta_1 d$, $2\beta_1^2 d$ and $8\beta_1^3 d$, while the first three cumulants of $\tilde{T}$ are

$$E(\tilde{T}) = 0,\ \text{var}(\tilde{T}) = 2M_2,\ E(\tilde{T}^3) = 8M_3$$

respectively, where $M_l = \sum_{r=1}^{\infty} \lambda_r^l$ $(l = 2, 3, \dots)$. Equating the first three-cumulants of $\tilde{T}$ and R then leads to

$$\beta_0 = -\frac{M_2^2}{M_3},\ \beta_1 = \frac{M_3}{M_2},\ d = \frac{M_2^3}{M_3^2}. \tag{20}$$

Since $\tilde{K}(y, y')$ is nonnegative definite, we have $\lambda_r \geq 0$ $(r = 1, 2, \dots)$ and $\lambda_{\max} = \max_r \lambda_r > 0$. Thus, $M_l > 0$ $(l = 1, 2, \dots)$. It follows that $\beta_0 < 0, \beta_1 > 0$ and $d > 0$. Actually, we can show that $d \geq 1$. Note that the skewness of $\tilde{T}$ can be expressed as

$$\frac{E(\tilde{T}^3)}{\text{var}^{3/2}(\tilde{T})} = \frac{8M_3}{(2M_2)^{3/2}} = (8/d)^{1/2}.$$

Thus the skewness of $\tilde{T}$ will become small as d increases.

Remark 4.1. *The 3-c matched χ^2-approximation to the distribution of $\tilde{T}$ is very accurate. In fact, Zhang [24] showed that the upper density approximation error bound for the 3-c matched χ^2-approximation to the distribution of $\tilde{T}$ is $O(M) + O(1/d)$, where $M = M_4/M_2^2$, showing that this upper density approximation error bound will disappear as M and $1/d$ tend to 0. The good performance of the 3-c matched χ^2-approximation is also partially verified by the simulation results presented in Section 5.*

Let $\hat{M}_2$ and $\hat{M}_3$ be the consistent estimators of M_2 and M_3. Plug them into (20), the consistent estimators of β_0, β_1 and d are then obtained as

$$\hat{\beta}_0 = -\frac{\hat{M}_2^2}{\hat{M}_3}, \quad \hat{\beta}_1 = \frac{\hat{M}_3}{\hat{M}_2}, \quad \hat{d} = \frac{\hat{M}_2^3}{\hat{M}_3^2}. \tag{21}$$

Then for any nominal significance level $\epsilon > 0$, let $\chi_d^2(\epsilon)$ denote the upper 100ϵ percentile of χ_d^2. Then using (21), the proposed test with the 3-c matched χ^2-approximation can then be conducted via using the approximate critical value $\hat{\beta}_0 + \hat{\beta}_1 \chi_{\hat{d}}^2(\epsilon)$ or the approximate p-value $P(\chi_{\hat{d}}^2 \geq (T_n - \hat{\beta}_0)/\hat{\beta}_1)$.

To implement the above 3-c matched χ^2-approximation, we need to estimate M_2 and M_3 consistently. For this purpose, we propose two methods. First, this can be done via estimating the unknown eigenvalues λ_r $(r = 1, 2, \dots)$ of $\tilde{K}(y, y')$ consistently, where $y, y' \overset{\text{i.i.d.}}{\sim} P$ with P being the common Borel probability measure of the two samples when the null hypothesis holds. Gretton et al. [13] pointed out that the empirical eigenvalues of the centered Gram matrix can be used to construct the consistent estimators of λ_r $(r = 1, 2, \dots)$. To this end, we pool the two samples (1) and denote it as

$$y_1, \dots, y_n. \tag{22}$$

Under the null hypothesis, we have $y_1, \dots, y_n \overset{\text{i.i.d.}}{\sim} P$. Let K be the $n \times n$ Gram matrix whose (i, j)th entry is $K(y_i, y_j)$ $(i, j = 1, \dots, n)$. Let 1_n denote an $n \times 1$ vector of ones and I_n denote the $n \times n$ identity matrix. Then $H_n = I_n - 1_n 1_n^\top / n$ is an $n \times n$ projection matrix of rank $n - 1$. Set $\tilde{K}^* = H_n K H_n$, which is usually called the centered Gram matrix whose (i, j)th entry is

$$\begin{aligned}\tilde{K}^*(y_i, y_j) =& K(y_i, y_j) - n^{-1} \sum_{v=1}^n K(y_i, y_v) \\ & - n^{-1} \sum_{u=1}^n K(y_u, y_j) + n^{-2} \sum_{u=1}^n \sum_{v=1}^n K(y_u, y_v)\end{aligned}$$

$(i, j = 1, \dots, n)$. For any fixed i and j, it is easily seen that as $n \to \infty$, by the law of large numbers, we have

$$\begin{aligned}\tilde{K}^*(y_i, y_j) &\xrightarrow{d} \tilde{K}(y_i, y_j) \\ &= K(y_i, y_j) - E_{y'}(K(y_i, y')) - E_y(K(y, y_j)) + E_{y,y'}(K(y, y')).\end{aligned}$$

The following theorem gives the uniform convergence rate of $\tilde{K}^*(y_i, y_j)$ to $\tilde{K}(y_i, y_j)$.

Theorem 4.1. *Assume that $|K(y, y')| \leq B_K$ for all $y, y' \in \mathcal{Y}$ for some $B_K < \infty$. Then under Assumption 1, as $n \to \infty$, we have*

$$|\tilde{K}^*(y_i, y_j) - \tilde{K}(y_i, y_j)| = O_p(n^{-1/2}) \qquad \textit{uniformly for all } y_i, y_j. \tag{23}$$

Let $\varpi_1, \dots, \varpi_q$ be all the non-zero eigenvalues of $\tilde{K}^*$, which can be obtained via an eigen-decomposition of $\tilde{K}^*$. Then $\hat{\lambda}_r = \varpi_r/n$ $(r = 1, \dots, q)$ are consistent estimators of λ_r $(r = 1, 2, \dots)$ [20]. The consistent estimators of M_2 and M_3 are then obtained as

$$\hat{M}_2 = \sum_{r=1}^{q} \hat{\lambda}_r^2, \;\; \hat{M}_3 = \sum_{r=1}^{q} \hat{\lambda}_r^3. \tag{24}$$

To show the consistency of $\hat{M}_2$ and $\hat{M}_3$, following Gretton et al. [13, Theorem 1], we impose the following condition:

$$\sum_{r=1}^{\infty} \sqrt{\lambda_r} < \infty. \tag{25}$$

Note that Condition (25) is stronger than Assumption 3, since we have

$$E[\tilde{K}(y, y)] = \sum_{r=1}^{\infty} \lambda_r \leq \left[\sum_{r=1}^{\infty} \sqrt{\lambda_r} \right]^2 < \infty. \tag{26}$$

Theorem 4.2. *Under Assumptions 1, 2 and Condition (25), as $n \to \infty$, we have $\hat{M}_\ell \xrightarrow{p} M_\ell, \ell = 2, 3$ and*

$$\hat{\beta}_0 \xrightarrow{p} \beta_0, \hat{\beta}_1 \xrightarrow{p} \beta_1, \hat{d} \xrightarrow{p} d.$$

Remark 4.2. *The above implementation depends on the large sample property of $\tilde{T}_n$ as stated in Theorem 3.1. In fact, in the experiments conducted in Gretton et al. [13, Section 4], the sample sizes of the two groups equal* 5000 *each, which is extremely large. Thus, for small and moderate sample sizes which are realistic in real data analysis, the above implementation can be very inaccurate. The simulation studies conducted in Section 5 indicate that the above implementation results in very conservative empirical sizes and this may also affect the associate power performance.*

Remark 4.2 motivates our second method of estimating M_2 and M_3. To take the moderate or small sample sizes into account, we can estimate the distribution of $\tilde{T}_n$ directly via approximating it using the distribution of R by matching the first three cumulants of $\tilde{T}_n$ and R. To this end, we first find out the first three cumulants of $\tilde{T}_n$ under the null hypothesis (Assumption 1) as in the following theorem. Its proof is deferred to Appendix A.

Theorem 4.3. *Under Assumption 1, the first three cumulants of $\tilde{T}_n$ are given by*

$$\begin{aligned}
E(\tilde{T}_n) = &\;\; 0, \\
\operatorname{var}(\tilde{T}_n) = &\;\; 2\left\{1 + \left(\frac{n_2^2}{n^2(n_1-1)} + \frac{n_1^2}{n^2(n_2-1)}\right)\right\} E(\tilde{K}^2(y, y')), \\
E(\tilde{T}_n^3) = &\;\; 8\left\{1 - \left(\frac{n_2^3}{n^3(n_1-1)^2} + \frac{n_1^3}{n^3(n_2-1)^2}\right)\right\} E(\tilde{K}(y, y')\tilde{K}(y, y'')\tilde{K}(y', y'')) \\
&+4\left(\frac{n_2^3 n_1}{n^3(n_1-1)^2} - \frac{2n_1n_2}{n^3} + \frac{n_1^3 n_2}{n^3(n_2-1)^2}\right) E(\tilde{K}^3(y, y')),
\end{aligned}$$

where $y, y', y'' \overset{i.i.d.}{\sim} P$. *Furthermore, under Assumptions 1–3, as* $n \to \infty$, *we have*

$$\begin{aligned} \operatorname{var}(\tilde{T}_n) = &\ 2E(\tilde{K}^2(y,y'))(1+o(1)), \\ E(\tilde{T}_n^3) = &\ 8E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y''))(1+o(1)). \end{aligned}$$

By Theorem 4.3, we may estimate M_2 and M_3 using the following estimators

$$\begin{aligned} \hat{M}_2 = &\ \left\{1+\left(\tfrac{n_2^2}{n^2(n_1-1)}+\tfrac{n_1^2}{n^2(n_2-1)}\right)\right\}\widehat{E}(\tilde{K}^2(y,y')), \\ \hat{M}_3 = &\ \left\{1-\left(\tfrac{n_2^3}{n^3(n_1-1)^2}+\tfrac{n_1^3}{n^3(n_2-1)^2}\right)\right\}\widehat{E}(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')) \\ &+\tfrac{1}{2}\left(\tfrac{n_2^3 n_1}{n^3(n_1-1)^2}-\tfrac{2n_1n_2}{n^3}+\tfrac{n_1^3 n_2}{n^3(n_2-1)^2}\right)\widehat{E}(\tilde{K}^3(y,y')), \end{aligned} \tag{27}$$

where

$$\begin{aligned} \widehat{E}(\tilde{K}^2(y,y')) =&\frac{2}{n(n-1)}\sum_{1\le i<j\le n}(\tilde{K}^*(y_i,y_j))^2, \\ \widehat{E}(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')) =&\frac{6}{n(n-1)(n-2)} \\ &\cdot\sum_{1\le i<j<k\le n}\tilde{K}^*(y_i,y_j)\tilde{K}^*(y_j,y_k)\tilde{K}^*(y_k,y_i), \\ \widehat{E}(\tilde{K}^3(y,y')) =&\frac{2}{n(n-1)}\sum_{1\le i<j\le n}(\tilde{K}^*(y_i,y_j))^3. \end{aligned}$$

Notice that the second term of $\hat{M}_3$ in (27) can be very small, even for a small sample size. The tests with and without this second term perform almost the same in simulations (data not shown). Thus we will throughout ignore this second term.

The following theorem shows that under some regularity conditions, $\hat{M}_\ell, \ell = 2, 3$ are consistent estimators of $M_\ell, \ell = 2, 3$ and hence $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{d}$ are consistent estimators of β_0, β_1 and d respectively.

Theorem 4.4. *Assume that* $|K(y, y')| \le B_K$ *for all* $y, y' \in \mathcal{Y}$ *for some* $B_K < \infty$. *Then under Assumptions 1 and 2, as* $n \to \infty$, *we have* $\hat{M}_\ell \xrightarrow{p} M_\ell, \ell = 2, 3$ *and*

$$\hat{\beta}_0 \xrightarrow{p} \beta_0, \hat{\beta}_1 \xrightarrow{p} \beta_1, \hat{d} \xrightarrow{p} d.$$

To sum up, we propose two tests using the 3-c matched χ^2-approximation methods with the estimators (24) and (27). The resulting tests are denoted as T_{3c1} and T_{3c2} respectively. We check their finite sample behavior in the next section.

5. Numerical experiments

5.1. Tests considered and general set-up

In this section, we conducted intensive numerical experiments to examine the performance of the proposed T_{3c1} and T_{3c2} tests in terms of controlling the type I error level, powers, and computational time against

- the energy test proposed in [22], denoted as T_{SR};
- the modified Cramér test with $\psi(t) = 1 - \exp(-t/2)$ proposed in [17, Page 2], denoted as T_{SG1};
- the marginal-modified Cramér test with $\psi(t) = 1 - \exp(-t/2)$ proposed in [17, Section 2], denoted as T_{SG2};
- the block-modified Cramér test with $\psi(t) = 1 - \exp(-t/2)$ proposed in [17, Section 3], denoted as T_{SG3}.

We used the energy test as a competitor since it is known to have good finite sample properties and is very popular in practice. T_{SG1} has the same test statistic as our test with the Gaussian radial basis function (RBF) kernel (see below). T_{SG2} and T_{SG3} are modifications of the Cramér test, constructed for high-dimensional settings by comparing marginal distributions. The null distributions of T_{SR} and T_{SGi} ($i = 1, 2, 3$) are approximated by permutation. Additionally, we compare our tests with the test procedures by Gretton et al., as was suggested by the reviewer (see Section 5.4).

In the implementation of T_{3c1} and T_{3c2}, we chose the kernel $K(\cdot, \cdot)$ to be the Gaussian RBF kernel $K(y, y') = \exp\{-\|y - y'\|^2/(2\sigma^2)\}$, where σ^2 is called a kernel width. It is easy to see that the above Gaussian RBF kernel is bounded above by 1 so that Assumption 3 is always satisfied and the conditions of Theorems 3.2 and 3.3 are also satisfied. Other kernels are also applicable. For an extensive list of kernels, we refer to [18]. Following [13], we may take σ^2 to be the squared median distance between observations in the pooled sample (22). The resulting tests are denoted as T_{3c1m} and T_{3c2m} respectively. Alternatively, following [17], we may take σ^2 to be the data dimension p and the resulting tests are denoted as T_{3c1p} and T_{3c2p} respectively. Note that the dimension based kernel width p may not be applicable for all the data (see Section 5.3).

Throughout this section, we took $\epsilon = 5\%$ to be the nominal significance level. For a test implemented by permutation, we used 1000 permutation runs for estimating the p-value. As usual, the empirical size and power of a test were computed as the proportion of rejections of the null hypothesis based on 1000 simulation runs. The simulation experiments were performed using the R program [23]. For T_{SR}, its implementation in the energy package [16] was used. The R codes for conducting the simulations are available in the Supplement Material.

5.2. *Artificial data sets*

Set-up: Let $\mu_* = (1, \ldots, p)^\top/(\sum_{i=1}^p i^2)^{1/2}$ and $\Sigma_\rho = (1 - \rho)I_p + \rho J_p$, where I_p and J_p denote the identity matrix and the matrix of ones of size $p \times p$ and $\rho \in (0, 1)$. We generated the two samples (1) as follows.

For size control, we set $y_{1i} = \mu + \Gamma^{1/2} u_{1i}$ ($i = 1, \ldots, n_1$) and $y_{2i} = \mu + \Gamma^{1/2} u_{2i}$ ($i = 1, \ldots, n_2$), where $u_{\alpha i} = (u_{\alpha i1}, \ldots, u_{\alpha ip})^\top$ ($\alpha = 1, 2$) and the components $u_{\alpha ir}$ are i.i.d. ($r = 1, \ldots, p$). As their distributions, we considered the normal distribution $\mathcal{N}(0, 1)$, the Student distribution with four degrees of freedom, t_4, and the chi-square distribution with one degree of freedom, χ^2_1. These three cases of distributions were denoted as Models 1, 2 and 3 respectively. We set $\mu = \mu_*$ and $\Gamma = \Sigma_\rho$ with $\rho = 0.2, 0.5, 0.8$.

On the other hand for power comparison, we set $y_{1i} = \mu_1 + \Gamma_1^{1/2} u_{1i}$ $(i = 1, \ldots, n_1)$ and $y_{2i} = \mu_2 + \Gamma_2^{1/2} u_{2i}$ $(i = 1, \ldots, n_2)$, where $\mu_1, \mu_2, \Gamma_1, \Gamma_2, u_{\alpha i} = (u_{\alpha i1}, \ldots, u_{\alpha ip})^\top$ were generated using the following five models ($\alpha = 1, 2$; $i = 1, \ldots, n_\alpha$; $r = 1, \ldots, p$):

- Model 4 with $\mu_1 = \mu_*$, $\mu_2 = \mu_* + \delta 1_p$, $\delta = 0.2, 0.4, 0.6$, 1_p is $p \times 1$ vector of ones, $\Gamma_1 = \Gamma_2 = \Sigma_\rho$, $\rho = 0.2, 0.5, 0.8$, $u_{\alpha ir} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$;
- Model 5 with $\mu_1 = \mu_2 = \mu_*$, $\Gamma_1 = \Sigma_{0.95}$, $\Gamma_2 = \Sigma_\rho$, $\rho = 0.3, 0.5, 0.7$, $u_{\alpha ir} \overset{\text{i.i.d.}}{\sim} t_4$;
- Model 6 with $\mu_1 = \mu_2 = \mu_*$, $\Gamma_1 = (\rho^{|i-j|})_{i,j=1}^p$, $\Gamma_2 = \Sigma_\rho$, $\rho = 0.2, 0.5, 0.8$, $u_{\alpha ir} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ or $u_{\alpha ir} \overset{\text{i.i.d.}}{\sim} t_4$;
- Model 7 with $y_{2i} = \mu_2 + \Gamma_2^{1/2}(u_{2i} + \delta v_i)$, where $\mu_1 = \mu_2 = \mu_*$, $\Gamma_1 = \Gamma_2 = \Sigma_\rho$, $\rho = 0.2, 0.5, 0.8$, $u_{\alpha i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, I_p)$, $\delta = 0.5, 0.75, 1$, $v_i = (v_{i1}, \ldots, v_{ip})^\top$ and $v_{ir} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$;
- Model 8 is Model 7 with $v_{ir} \overset{\text{i.i.d.}}{\sim} t_4$.

Under Models 4–8, the two generated samples do not have the same distribution so that the alternative hypothesis holds. Concretely speaking, under Model 4, the two mean vectors are different. Under Models 5 and 6, the two covariance matrices are different, but the one or two-dimensional marginal distributions are the same. On the other hand under Models 7 and 8, the one or two-dimensional marginal distributions are not the same. It is expected that the tests may have different performances under different models. We considered $p = 10, 100, 500$ and $(n_1, n_2) \in \{(20, 30), (40, 60)\}$.

Results: To save space, the empirical sizes and powers of the tests obtained under Models 1–8 are given in Tables 10–15 of Appendix B. Here Figure 1 presents the most important findings of these all results, which we describe in the following.

We first study the behavior of the tests under the null hypothesis (Models 1–3). Row 1 of Figure 1 (see also Table 10 in Appendix B) shows that when the null hypothesis holds, almost all tests control the type I error well except that T_{3c1p} is very conservative, most of its empirical sizes being smaller than 5% and T_{3c1m} is also quite conservative. That is, T_{3c1} is generally conservative regardless of which of the two kernel width choices is used. As mentioned in Remark 4.2, this is due to the fact that the sample sizes $(n_1, n_2) \in \{(20, 30), (40, 60)\}$ are too small for the asymptotical property of T_{3c1p} and T_{3c1m} as stated in Theorem 3.1 to take effect.

We now study the power behavior of the tests under various alternative hypotheses (Models 4–8) based on Rows 2–6 of Figure 1 (see also Tables 11–15 in Appendix B). We have several conclusions. First of all, T_{3c1p} and T_{3c2p} perform generally quite well under all the models and they are generally comparable with T_{SG1} and outperform the other tests with a good margin. Notice that

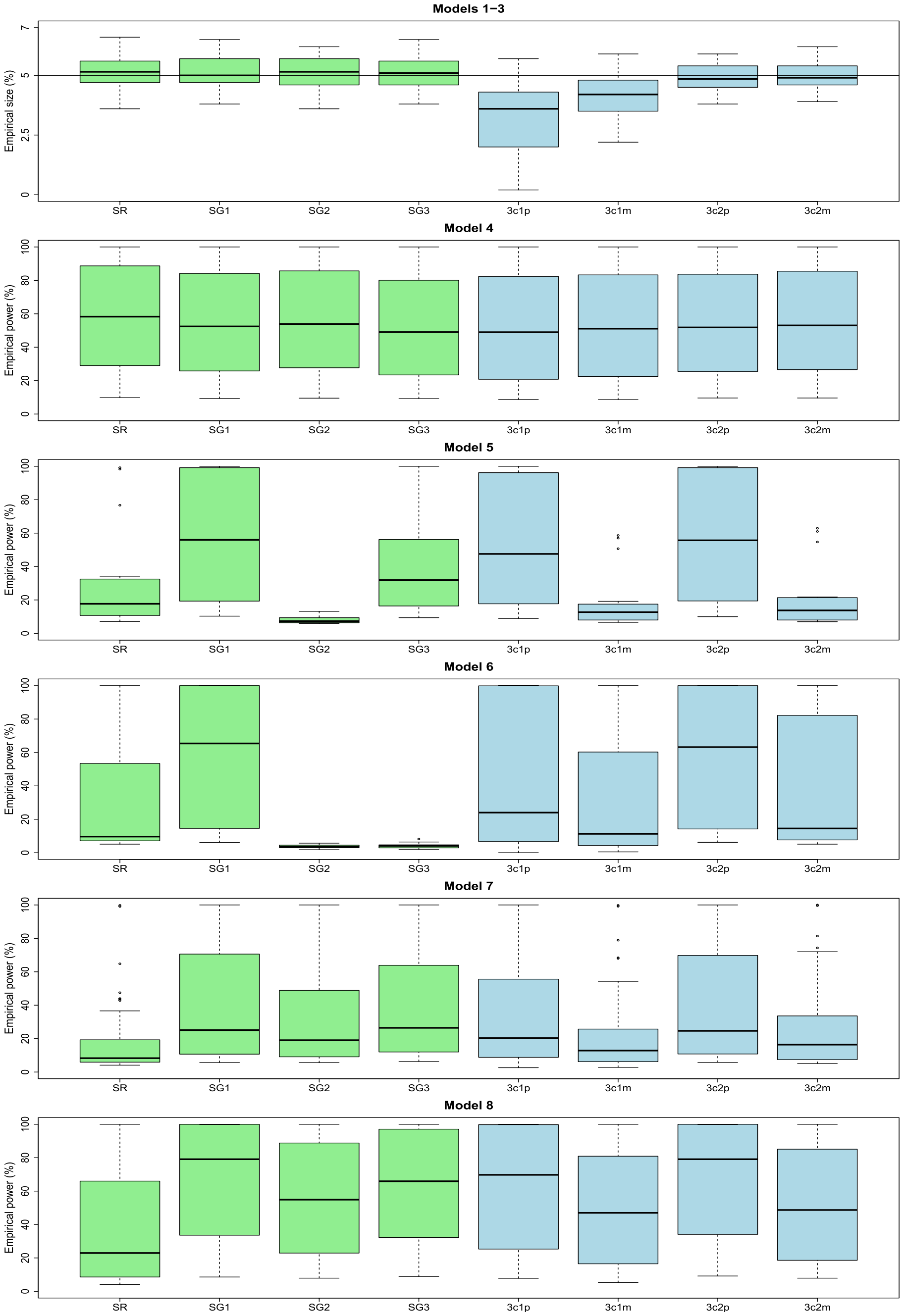

Fig 1: Box-and-whisker plots for the empirical sizes and powers obtained in Models 1–3 and Models 4–8 respectively. Green (respectively blue) box-and-whisker plots correspond to the known tests T_{SR}, T_{SG1}, T_{SG2} and T_{SG3} (respectively new tests T_{3c1p}, T_{3c1m}, T_{3c2p} and T_{3c2m}).

the test statistics of T_{3c1p}, T_{3c2p} and T_{SG1} are identical, but their implementations are quite different. The implementation of T_{SG1} depends on permutation, while the null distributions of T_{3c1p} and T_{3c2p} are approximated via the three-cumulant matched chi-square approximations without using any permutation or bootstrapping as described in Section 4. This means that our chi-square approximation approaches can work as well as permutation method, but they need much less time to conduct (see Section 5.5). Secondly, T_{3c1m}, T_{3c2m} and T_{SR} perform comparably under all the models. They perform comparably with T_{3c1p}, T_{3c2p}, and T_{SG1} in Model 4, but they perform generally worse than the latter tests in Models 5–8. This means that the power behavior of our tests is strongly affected by the kernel width choice. Thus, a study for the kernel width choice for our tests is interesting and warranted. Thirdly, T_{SG2} and T_{SG3} perform reasonably well under all the models except in Models 5 and 6. T_{SG2} has nearly no power in both Models 5 and 6 and T_{SG3} has nearly no powers in Model 6. This is not a surprise, because T_{SG2} and T_{SG3} are constructed via comparing the one or two-dimensional marginal distributions only. By the way, these two tests are very time-consuming, which is caused by their construction (see Section 5.5).

5.3. Real data sets

Data sets: Here, we considered two real data sets. The first one is the glass data set described in [8]. It contains two types of glass, having 70 and 76 observations respectively, which are characterized by $p = 9$ variables. The second one is the well-known colon data set [1] available at:
http://genomics-pubs.princeton.edu/oncology/affydata/index.html. It contains 40 tumor and 22 normal colon tissues, each having $p = 2000$ gene expression levels.

Two-sample tests: Notice that in each of the above data sets, there are two natural groups of observations. We first check if the two groups of each data set are generated from different distributions. The first two rows of Table 1 display the p-values of all tests. It is seen that all tests reject the null hypothesis for the glass data set, showing that the two natural groups of the glass data set are unlikely to have the same distribution. However, for the colon data, the results of the tests are not consistent: the null hypothesis was strongly rejected by $T_{SR}, T_{SG1}, T_{3c1m}$ and T_{3c2m}, but it was not rejected by $T_{SG2}, T_{SG3}, T_{3c1p}$ and T_{3c2p}. Notice that T_{3c1p} and T_{3c1m} (resp. T_{3c2p} and T_{3c2m}) are constructed in the same way except that the dimension based kernel width p used in T_{3c1p} and T_{3c2p} may not be at the same scale level as the squared distances of the observations while the squared median-distance based kernel width used in T_{3c1m} and T_{3c2m} are always at the same scale level as the squared distances of the observations. Thus, this problem may be solved via re-scaling each variable of the colon data using the pooled standard deviation of the variable so that the dimension-based kernel width p is at the same scale level as the squared distances of the transformed observations. The last row of Table 1 displays the p-values of all tests for the scaled colon data set. It is seen that the p-values of T_{3c1p} and T_{3c2p} are

TABLE 1
P-values of all tests for the glass and colon data sets.

Data set	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
glass	0.0010	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
colon	0.0010	0.0000	0.2070	0.2820	0.4759	0.0017	0.3229	0.0009
colon (scaled)	0.0420	0.0210	0.0550	0.0340	0.0365	0.0401	0.0265	0.0311

generally comparable with those of T_{3c1m} and T_{3c2m}, respectively, and almost all tests except T_{SG2} reject the null hypothesis at 5% significance level. This shows that the two natural groups of the colon data set are unlikely to have the same distribution.

Numerical experiments based on real data sets: It is often of interest to determine if all the tests have good size control and power for the real data sets. Here we present the numerical experiments based on the above glass and colon data sets.

Set-up: For each of the glass and colon data sets, we selected without replacement observations for two samples. For the type I error level (respectively power) study, we selected two samples of sizes n_1 and n_2 without replacement from the first group only (respectively from the first and second groups separately). For simplicity, we considered the balanced design with $n_1 = n_2$. The sample sizes depended on the data set and are reported with results below. For T_{SR} and T_{SGi} $(i = 1, 2, 3)$, the number of permutation runs was 1000. The empirical size and power of a test were computed using 1000 simulation runs.

Results: For the glass data set, the resulting empirical sizes and powers of the tests are presented in Table 2. It is seen that in terms of size control, all tests are comparable and are largely close to the nominal size 5% most of the time. These results are similar to those obtained from the numerical investigation based on the artificial data sets. In terms of power, T_{3c2m} performs best, followed by T_{SG3}, T_{SG2}, T_{3c1m}, T_{3c2p}, T_{SG1}, T_{SR}, while T_{3c1p} performs worst. Notice that unlike in the artificial data based experiment, in this real data based experiment, the squared median-distance based kernel width performs better than the dimension-based kernel width, since T_{3c2m} and T_{3c1m} generally have larger powers than T_{3c2p} and T_{3c1p} respectively.

For the unscaled colon data set, the resulting empirical sizes and powers of the tests are presented in Table 3. We have the following observations. First, both T_{SR} and T_{3c2m} have good size control and powers. Second, T_{3c1m} is rather conservative since many of its empirical sizes are less than 2%. This conservativity is expected since as mentioned in Remark 4.2, for small and moderate sample sizes, the asymptotic properties of the test statistic derived in Theorem 3.1 and used in the implementation of T_{3c1} do not take effect. Due to this conservativity, the empirical powers of T_{3c1m} are also affected and are generally smaller than those of T_{SR} and T_{3c2m}. Third, both T_{SG2} and T_{SG3} have good size control but their empirical powers are around 5% or even smaller than 5%, showing that both T_{SG2} and T_{SG3} do not perform well in the colon data set based experiments. Fourth, the empirical sizes and powers of T_{SG1}, T_{3c1p} and T_{3c2p} totally do not make sense since the empirical sizes and powers of T_{SG1}

TABLE 2
Empirical sizes and powers (in %) of all tests for the experiment based on the glass data set $(n = n_1 = n_2)$.

n	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
	Empirical sizes							
11	4.7	5.0	4.8	4.7	4.8	4.6	5.1	5.1
12	5.4	5.2	5.0	5.0	5.7	5.1	5.9	5.6
13	4.7	4.4	5.0	4.9	5.0	4.8	5.1	5.5
14	5.1	4.7	5.2	5.2	4.6	5.1	5.0	5.6
15	4.6	4.6	4.8	4.8	5.0	4.8	5.3	5.4
16	3.8	4.3	4.7	4.5	4.5	3.5	5.0	3.6
17	4.0	4.2	3.7	3.4	4.1	3.4	4.4	4.3
18	4.4	4.4	4.2	3.9	4.2	4.3	4.2	4.8
19	4.9	5.4	5.4	5.3	5.5	5.1	5.5	5.4
20	5.7	5.4	6.1	6.0	5.5	5.7	5.8	6.1
21	4.9	4.8	4.9	5.0	4.6	4.7	4.8	4.8
22	5.3	5.1	5.0	5.0	4.8	5.0	4.8	5.5
23	5.3	5.4	5.3	5.4	5.4	5.6	5.6	6.1
24	5.7	5.7	5.6	5.4	5.3	5.2	5.5	5.6
25	5.5	5.6	5.0	5.1	5.3	4.4	5.5	4.8
26	4.4	4.4	4.4	4.5	4.5	4.4	4.6	4.5
27	3.9	4.1	4.4	4.2	4.1	3.6	4.2	3.7
28	4.6	4.9	5.0	5.1	4.8	4.5	5.0	4.7
29	6.1	5.6	6.3	6.3	5.9	5.7	6.1	5.7
30	4.0	4.3	4.5	4.4	4.1	4.2	4.1	4.5
	Empirical powers							
11	18.1	24.2	28.5	31.0	12.7	27.4	25.9	38.4
12	23.2	27.4	33.6	36.1	15.5	33.5	29.4	43.9
13	26.2	29.6	39.0	40.4	17.9	36.2	31.3	48.4
14	30.5	35.1	43.3	47.0	22.1	44.1	38.1	55.0
15	38.4	42.5	51.2	54.3	26.3	49.8	44.5	61.4
16	43.2	48.4	56.9	58.6	32.5	54.3	49.3	64.4
17	46.9	52.6	62.1	64.4	36.0	58.6	54.2	68.8
18	55.3	59.6	65.4	69.2	42.8	65.6	60.8	74.8
19	63.9	63.6	71.8	74.5	49.5	72.1	66.5	79.7
20	64.8	63.9	71.4	73.8	52.4	73.4	66.1	80.8
21	72.4	69.9	78.1	80.3	55.8	77.3	71.9	85.3
22	77.8	75.6	80.7	84.0	64.1	82.7	77.0	86.7
23	80.5	79.3	83.3	86.2	66.0	84.2	79.6	89.9
24	86.7	83.7	90.1	91.9	74.1	90.5	84.5	93.7
25	90.7	86.3	91.4	93.2	79.6	93.0	87.7	95.7
26	92.4	90.5	94.1	94.9	84.1	94.6	90.7	96.4
27	94.6	91.2	94.0	95.8	84.4	94.8	91.4	96.6
28	96.5	94.4	96.3	97.7	89.1	97.0	94.9	98.1
29	98.1	94.8	97.9	98.2	91.2	97.8	96.0	98.9
30	99.1	96.3	98.4	98.6	94.1	98.7	97.1	99.4

are always 100% while the empirical sizes and powers of T_{3c1p} and T_{3c2p} are always 0%. The problems with T_{SG1}, T_{3c1p} and T_{3c2p} are probably due to the fact that they all use the dimension based kernel width p which is not adaptive to the scale level of the squared distances between the observations. As mentioned in the "two-sample tests" above, these problems may be solved via re-scaling each variable of the colon data set using the pooled standard deviation of the variable.

TABLE 3

Empirical sizes and powers (in %) of all tests for the experiments based on the colon data set ($n = n_1 = n_2$). The results for the T_{SG1}, T_{3c1p}, and T_{3c2p} tests are clearly explained in the text.

n	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
	Empirical sizes							
5	4.5	100	5.3	4.9	0	1.7	0	4.7
6	3.9	100	4.6	3.8	0	2.3	0	5.1
7	5.0	100	5.4	5.6	0	2.9	0	5.7
8	4.3	100	4.8	4.5	0	1.9	0	4.6
9	5.0	100	4.8	6.2	0	3.0	0	5.1
10	5.5	100	4.8	5.7	0	3.7	0	5.7
11	4.1	100	4.3	5.4	0	2.7	0	4.4
12	5.3	100	4.3	5.3	0	3.2	0	5.4
13	5.1	100	4.9	4.6	0	3.7	0	5.3
14	4.6	100	4.1	4.8	0	2.7	0	4.7
15	4.9	100	5.6	6.1	0	3.3	0	5.0
16	5.7	100	5.0	5.8	0	4.4	0	5.6
17	3.9	100	4.7	4.7	0	2.5	0	3.8
18	5.0	100	5.3	4.3	0	3.6	0	5.1
19	5.9	100	5.4	4.3	0	4.5	0	5.9
20	6.4	100	6.5	4.7	0	4.8	0	6.2
	Empirical powers							
5	16.5	100	5.0	5.4	0	4.8	0	17.5
6	24.7	100	4.2	5.7	0	8.2	0	27.6
7	28.8	100	4.8	5.4	0	11.5	0	32.0
8	35.9	100	4.1	5.2	0	17.7	0	37.6
9	43.3	100	4.4	5.1	0	22.0	0	46.4
10	52.1	100	4.5	4.4	0	28.9	0	53.2
11	61.2	100	4.8	4.4	0	38.0	0	63.1
12	71.2	100	4.3	4.4	0	48.8	0	72.3
13	78.4	100	3.0	5.5	0	58.6	0	77.7
14	84.1	100	2.9	3.9	0	67.3	0	84.4
15	88.7	100	1.9	4.5	0	72.8	0	88.9
16	92.5	100	2.3	3.9	0	81.1	0	92.8
17	96.8	100	3.1	3.5	0	89.2	0	96.4
18	98.7	100	2.4	3.1	0	94.2	0	98.4
19	99.5	100	2.2	2.6	0	97.8	0	99.5
20	99.8	100	2.1	3.1	0	98.5	0	99.8

We then repeated the above numerical experiment based on the scaled colon data set. The resulting empirical sizes and powers of the tests are presented in Table 4. We now have the following observations. First, the empirical sizes of $T_{SR}, T_{SG1}, T_{SG2}, T_{SG3}, T_{3c2p}$, and T_{3c2m} are generally comparable and they are all around 5% while the empirical powers of $T_{SR}, T_{SG1}, T_{3c2m}$ and T_{3c2p} are generally comparable. This means that the empirical sizes and powers of T_{SG1}, T_{3c1p} and T_{3c2p} have been improved substantially and to make T_{SG1}, T_{3c1p} and T_{3c2p} work well, it is indeed very important to make the squared distances of the observations and the dimension based kernel width p have the same scale level. Second, the empirical sizes of T_{3c1m} and T_{3c1p} are now generally comparable and so are their empirical powers. Nevertheless, T_{3c1m} and T_{3c1p} are still rather conservative due to the same reason as mentioned in the previous paragraph for the conservativity of T_{3c1m} and due to this conservativity, the empirical pow-

Table 4

Empirical sizes and powers (in %) of all tests for the experiments based on the scaled colon data set ($n = n_1 = n_2$).

n	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
	Empirical sizes							
5	5.6	5.7	5.2	5.1	1.2	1.7	4.8	5.1
6	5.2	5.1	5.5	6.0	2.0	2.9	5.5	6.5
7	4.3	4.0	4.5	4.1	2.0	2.6	4.1	4.7
8	5.7	5.3	6.0	5.7	3.1	3.8	6.0	5.8
9	4.9	4.7	5.2	5.1	2.8	3.4	5.2	5.7
10	5.7	5.3	5.2	4.9	3.1	4.4	5.6	5.8
11	6.4	6.1	5.2	5.7	3.5	4.1	6.2	6.3
12	4.8	4.9	5.3	5.1	2.7	3.0	5.0	5.0
13	6.7	5.5	5.8	5.3	4.0	4.8	5.9	6.2
14	4.8	5.5	5.3	5.2	3.3	3.6	5.3	5.1
15	5.3	5.0	4.9	5.3	2.9	3.3	5.4	5.5
16	4.8	5.0	4.8	4.9	3.1	3.3	4.8	4.6
17	4.7	5.0	4.9	5.3	3.2	3.5	5.3	4.8
18	4.5	5.0	4.6	5.2	3.0	3.1	5.0	4.8
19	4.2	4.6	4.5	4.1	2.6	3.3	4.2	4.6
20	5.7	5.6	5.8	5.7	3.7	4.2	5.8	5.6
	Empirical powers							
5	6.7	8.0	6.5	7.2	1.6	2.0	7.5	7.4
6	6.0	6.9	5.2	6.0	1.9	2.6	7.5	7.7
7	7.9	8.8	6.0	7.3	2.8	3.5	9.7	9.7
8	8.5	8.9	5.3	7.0	2.8	3.6	9.4	9.8
9	9.8	11.0	6.5	7.8	3.1	4.2	11.0	10.6
10	10.7	12.7	8.5	9.7	5.0	5.1	13.9	13.2
11	10.8	13.5	6.8	8.1	4.5	5.3	12.9	12.6
12	11.4	13.8	6.8	8.6	4.7	5.6	14.4	13.7
13	13.5	17.8	6.2	9.2	5.4	5.9	16.5	15.8
14	17.4	18.4	8.6	11.0	7.5	9.0	18.8	17.5
15	17.1	21.2	8.8	11.8	7.5	8.6	22.9	20.0
16	20.6	24.4	8.8	12.1	9.5	10.1	25.2	24.1
17	21.7	26.2	8.5	12.5	9.3	10.4	26.2	23.3
18	26.4	31.5	8.4	13.7	11.9	12.2	30.7	26.1
19	30.2	37.6	9.1	16.6	15.3	14.8	36.3	32.2
20	34.7	41.8	10.8	18.1	17.2	18.3	40.8	36.8

ers of T_{3c1m} and T_{3c1p} are generally smaller than those of $T_{SR}, T_{SG1}, T_{3c2p}$ and T_{3c2m}. Third, both T_{SG2} and T_{SG3} have good size control but their empirical powers are still much smaller than those of $T_{SR}, T_{SG1}, T_{3c2p}$ and T_{3c2m}. This again shows that T_{SG2} and T_{SG3} do not perform well in the colon data set based experiments.

5.4. Comparison with the Gretton et al. tests

To address a concern from the reviewer, some simulation studies are conducted to compare our tests T_{3c1q} and T_{3c2q} against two of the Gretton et al. tests, namely the test (denoted as T^q_{spec}) based on the null distribution estimate using the empirical Gram matrix spectrum [13, Section 3.2], and the test (denoted as T^q_{pear}) based on the Pearson curve approximation [11, Section 4], for $q = p, m$.

Table 5

Empirical sizes (under Models 1 and 2) and powers (under Models $6[\mathcal{N}]$ *and* $6[t_4]$*) (in %) of* T^p_{spec}*,* T^p_{pear}*,* T_{3c1p}*,* T_{3c2p}*,* T^m_{spec}*,* T^m_{pear}*,* T_{3c1m} *and* T_{3c2m} *for* $\mathbf{n} = (30, 30)$*.*

Model	p	ρ	T^p_{spec}	T^p_{pear}	T_{3c1p}	T_{3c2p}	T^m_{spec}	T^m_{pear}	T_{3c1m}	T_{3c2m}
1	10	0.2	3.6	2.9	3.3	4.7	3.6	3.4	3.7	4.6
		0.5	4.1	3.9	4.1	4.9	4.0	3.5	3.6	4.5
		0.8	4.4	3.8	4.0	4.9	4.9	4.3	5.0	5.3
	100	0.2	1.9	1.0	1.6	5.0	3.3	2.2	2.8	5.1
		0.5	3.7	3.4	3.5	4.4	3.9	3.4	3.7	4.3
		0.8	4.9	3.8	4.3	4.9	4.6	4.1	4.7	5.0
	500	0.2	2.9	1.1	2.3	4.4	3.0	1.6	2.3	4.5
		0.5	4.0	3.1	4.1	4.7	3.9	3.2	3.4	4.1
		0.8	4.3	3.7	3.7	4.1	3.9	3.4	3.8	4.0
2	10	0.2	2.8	2.2	2.5	4.4	3.1	2.2	2.7	4.1
		0.5	3.8	3.5	3.4	4.9	4.4	3.7	4.1	5.2
		0.8	3.9	3.3	3.6	4.6	4.3	3.6	4.4	4.9
	100	0.2	0.3	0.1	0.3	4.2	3.4	1.9	2.7	5.1
		0.5	4.1	2.9	3.6	5.2	4.1	3.3	4.0	5.0
		0.8	4.7	4.5	4.4	5.3	4.3	3.4	4.1	4.4
	500	0.2	0.2	0.1	0.3	4.2	3.5	2.4	3.1	5.7
		0.5	2.8	1.9	2.5	4.8	4.1	3.3	4.0	4.4
		0.8	4.1	3.4	3.5	4.8	4.3	3.8	4.4	4.6
$6[\mathcal{N}]$	10	0.2	4.7	2.8	4.5	7.5	4.1	3.3	4.2	6.1
		0.5	12.1	10.4	11.0	15.0	9.2	8.1	7.8	9.8
		0.8	8.5	7.7	8.0	9.3	7.2	6.4	6.5	7.5
	100	0.2	2.8	1.4	2.4	18.9	3.5	1.7	3.0	9.8
		0.5	67.4	47.7	63.0	90.7	31.1	22.0	27.5	42.4
		0.8	99.4	98.1	99.2	99.7	76.2	64.9	72.0	83.7
	500	0.2	2.2	0.3	1.9	28.4	3.4	0.9	2.6	12.7
		0.5	85.9	63.0	81.6	99.9	38.9	25.2	32.9	57.6
		0.8	100.0	100.0	100.0	100.0	99.0	95.6	99.3	100.0
$6[t_4]$	10	0.2	3.4	1.8	3.0	7.9	4.5	3.8	4.1	6.6
		0.5	13.3	10.3	12.4	19.6	8.0	6.6	7.5	9.4
		0.8	16.4	12.6	14.4	19.3	8.9	7.3	8.3	8.7
	100	0.2	0.4	0.1	0.5	37.4	3.6	1.5	3.1	10.7
		0.5	86.8	64.3	86.6	99.7	30.5	22.6	27.7	43.4
		0.8	100.0	100.0	100.0	100.0	75.7	67.0	71.4	83.4
	500	0.2	0.2	0.0	0.1	71.5	2.6	1.1	2.5	12.8
		0.5	98.1	80.9	98.8	100.0	37.7	24.9	32.5	58.0
		0.8	100.0	100.0	100.0	100.0	98.3	94.9	98.09	100.0

We do not consider the test based on the Gamma approximation, since *"Gamma approximation to the null distribution, which has a smaller computational cost, is generally less accurate"* [10, p. 738].

Table 5 presents the empirical sizes (under Models 1 and 2 in Section 5.2) and powers (under Models $6[\mathcal{N}]$ and $6[t_4]$) of T^p_{spec}, T^p_{pear}, T_{3c1p}, T_{3c2p}, T^m_{spec}, T^m_{pear}, T_{3c1m} and T_{3c2m} for $\mathbf{n} = (30, 30)$. It is seen that in terms of size control and power, for $q = p, m$, T_{3c2q} substantially outperforms T^q_{spec}, T^q_{pear} and T_{3c1q} which are generally comparable and are all quite conservative. As mentioned in Remark 4.2, the conservativity of T_{3c1p} and T_{3c1m} is due to the fact that the null distributions of T_{3c1p} and T_{3c1m} are estimated using the estimated eigenvalues of the Gram matrix whose asymptotic properties may not take effect for small and

Table 6

Empirical sizes and powers (in %) of T^p_{spec}, T^p_{pear}, T_{3c1p}, T_{3c2p}, T^m_{spec}, T^m_{pear}, T_{3c1m} and T_{3c2m} for the experiments based on the glass data set ($n = n_1 = n_2$).

n	T^p_{spec}	T^p_{pear}	T_{3c1p}	T_{3c2p}	T^p_{spec}	T^p_{pear}	T_{3c1p}	T_{3c2p}
	Empirical sizes				Empirical powers			
11	6.2	4.9	5.5	5.8	13.0	12.0	13.5	24.8
12	3.3	3.2	3.1	4.3	19.1	17.6	17.4	33.2
13	4.7	4.3	4.8	4.9	21.0	20.1	19.2	35.1
14	5.2	4.0	5.0	5.1	25.2	24.1	24.8	41.2
15	5.6	5.3	5.6	5.9	29.4	27.9	27.5	45.2
16	4.9	4.1	4.5	5.0	34.1	33.7	32.6	49.8
17	4.9	4.8	4.7	4.9	37.9	38.4	36.2	55.0
18	4.7	3.7	4.7	4.7	41.9	42.7	39.5	56.8
19	5.2	4.5	4.8	5.0	48.5	48.1	46.5	63.4
20	3.8	3.6	3.4	4.2	54.5	55.4	52.7	68.6
21	5.2	4.2	5.3	5.3	61.4	62.2	61.3	74.3
22	5.5	5.0	5.8	5.9	64.0	64.7	62.4	76.2
23	5.7	4.6	5.7	5.9	68.5	69.2	68.2	81.4
24	5.7	5.0	5.2	5.4	74.6	77.2	74.5	86.4
25	5.4	4.8	5.2	5.3	78.8	79.7	78.1	87.4
26	5.0	4.1	5.1	5.1	83.3	82.5	82.5	89.7
27	4.2	3.6	4.1	4.3	86.1	85.3	85.9	92.8
28	5.2	4.6	5.2	5.2	88.0	88.3	87.9	93.5
29	6.8	6.2	6.4	5.6	92.5	91.8	91.9	95.4
30	5.0	4.7	5.0	5.1	92.8	92.8	92.2	95.8
n	T^m_{spec}	T^m_{pear}	T_{3c1m}	T_{3c2m}	T^m_{spec}	T^m_{pear}	T_{3c1m}	T_{3c2m}
	Empirical sizes				Empirical powers			
11	5.4	5.0	4.9	5.7	29.2	23.1	28.0	38.4
12	3.3	3.1	3.1	4.2	34.1	30.0	32.8	43.8
13	4.5	4.1	4.6	4.8	38.7	33.1	37.8	48.8
14	4.2	4.0	3.9	4.6	45.4	41.1	44.3	54.6
15	5.7	4.2	5.3	5.7	48.3	44.5	46.8	57.4
16	4.5	4.3	4.0	4.5	57.6	51.6	54.9	65.7
17	3.9	3.8	3.9	4.2	62.8	57.5	61.3	70.1
18	3.6	3.6	3.5	4.3	64.6	60.0	62.3	72.7
19	5.2	4.6	5.0	5.1	71.0	65.7	69.0	76.6
20	3.7	3.0	2.9	4.2	75.1	71.0	75.4	82.5
21	5.8	5.1	5.8	5.9	80.3	77.1	80.0	85.5
22	5.9	5.1	6.2	5.3	81.6	77.3	81.2	87.1
23	5.6	5.1	5.2	5.7	86.8	85.0	86.5	90.4
24	5.1	4.8	4.5	5.0	89.0	87.1	89.1	92.9
25	5.3	4.8	5.2	5.3	92.1	89.3	91.6	93.8
26	5.0	4.3	5.0	5.2	94.9	92.9	94.3	96.5
27	4.8	3.7	3.9	4.4	95.7	94.9	95.8	97.4
28	5.5	4.7	5.4	5.7	96.2	94.9	96.0	97.3
29	6.3	6.1	6.0	5.6	98.6	97.7	98.5	99.1
30	5.3	4.7	4.8	5.0	98.7	97.9	98.7	98.9

moderate sample sizes. Obviously, this reason is also applicable to explain the conservativity of T^p_{spec} and T^m_{spec}; see also [13, p. 7]. Tables 6, 7 and 8 present the empirical sizes and powers of T^p_{spec}, T^p_{pear}, T_{3c1p}, T_{3c2p}, T^m_{spec}, T^m_{pear}, T_{3c1m} and T_{3c2m} for the experiments based on the glass data set, the colon data set, and the scaled colon data set respectively as described in Section 5.3. For the colon data

TABLE 7

Empirical sizes and powers (in %) of T^m_{spec}, T^m_{pear}, T_{3c1m} and T_{3c2m} for the experiments based on the colon data set ($n = n_1 = n_2$).

	Empirical sizes				Empirical powers			
n	T^m_{spec}	T^m_{pear}	T_{3c1m}	T_{3c2m}	T^m_{spec}	T^m_{pear}	T_{3c1m}	T_{3c2m}
5	2.8	0.6	2.8	5.6	6.0	1.0	5.0	19.3
6	2.3	0.4	1.9	4.8	8.4	2.1	7.7	25.4
7	3.4	1.7	2.8	4.9	12.4	5.7	11.4	31.9
8	3.2	1.9	2.8	5.4	18.4	8.5	16.2	38.2
9	4.2	2.3	3.6	5.9	25.2	13.4	21.8	46.8
10	3.3	2.5	2.7	5.2	31.7	20.1	30.3	52.7
11	3.6	2.1	3.0	5.2	42.8	28.8	41.4	65.3
12	3.9	2.2	3.5	5.7	51.9	36.8	48.9	73.2
13	3.3	2.3	3.1	4.5	57.7	44.9	55.0	76.8
14	2.9	2.1	2.9	4.4	67.2	55.7	64.4	83.7
15	4.2	2.9	3.8	4.9	77.8	66.6	75.8	90.4
16	3.5	2.4	3.1	4.9	84.5	73.5	82.8	93.5
17	3.0	2.7	3.1	4.4	90.4	82.1	89.2	96.0
18	4.1	3.6	3.7	5.2	93.6	87.2	92.6	98.7
19	4.9	3.8	4.6	5.2	96.8	92.9	96.8	99.7
20	3.6	3.3	3.1	4.2	99.0	97.1	99.3	99.9

set, the empirical sizes and powers of T^p_{spec}, T^p_{pear}, T_{3c1p} and T_{3c2p} were all equal to zero, and thus they are omitted in Table 7 (see Section 5.3 for explanation). It is seen that in terms of size control and power, T_{3c2q} outperforms T^q_{spec}, T^q_{pear} and T_{3c1q} generally, for $q = p, m$. In addition, in terms of size control, we can also observe the conservativity of T^p_{pear} and T^m_{pear} in some cases in Table 6 and the conservativity of T^p_{spec}, T^m_{spec}, T^p_{pear}, T^m_{pear}, T_{3c1p} and T_{3c1m} in almost all cases in Tables 7-8. The above results show that in terms of size control and power, our tests T_{3c1p}, T_{3c2p}, T_{3c1m} and T_{3c2m} outperform or perform not worse than the Gretton et al. tests under consideration.

5.5. Computational time comparison of the tests

To address a question from the reviewer, we display the computational time (in minutes) of some considered tests in Figures 2 and 3. The experiments were performed for one physical computer with 4 cores, 8 GB RAM, Ubuntu 18.04 64-bit, R 4.1.1. In Figure 2, we present the fastest tests, i.e., T_{3c1m}, T_{3c2m} and T_{SR}, with T_{3c1m} the fastest, followed T_{3c2m}, which is significantly faster than T_{SR}. It is seen that compared with T_{3c1m}, the U-statistics used in estimating the first three cumulants in T_{3c2m} does add some amount of extra computation. In Figure 3, we present the computational time of T_{SG1}, T_{3c1m}, T_{3c2m} and T_{SR}. It is seen that T_{SG1} is much more time-consuming than T_{3c1m}, T_{3c2m}, and T_{SR}. Moreover, from Table 9, we obtain that T_{SG2} and T_{SG3} are much more time-consuming than $T_{SG1}, T_{SR}, T_{3c1m}$ and T_{3c2m}. Note that the implementations of T_{SG1}, T_{SG2} and T_{SG3} are based on the C++ implementation of the most time-consuming parts of their constructions.

Table 8

Empirical sizes and powers (in %) of T^p_{spec}, T^p_{pear}, T_{3c1p}, T_{3c2p}, T^m_{spec}, T^m_{pear}, T_{3c1m} and T_{3c2m} for the experiments based on the scaled colon data set ($n = n_1 = n_2$).

n	T^p_{spec}	T^p_{pear}	T_{3c1p}	T_{3c2p}	T^p_{spec}	T^p_{pear}	T_{3c1p}	T_{3c2p}
	Empirical sizes				Empirical powers			
5	1.3	0.2	1.1	4.8	1.1	0.0	0.8	5.8
6	2.3	0.4	2.0	4.8	1.7	0.1	1.5	7.2
7	3.3	1.8	3.2	5.6	2.0	0.6	1.9	8.9
8	3.0	1.8	2.6	5.8	2.5	1.0	2.0	8.0
9	3.1	2.2	3.1	5.4	3.1	1.3	2.5	8.8
10	3.6	1.6	2.8	5.2	3.3	1.8	2.5	10.6
11	2.6	1.6	2.5	4.7	3.3	1.2	3.0	12.0
12	3.2	1.9	2.5	4.8	5.4	2.4	4.4	13.0
13	3.5	1.9	3.0	5.3	6.5	3.7	5.3	17.3
14	3.0	2.0	3.2	5.2	7.5	4.3	6.1	16.9
15	3.5	2.4	3.0	5.1	10.0	6.4	7.3	21.0
16	3.0	2.3	2.7	4.6	9.3	6.1	7.5	20.5
17	3.7	2.5	2.9	4.5	10.4	6.7	7.6	23.7
18	3.0	2.4	2.6	4.5	15.7	9.6	10.7	30.8
19	3.8	2.8	3.4	4.6	17.7	12.0	12.1	35.3
20	4.6	3.1	4.5	5.2	19.3	14.4	14.8	39.3
n	T^m_{spec}	T^m_{pear}	T_{3c1m}	T_{3c2m}	T^m_{spec}	T^m_{pear}	T_{3c1m}	T_{3c2m}
	Empirical sizes				Empirical powers			
5	1.7	0.3	1.6	4.3	1.4	0.2	1.2	5.6
6	2.3	0.8	2.3	5.2	2.5	0.4	2.1	6.1
7	4.0	2.2	3.7	5.8	3.7	0.9	3.0	8.6
8	3.3	2.0	3.1	5.9	3.2	1.6	2.5	7.7
9	3.2	2.4	3.1	4.8	3.1	1.6	2.8	8.5
10	4.0	2.0	3.4	5.1	4.1	2.4	3.5	8.7
11	3.2	2.0	3.0	4.7	4.5	2.3	3.7	10.4
12	3.3	2.3	3.0	4.6	5.7	3.9	4.8	11.4
13	3.9	2.7	3.4	5.4	6.4	5.0	6.1	13.5
14	3.3	2.3	3.2	5.3	8.1	5.5	6.7	15.2
15	3.8	2.7	3.7	5.4	9.3	8.3	7.4	18.6
16	3.5	3.0	3.0	4.6	9.9	7.7	7.5	16.9
17	3.6	2.9	3.2	5.1	10.8	8.1	7.3	19.5
18	3.4	2.9	3.0	3.7	15.4	11.7	10.9	26.2
19	4.2	3.4	3.9	4.7	17.2	14.6	12.5	29.6
20	4.7	3.4	4.1	5.7	18.9	16.2	14.6	32.2

In conclusion, our tests T_{3c1p},T_{3c2p}, T_{3c1m}, and T_{3c2m} are faster to compute than the existing tests under consideration.

5.6. Testing procedure recommendation

Based on the above numerical experiments using artificial and real data, we can recommend T_{3c2p} and T_{3c2m} for practical applications when the sample sizes are small or moderate. They keep the type I error level very well and have high powers generally. They are fast to compute and outperform several existing tests. The squared median-distance based kernel width can be used generally since it is scale-invariant in the sense that the associated test statistic does not change if

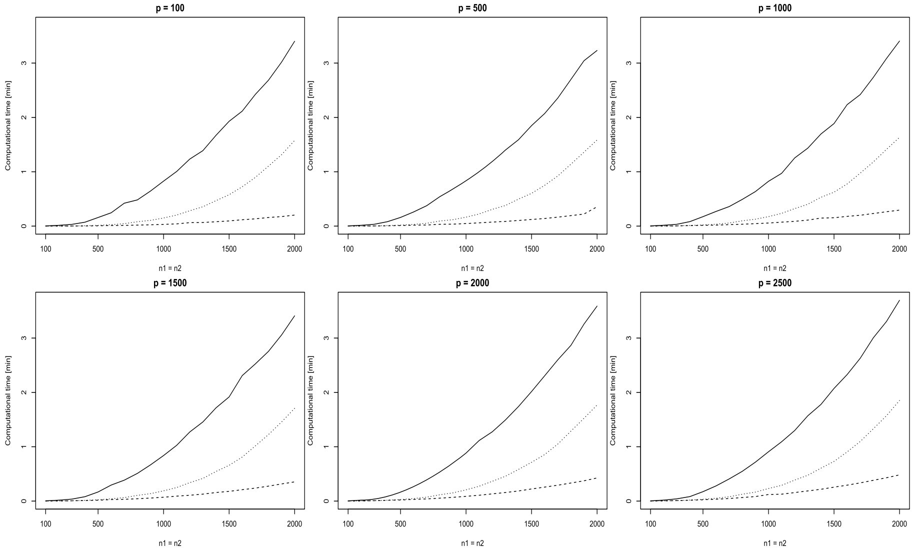

Fig 2: Computational time (in minutes) of T_{SR} (solid line), T_{3c1m} (dashed line), and T_{3c2m} (dotted line) for different sample sizes ($n_1 = n_2$) and dimensions $p = 100, 500, 1000, 1500, 2000, 2500$.

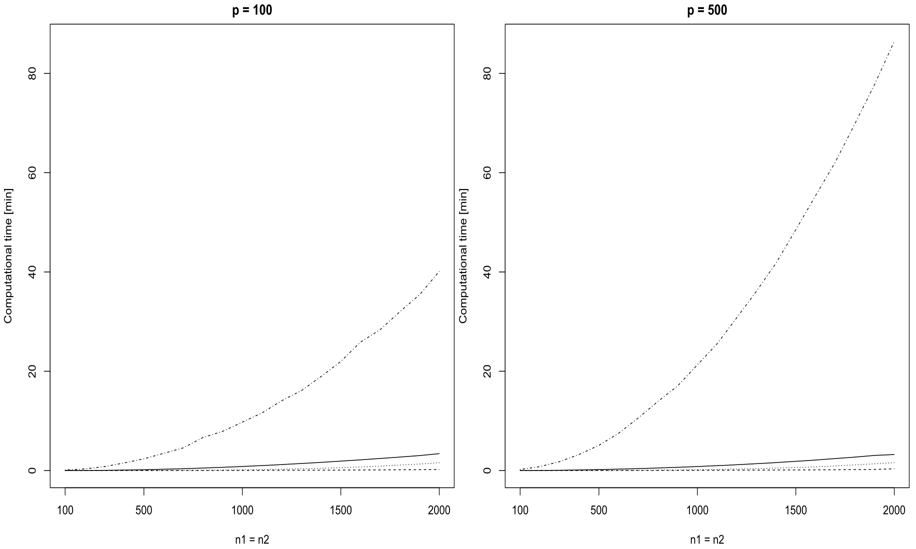

Fig 3: Computational time (in minutes) of T_{SG1} (dash-dotted line), T_{SR} (solid line), T_{3c1m} (dashed line), and T_{3c2m} (dotted line) for different sample sizes ($n_1 = n_2$) and dimensions $p = 100, 500$.

Table 9
Computational time (in minutes) of T_{SR}, T_{SG1}, T_{SG2}, T_{SG3}, T_{3c1m}, and T_{3c2m} for different sample sizes ($n_1 = n_2$) and dimension $p = 100$.

$n_1 = n_2$	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1m}	T_{3c2m}
100	0.00	0.10	1.46	1.40	0.00	0.00
200	0.01	0.34	5.84	5.53	0.00	0.00
300	0.03	0.79	13.24	12.70	0.00	0.00
400	0.07	1.54	23.81	22.90	0.00	0.01
500	0.16	2.37	37.50	35.95	0.01	0.01
600	0.25	3.46	53.81	51.80	0.01	0.03
700	0.42	4.57	72.75	71.20	0.01	0.04
800	0.48	6.65	95.68	94.01	0.02	0.08
900	0.64	7.90	122.63	116.43	0.02	0.11
1000	0.83	9.75	152.26	141.12	0.03	0.15
1100	1.01	11.61	183.06	170.18	0.04	0.20
1200	1.23	14.07	221.46	200.87	0.06	0.28
1300	1.39	16.10	252.16	237.65	0.07	0.36
1400	1.67	18.94	290.86	272.37	0.08	0.47

each observation is multiplied by a nonzero constant, while the dimension based kernel width p in T_{3c1p} and T_{3c2p} should be used with caution, since it is not scale-invariant and not always applicable.

6. Concluding remarks

We have considered the two-sample problem in a separable metric space, which includes many data types. For this problem, we have proposed two new tests based on the maximum mean discrepancy. In contrast to the existing tests, we allowed different sample sizes, which is more realistic, in both theoretical and practical issues. In addition, unlike the existing tests, which are implemented by permutation, the two new tests were based on the three-cumulant matched χ^2-approximation. The first test, T_{3c1}, used the cumulants of the asymptotic null distribution of the test statistic, which resulted in its good behavior for large sample sizes, but for small and moderate sample sizes, conservativity appeared and resulted in less power. On the other hand, the second test, T_{3c2}, used the cumulants of the null distribution, which resulted in an accurate and fast test for small and moderate sample sizes. In the numerical experiments presented in Section 5, we considered a squared median distance based kernel width and a dimension-based kernel width. The former kernel width is scale-invariant while the latter one is not. The resulting tests are denoted as T_{3c1m}, T_{3c2m} and T_{3c1p}, T_{3c2p} respectively. In the numerical experiments based on artificial data, T_{3c1p} and T_{3c2p} outperform T_{3c1m} and T_{3c2m} substantially while in the numerical experiments based on the two real data sets, T_{3c1p} and T_{3c2p} can perform much worse than T_{3c1m} and T_{3c2m} unless the data are properly re-scaled so that the squared distances of the transformed observations and the dimension based kernel width p are at the same scale level. This means that we should use the dimension-based kernel width with caution. It also says that it is important to select a good kernel width for the proposed new tests. Further research in this direction is interesting and warranted. Of course, the performance of the new tests needs to be further evaluated on additional artificial and real data sets. In particular, we can examine the behavior of the new tests for other types of data including strings and graphs. This may constitute a direction for our future research.

Appendix A: Proofs

The equalities (8) of the form

$$\begin{array}{rl} S_{\alpha\alpha} = & \tilde{S}_{\alpha\alpha} + 2\langle \bar{x}_\alpha - \mu_\alpha, \mu_\alpha\rangle + \|\mu_\alpha\|^2, \\ S_{12} = & \tilde{S}_{12} + \langle \bar{x}_1 - \mu_1, \mu_2\rangle + \langle \bar{x}_2 - \mu_2, \mu_1\rangle + \langle \mu_1, \mu_2\rangle, \end{array}$$

where $\alpha = 1, 2$, follow from

$$\begin{array}{rl} S_{\alpha\alpha} = & \frac{2}{n_\alpha(n_\alpha-1)} \sum_{1\le i<j\le n_\alpha} \langle (x_{\alpha i} - \mu_\alpha) + \mu_\alpha, (x_{\alpha j} - \mu_\alpha) + \mu_\alpha\rangle \\ = & \frac{2}{n_\alpha(n_\alpha-1)} \sum_{1\le i<j\le n_\alpha} \big(\langle x_{\alpha i} - \mu_\alpha, x_{\alpha j} - \mu_\alpha\rangle + \langle x_{\alpha i} - \mu_\alpha, \mu_\alpha\rangle \\ & +\langle \mu_\alpha, x_{\alpha j} - \mu_\alpha\rangle + \|\mu_\alpha\|^2\big) \\ = & \tilde{S}_{\alpha\alpha} + 2\langle \bar{x}_\alpha - \mu_\alpha, \mu_\alpha\rangle + \|\mu_\alpha\|^2, \\ S_{12} = & \langle (\bar{x}_1 - \mu_1) + \mu_1, (\bar{x}_2 - \mu_2) + \mu_2\rangle \\ = & \langle \bar{x}_1 - \mu_1, \bar{x}_2 - \mu_2\rangle + \langle \bar{x}_1 - \mu_1, \mu_2\rangle + \langle \mu_1, \bar{x}_2 - \mu_2\rangle + \langle \mu_1, \mu_2\rangle \\ = & \tilde{S}_{12} + \langle \bar{x}_1 - \mu_1, \mu_2\rangle + \langle \bar{x}_2 - \mu_2, \mu_1\rangle + \langle \mu_1, \mu_2\rangle, \end{array}$$

where $\alpha = 1, 2$ and

$$\begin{array}{rl} \tilde{S}_{\alpha\alpha} = & \frac{2}{n_\alpha(n_\alpha-1)} \sum_{1\le i<j\le n_\alpha} \langle x_{\alpha i} - \mu_\alpha, x_{\alpha j} - \mu_\alpha\rangle \\ = & \frac{2}{n_\alpha(n_\alpha-1)} \sum_{1\le i<j\le n_\alpha} \tilde{K}(y_{\alpha i}, y_{\alpha j}), \\ \tilde{S}_{12} = & \langle \bar{x}_1 - \mu_1, \bar{x}_2 - \mu_2\rangle = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \tilde{K}(y_{1i}, y_{2j}). \end{array}$$

Lemma A.1. *Let $\alpha, \beta = 1, 2$ and $\alpha \neq \beta$. We have $E(\tilde{S}_{\alpha\alpha}) = 0$, $E(\tilde{S}_{\alpha\beta}) = 0$, and $\operatorname{cov}(\tilde{S}_{\alpha\alpha}, \tilde{S}_{\alpha\beta}) = 0$. Further,*

$$\operatorname{var}(\tilde{S}_{\alpha\alpha}) = \frac{2}{n_\alpha(n_\alpha - 1)} E(\tilde{K}^2(y_{\alpha 1}, y_{\alpha 2})),\ \operatorname{var}(\tilde{S}_{\alpha\beta}) = \frac{1}{n_\alpha n_\beta} E(\tilde{K}^2(y_{\alpha 1}, y_{\beta 1})).$$

Proof of Lemma A.1. We have the following useful properties. By (7), when $y' = y$, we have

$$E_y(\tilde{K}(y, y)) = E_y(K(y, y)) - E_{z,z'}(K(z, z')) > 0, \tag{28}$$

where z and z' are independent copies of y, and when y and y' are independent, we have

$$E_y(\tilde{K}(y, y')) = E_{y'}(\tilde{K}(y, y')) = E_{y,y'}(\tilde{K}(y, y')) = 0. \tag{29}$$

By (28), (29) and (9), we have

$$\begin{aligned}
E(\tilde{S}_{\alpha\alpha}) &= \tfrac{2}{n_\alpha(n_\alpha-1)}\textstyle\sum_{1\le i<j\le n_\alpha} E(\tilde{K}(y_{\alpha i}, y_{\alpha j})) = 0,\\
E(\tilde{S}_{\alpha\beta}) &= \tfrac{1}{n_\alpha n_\beta}\textstyle\sum_{i=1}^{n_\alpha}\sum_{j=1}^{n_\beta} E(\tilde{K}(y_{\alpha i}, y_{\beta j})) = 0,\\
\operatorname{cov}\left(\tilde{S}_{\alpha\alpha}, \tilde{S}_{\alpha\beta}\right) &= E(\tilde{S}_{\alpha\alpha}\tilde{S}_{\alpha\beta}) = 0,\\
\operatorname{var}(\tilde{S}_{\alpha\alpha}) &= \tfrac{4}{n_\alpha^2(n_\alpha-1)^2}\textstyle\sum_{1\le i<j\le n_\alpha} \operatorname{var}(\tilde{K}(y_{\alpha i}, y_{\alpha j}))\\
&= \tfrac{2}{n_\alpha(n_\alpha-1)} E(\tilde{K}^2(y_{\alpha 1}, y_{\alpha 2})),\\
\operatorname{var}(\tilde{S}_{\alpha\beta}) &= \tfrac{1}{n_\alpha^2 n_\beta^2}\textstyle\sum_{i=1}^{n_\alpha}\sum_{j=1}^{n_\beta} \operatorname{var}(\tilde{K}(y_{\alpha i}, y_{\beta j}))\\
&= \tfrac{1}{n_\alpha n_\beta} E(\tilde{K}^2(y_{\alpha 1}, y_{\beta 1})).
\end{aligned}$$

□

Theorem 3.1. *Under Assumptions 1–3, as $n \to \infty$, we have $\tilde{T}_n \xrightarrow{d} \tilde{T}$, where*

$$\tilde{T} \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r (A_r - 1), \;\; A_r \overset{i.i.d.}{\sim} \chi_1^2.$$

Proof of Theorem 3.1. Under Assumption 1, we have $P_1 = P_2 = P$. Let $y, y' \overset{\text{i.i.d.}}{\sim} P$. Under Assumption 3, we have the Mercer's expansion (12). By (29) and (13), we have

$$\lambda_r E_y(\psi_r(y)) = \int_{\mathcal{Y}} E_y(\tilde{K}(y, y'))\psi_r(y')P(dy') = 0.$$

This, together with (13), implies that $E(\psi_r(y)) = 0$ whenever $\lambda_r \neq 0$ and $\operatorname{var}(\psi_r(y)) = \int_{\mathcal{Y}} \psi_r^2(y)P(dy) = 1$. Set $z_{r,\alpha i} = \psi_r(y_{\alpha i})$ $(i = 1, \dots, n_\alpha; \alpha = 1, 2)$. Under Assumption 1, we have $y_{\alpha i} \overset{\text{i.i.d.}}{\sim} P$. It follows that for a fixed $r = 1, 2, \dots$, $z_{r,\alpha i}$ $(i = 1, \dots, n_\alpha; \alpha = 1, 2)$ are i.i.d. with mean 0 and variance 1. For different r, $z_{r,\alpha i}$ $(i = 1, \dots, n_\alpha; \alpha = 1, 2)$ are uncorrelated. Then by (9) and (12), we have

$$\begin{aligned}
\tilde{S}_{12} &= \tfrac{1}{n_1 n_2}\textstyle\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(\sum_{r=1}^{\infty}\lambda_r z_{r,1i} z_{r,2j}\right)\\
&= \textstyle\sum_{r=1}^{\infty}\lambda_r \bar{z}_{r,1}\bar{z}_{r,2},\\
\tilde{S}_{\alpha\alpha} &= \tfrac{2}{n_\alpha(n_\alpha-1)}\textstyle\sum_{1\le i<j\le n_\alpha}\left(\sum_{r=1}^{\infty}\lambda_r z_{r,\alpha i} z_{r,\alpha j}\right)\\
&= \textstyle\sum_{r=1}^{\infty}\lambda_r\left(\tfrac{2}{n_\alpha(n_\alpha-1)}\sum_{1\le i<j\le n_\alpha} z_{r,\alpha i} z_{r,\alpha j}\right)\\
&= \textstyle\sum_{r=1}^{\infty}\lambda_r(\bar{z}_{r,\alpha}^2 - 1/n_\alpha)(1 + o_p(1)),
\end{aligned}$$

where $\bar{z}_{r,\alpha} = n_\alpha^{-1}\sum_{i=1}^{n_\alpha} z_{r,\alpha i}$ $(\alpha = 1, 2; r = 1, 2, \dots)$. In the last equality, we have used the following facts

$$\begin{aligned}
\frac{2}{n_\alpha(n_\alpha-1)}\sum_{1\le i<j\le n_\alpha} z_{r,\alpha i} z_{r,\alpha j} =& \frac{1}{n_\alpha^2}\left(\sum_{i=1}^{n_\alpha}\sum_{j=1}^{n_\alpha} z_{r,\alpha i} z_{r,\alpha j} - \sum_{i=1}^{n_\alpha} z_{r,\alpha i}^2\right)(1 + o(1))\\
=& \left(\bar{z}_{r,\alpha}^2 - \frac{1}{n_\alpha}\right)(1 + o_p(1)),
\end{aligned}$$

since as $n_\alpha \to \infty$, we have $n_\alpha^{-1}\sum_{i=1}^{n_\alpha} z_{r,\alpha i}^2 \xrightarrow{p} E(z_{r,\alpha 1}^2) = 1$. By (11), we then have

$$\begin{aligned}\tilde{T}_n &= \textstyle\sum_{r=1}^\infty \lambda_r \frac{n_1 n_2}{n}\big((\bar{z}_{r,1}^2 - 1/n_1) + (\bar{z}_{r,2}^2 - 1/n_2) - 2\bar{z}_{r,1}\bar{z}_{r,2}\big)(1+o_p(1))\\ &= \textstyle\sum_{r=1}^\infty \lambda_r \left(\frac{n_1 n_2}{n}(\bar{z}_{r,1} - \bar{z}_{r,2})^2 - 1\right)(1+o_p(1))\\ &= \textstyle\sum_{r=1}^\infty \lambda_r (A_{n,r} - 1)(1+o_p(1)),\end{aligned}$$

where $A_{n,r} = w_{n,r}^2$ with $w_{n,r} = (n_1 n_2/n)^{1/2}(\bar{z}_{r,1} - \bar{z}_{r,2})$ $(r = 1, 2, \dots)$, which are uncorrelated. For any given $r = 1, 2, \dots$, by the central limit theorem, under Assumptions 2 and 3, as $n \to \infty$, we have $n_\alpha^{1/2}\bar{z}_{r,\alpha} \xrightarrow{d} \mathcal{N}(0,1)$ $(\alpha = 1, 2)$ and $\bar{z}_{r,1}$ and $\bar{z}_{r,2}$ are independent. It follows that under Assumptions 2 and 3, as $n \to \infty$, $w_{n,r} \xrightarrow{d} w_r \sim \mathcal{N}(0,1)$, and hence

$$A_{n,r} = w_{n,r}^2 \xrightarrow{d} A_r = w_r^2 \overset{\text{i.i.d.}}{\sim} \chi_1^2 \ (r = 1, 2, \dots). \tag{30}$$

It follows that as $n \to \infty$, $E(A_{n,r}) = 1 + o(1)$ and $\text{var}(A_{n,r}) = 2 + o(1)$.

Let $\varphi_X(t) = E(e^{itX})$ denote the characteristic function of a random variable X. Set $\tilde{T}_n^{(q)} = \sum_{r=1}^q \lambda_r (A_{n,r} - 1)$. Then $|\varphi_{\tilde{T}_n}(t) - \varphi_{\tilde{T}_n^{(q)}}(t)| \le |t|\big(E(\tilde{T}_n - \tilde{T}_n^{(q)})^2\big)^{1/2}$. Therefore, as $n \to \infty$, we have

$$\begin{aligned}E(\tilde{T}_n - \tilde{T}_n^{(q)})^2 = &\ E\left(\textstyle\sum_{r=q+1}^\infty \lambda_r (A_{n,r} - 1)\right)^2 (1+o_p(1))\\ = &\ \text{var}\left(\textstyle\sum_{r=q+1}^\infty \lambda_r A_{n,r}\right)(1+o_p(1))\\ \le &\ \left(\textstyle\sum_{r=q+1}^\infty \text{var}^{1/2}(\lambda_r A_{n,r})\right)^2 (1+o_p(1))\\ = &\ 2\left(\textstyle\sum_{r=q+1}^\infty \lambda_r\right)^2 (1+o_p(1)).\end{aligned}$$

It follows that

$$|\varphi_{\tilde{T}_n}(t) - \varphi_{\tilde{T}_n^{(q)}}(t)| \le |t|\sqrt{2}\left(\sum_{r=q+1}^\infty \lambda_r\right)(1+o(1)). \tag{31}$$

Let t be fixed. Under Assumption 3 and (14), as $q \to \infty$, we have $\sum_{r=q+1}^\infty \lambda_r \to 0$. Thus, by (31), for any given $\epsilon > 0$, there exist N_1 and Q_1, depending on $|t|$ and ϵ, such that as $n > N_1$ and $q > Q_1$, we have

$$|\varphi_{\tilde{T}_n}(t) - \varphi_{\tilde{T}_n^{(q)}}(t)| \le \epsilon. \tag{32}$$

For any fixed $q > Q_1$, by (30), as $n \to \infty$, we have $\tilde{T}_n^{(q)} \xrightarrow{d} \tilde{T}^{(q)} \overset{d}{=} \sum_{r=1}^q \lambda_r (A_r - 1)$, $A_r \overset{\text{i.i.d.}}{\sim} \chi_1^2$. Thus, there exists N_2, depending on q and ϵ such that as $n > N_2$, we have

$$|\varphi_{\tilde{T}_n^{(q)}}(t) - \varphi_{\tilde{T}^{(q)}}(t)| \le \epsilon. \tag{33}$$

Recall that $\tilde{T} = \sum_{r=1}^\infty \lambda_r (A_r - 1)$, $A_r \overset{\text{i.i.d.}}{\sim} \chi_1^2$. Along the same lines as those for proving (32), we can show that there exist Q_2, depending on $|t|$ and ϵ, such that as $q > Q_2$, we have

$$|\varphi_{\tilde{T}^{(q)}}(t) - \varphi_{\tilde{T}}(t)| \le \epsilon. \tag{34}$$

It follows from (32)–(34) that for any $n \geq \max(N_1, N_2)$ and $q \geq \max(Q_1, Q_2)$, we have

$$\left|\varphi_{\tilde{T}_n}(t) - \varphi_{\tilde{T}}(t)\right| \leq \left|\varphi_{\tilde{T}_n}(t) - \varphi_{\tilde{T}_n^{(q)}}(t)\right| + \left|\varphi_{\tilde{T}_n^{(q)}}(t) - \varphi_{\tilde{T}^{(q)}}(t)\right| + \left|\varphi_{\tilde{T}^{(q)}}(t) - \varphi_{\tilde{T}}(t)\right| \leq 3\epsilon.$$

The convergence in distribution of $\tilde{T}_n$ to $\tilde{T}$ follows as we can let $\epsilon \to 0$. □

Theorem 3.2. *Assume that $|K(y, y')| \leq B_K$ for all $y, y' \in \mathcal{Y}$ for some $B_K < \infty$. Then we have*

$$E(\tilde{T}_n) = 0, \ \operatorname{var}(\tilde{T}_n) \leq 64B_K^2,$$

and

$$\operatorname{var}(Q_n) = n_1^2 n_2^2 n^{2\Delta-3} \left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right),$$

where $\sigma_1^2, \sigma_2^2 \leq 4\|h\|^2 B_K$.

Proof of Theorem 3.2. First of all, by (7), $|\tilde{K}(y, y')| \leq 4B_K$ for all $y, y' \in \mathcal{Y}$. Then by Lemma A.1, for $\alpha \neq \beta, \alpha, \beta = 1, 2$, we have $E(\tilde{T}_n) = 0$ and

$$\begin{aligned} \operatorname{var}(\tilde{S}_{\alpha\alpha}) = &\ \tfrac{2}{n_\alpha(n_\alpha-1)} E(\tilde{K}^2(y_{\alpha 1}, y_{\alpha 2})) \leq \tfrac{32B_K^2}{n_\alpha(n_\alpha-1)}, \\ \operatorname{var}(\tilde{S}_{\alpha\beta}) = &\ \tfrac{1}{n_\alpha n_\beta} E(\tilde{K}^2(y_{\alpha 1}, y_{\beta 1})) \leq \tfrac{16B_K^2}{n_\alpha n_\beta}. \end{aligned}$$

It follows that

$$\begin{aligned} \operatorname{var}(\tilde{T}_n) = &\ \left(\tfrac{n_1 n_2}{n}\right)^2 \left(\operatorname{var}(\tilde{S}_{11}) + \operatorname{var}(\tilde{S}_{22}) + 4\operatorname{var}(\tilde{S}_{12})\right) \\ \leq &\ \left(\tfrac{n_1 n_2}{n}\right)^2 \left(\tfrac{32B_K^2}{n_1(n_1-1)} + \tfrac{32B_K^2}{n_2(n_2-1)} + \tfrac{64B_K^2}{n_1 n_2}\right) \\ = &\ 32B_K^2 \left(1 + \tfrac{n_2^2}{n^2(n_1-1)} + \tfrac{n_1^2}{n^2(n_2-1)}\right) \leq 64B_K^2. \end{aligned}$$

By the Cauchy-Schwarz inequality and (17), we have

$$\sigma_\alpha^2 \leq E(\|x_{\alpha 1} - \mu_\alpha\|^2)\|h\|^2 = E(\tilde{K}(y_{\alpha 1}, y_{\alpha 1}))\|h\|^2 \leq 4\|h\|^2 B_K.$$

Finally, $\operatorname{var}(Q_n) = n_1^2 n_2^2 n^{2\Delta-3}\left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right)$. □

Theorem 3.3. *Assume that $|K(y, y')| \leq B_K$ for all $y, y' \in \mathcal{Y}$ for some $B_K < \infty$. Then under Assumption 2 and the local alternative hypothesis (16), as $n \to \infty$, we have*

(a) $\tilde{T}_n/(\operatorname{var}(Q_n))^{1/2} \xrightarrow{p} 0$,

(b) $Q_n/(\operatorname{var}(Q_n))^{1/2} \xrightarrow{d} \mathcal{N}(0, 1)$,

(c) $[T_n - n_1 n_2 \|h\|^2/(n^{2-2\Delta})]/\sqrt{\operatorname{var}(T_n)} \xrightarrow{d} \mathcal{N}(0, 1)$, *and hence*

$$P(T_n \geq \hat{C}_\epsilon) = \Phi\left(\frac{n^\Delta \|h\|^2}{2(\sigma_1^2/\tau + \sigma_2^2/(1-\tau))^{1/2}}\right)(1 + o(1)) \to 1,$$

where $\hat{C}_\epsilon$ denotes a consistent estimator of C_ϵ, the upper 100ϵ percentile of $\tilde{T}_n$ with ϵ being the given significance level, τ is defined in Assumption 2, and $\Phi(\cdot)$ denotes the cumulative distribution of $\mathcal{N}(0, 1)$.

Proof of Theorem 3.3. Under the given conditions, by Theorem 3.2, we have $E(\tilde{T}_n) = 0$, $\text{var}(\tilde{T}_n) \leq 64B_K^2$, and as $n \to \infty$,

$$\text{var}(Q_n) = (\tau(1-\tau))^2 \left(\sigma_1^2/\tau + \sigma_2/(1-\tau)\right) n^{2\Delta}(1+o(1)),$$

where τ is defined in Assumption 2. It follows that as $n \to \infty$, $\text{var}(\tilde{T}_n)/\text{var}(Q_n) \to 0$. Thus, we have $\tilde{T}_n/(\text{var}(Q_n))^{1/2} \xrightarrow{p} 0$, and hence (a) is proved. To show (b), notice that by the central limit theorem, as $n_\alpha \to \infty$, we have

$$u_\alpha = n_\alpha^{1/2}\langle \bar{x}_\alpha - \mu_\alpha, \mu_1 - \mu_2\rangle \xrightarrow{d} N(0, \sigma_\alpha^2)\ (\alpha = 1, 2).$$

Since $\bar{x}_1$ and $\bar{x}_2$ are independent and by (19), we have

$$Q_n = n_1 n_2 n^{-(3/2-\Delta)}(u_1/n_1^{1/2} - u_2/n_2^{1/2})$$

and $Q_n/(\text{var}(Q_n))^{1/2} \xrightarrow{d} \mathcal{N}(0,1)$. To show (c), notice that as $n \to \infty$, by (18), we have

$$\frac{T_n - n_1 n_2\|h\|^2/(n^{2-2\Delta})}{\sqrt{\text{var}(T_n)}} = \left[\frac{\tilde{T}_n}{2\sqrt{\text{var}(Q_n)}} + \frac{Q_n}{\sqrt{\text{var}(Q_n)}}\right][1+o(1)] \xrightarrow{d} \mathcal{N}(0,1).$$

Note that $\hat{C}_\epsilon$ denotes a consistent estimator of C_ϵ. Therefore, as $n \to \infty$, we have

$$\begin{aligned}
&P(T_n \geq \hat{C}_\epsilon) = P\left(\frac{T_n - n_1 n_2\|h\|^2/(n^{2-2\Delta})}{2\sqrt{\text{var}(Q_n)}} \geq \frac{\hat{C}_\epsilon - n_1 n_2\|h\|^2/(n^{2-2\Delta})}{2\sqrt{\text{var}(Q_n)}}\right) \\
&= \left\{1 - \Phi\left(\frac{C_\epsilon}{\frac{2n_1n_2}{n^{3/2-\Delta}}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} - \frac{n_1 n_2\|h\|^2/(n^{2-2\Delta})}{\frac{2n_1n_2}{n^{3/2-\Delta}}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right)\right\}(1+o(1)) \\
&= \Phi\left(\frac{n^\Delta\|h\|^2}{2\sqrt{\sigma_1^2/\tau + \sigma_2^2/(1-\tau)}}\right)(1+o(1)) \\
&\to 1.
\end{aligned}$$

The theorem is proved. □

Theorem 4.1. *Assume that $|K(y,y')| \leq B_K$ for all $y, y' \in \mathcal{Y}$ for some $B_K < \infty$. Then under Assumption 1, as $n \to \infty$, we have*

$$|\tilde{K}^*(y_i,y_j) - \tilde{K}(y_i,y_j)| = O_p(n^{-1/2}) \qquad \textit{uniformly for all } y_i, y_j.$$

Proof of Theorem 4.1. Under Assumption 1, set $x_i = K(\cdot, y_i), i = 1, 2, \ldots, n$ and $\mu = E(x_i)$. Then we have $\tilde{K}(y_i,y_j) = \langle x_i - \mu, x_j - \mu\rangle$ and $\tilde{K}^*(y_i,y_j) = \langle x_i - \bar{x}, x_j - \bar{x}\rangle$. It follows that

$$\begin{aligned}
|\tilde{K}^*(y_i,y_j) - \tilde{K}(y_i,y_j)| &= |\langle x_i - \bar{x}, x_j - \bar{x}\rangle - \langle x_i - \mu, x_j - \mu\rangle| \\
&= |\langle \mu - \bar{x}, x_j - \mu\rangle + \langle x_i - \mu, \mu - \bar{x}\rangle + \langle \mu - \bar{x}, \mu - \bar{x}\rangle| \\
&\leq \|\bar{x} - \mu\|\|x_j - \mu\| + \|\bar{x} - \mu\|\|x_i - \mu\| + \|\bar{x} - \mu\|^2 \\
&= \|\bar{x} - \mu\|^2 + \|\bar{x} - \mu\|(\|x_i - \mu\| + \|x_j - \mu\|).
\end{aligned}$$

Note that we have $E[\tilde{K}(y,y')] = 0$ when y, y' are independent. We have

$$\begin{aligned} n\|\bar{x}-\mu\|^2 = & \quad n^{-1}\textstyle\sum_{i=1}^n \sum_{j=1}^n \tilde{K}(y_i,y_j) \\ = & \quad n^{-1}\textstyle\sum_{i=1}^n \tilde{K}(y_i,y_i) + 2(n-1)^{-1}\sum_{1\le i<j\le n}\tilde{K}(y_i,y_j). \end{aligned}$$

It follows that as $n \to \infty$, we have

$$\begin{aligned} E\left[n\|\bar{x}-\mu\|^2\right] = & \quad E\left[\tilde{K}(y,y)\right] \le 4B_K, \\ \text{var}\left[n\|\bar{x}-\mu\|^2\right] = & \quad n^{-1}\text{var}\left[\tilde{K}(y,y)\right] + 2n(n-1)^{-1}E\left[\tilde{K}(y,y')\right]^2 \\ = & \quad 2E\left[\tilde{K}(y,y')\right]^2 [1+o(1)] \le 32B_K^2[1+o(1)], \end{aligned}$$

where we use the fact that var $\left[\tilde{K}(y,y)\right] \le E\left[\tilde{K}(y,y)\right]^2 \le 16B_K^2$. Therefore, as $n \to \infty$, we have

$$\|\bar{x}-\mu\|^2 = O_p(1/n), \quad \text{and} \quad \|\bar{x}-\mu\| = O_p(1/\sqrt{n}),$$

uniformly. Since $\|x_i - \mu\| = \sqrt{\tilde{K}(y_i,y_i)} \le \sqrt{4B_K}$ and similarly $\|x_j-\mu\| \le \sqrt{4B_K}$. As $n \to \infty$, we have

$$|\tilde{K}^*(y_i,y_j) - \tilde{K}(y_i,y_j)| = O_p(n^{-1/2}),$$

uniformly for all y_i, y_j's. □

Theorem 4.2. *Under Assumptions 1, 2 and Condition (25), as $n \to \infty$, we have $\hat{M}_\ell \xrightarrow{p} M_\ell, \ell = 2,3$ and*

$$\hat{\beta}_0 \xrightarrow{p} \beta_0, \hat{\beta}_1 \xrightarrow{p} \beta_1, \hat{d} \xrightarrow{p} d.$$

Proof of Theorem 4.2. First of all, under Condition (25), by (26), we have

$$\sum_{r=1}^\infty \lambda_r \le \left[\sum_{r=1}^\infty \sqrt{\lambda_r}\right]^2 < \infty. \tag{35}$$

By Proposition 12 of [20], under the given conditions, as $n \to \infty$, we have

$$\sum_{r=1}^\infty |\hat{\lambda}_r - \lambda_r| \xrightarrow{p} 0. \tag{36}$$

It follows that as $n \to \infty$, we have

$$\left|\sum_{r=1}^\infty \hat{\lambda}_r - \sum_{r=1}^\infty \lambda_r\right| \le \sum_{r=1}^\infty |\hat{\lambda}_r - \lambda_r| \xrightarrow{p} 0.$$

Therefore, as $n \to \infty$, we have

$$\sum_{r=1}^{\infty} \hat{\lambda}_r \xrightarrow{p} \sum_{r=1}^{\infty} \lambda_r. \tag{37}$$

Then as $n \to \infty$, for $\ell = 2, 3$, we have

$$\begin{array}{rl} |\hat{M}_\ell - M_\ell| \leq & \sum_{r=1}^{\infty} |\hat{\lambda}_r^\ell - \lambda_r^\ell| \\ \leq & \sum_{r=1}^{\infty} |\hat{\lambda}_r - \lambda_r|(\hat{\lambda}_r^{\ell-1} + \hat{\lambda}_r^{\ell-2}\lambda_r + \cdots + \hat{\lambda}_r\lambda_r^{\ell-2} + \lambda_r^{\ell-1}) \\ \leq & \sum_{r=1}^{\infty} |\hat{\lambda}_r - \lambda_r|(\hat{\lambda}_r + \lambda_r)^{\ell-1} \\ \leq & \sum_{r=1}^{\infty} |\hat{\lambda}_r - \lambda_r|(\sum_{r=1}^{\infty} \hat{\lambda}_r + \sum_{r=1}^{\infty} \lambda_r)^{\ell-1} \\ = & (\sum_{r=1}^{\infty} \hat{\lambda}_r + \sum_{r=1}^{\infty} \lambda_r)^{\ell-1} \sum_{r=1}^{\infty} |\hat{\lambda}_r - \lambda_r| \xrightarrow{p} 0. \end{array}$$

That is, as $n \to \infty$, we have $\hat{M}_\ell \xrightarrow{p} M_\ell, \ell = 2, 3$. The remaining claims then follow. The theorem is complete. □

Theorem 4.3. *Under Assumption 1, the first three cumulants of $\tilde{T}_n$ are given by*

$$\begin{array}{rl} E(\tilde{T}_n) = & 0, \\ \mathrm{var}(\tilde{T}_n) = & 2\left\{1 + \left(\frac{n_2^2}{n^2(n_1-1)} + \frac{n_1^2}{n^2(n_2-1)}\right)\right\} E(\tilde{K}^2(y, y')), \\ E(\tilde{T}_n^3) = & 8\left\{1 - \left(\frac{n_2^3}{n^3(n_1-1)^2} + \frac{n_1^3}{n^3(n_2-1)^2}\right)\right\} E(\tilde{K}(y, y')\tilde{K}(y, y'')\tilde{K}(y', y'')) \\ & +4\left(\frac{n_2^3 n_1}{n^3(n_1-1)^2} - \frac{2n_1n_2}{n^3} + \frac{n_1^3 n_2}{n^3(n_2-1)^2}\right) E(\tilde{K}^3(y, y')), \end{array}$$

where $y, y', y'' \overset{i.i.d.}{\sim} P$. Furthermore, under Assumptions 1–3, as $n \to \infty$, we have

$$\begin{array}{rl} \mathrm{var}(\tilde{T}_n) = & 2E(\tilde{K}^2(y, y'))(1 + o(1)), \\ E(\tilde{T}_n^3) = & 8E(\tilde{K}(y, y')\tilde{K}(y, y'')\tilde{K}(y', y''))(1 + o(1)). \end{array}$$

Proof of Theorem 4.3. Under Assumption 1, Lemma A.1 implies that

$$\begin{array}{rl} E(\tilde{T}_n) = & \frac{n_1n_2}{n}(E(\tilde{S}_{11}) + E(\tilde{S}_{22}) - 2E(\tilde{S}_{12})) = 0, \\ \mathrm{var}(\tilde{T}_n) = & ((\frac{n_1n_2}{n})^2 (\mathrm{var}(\tilde{S}_{11}) + \mathrm{var}(\tilde{S}_{22}) + 4\mathrm{var}(\tilde{S}_{12})) \\ = & \left(\frac{n_1n_2}{n}\right)^2 \left(\frac{2}{n_1(n_1-1)} E(\tilde{K}^2(y, y')) + \frac{2}{n_2(n_2-1)} E(\tilde{K}^2(y, y')) \right. \\ & \left. + \frac{4}{n_1n_2} E(\tilde{K}^2(y, y'))\right) \\ = & 2\left(\frac{n_1n_2}{n}\right)^2 \left(\frac{1}{n_1(n_1-1)} + \frac{1}{n_2(n_2-1)} + \frac{2}{n_1n_2}\right) E(\tilde{K}^2(y, y')) \\ = & 2\left\{1 + \left(\frac{n_2^2}{n^2(n_1-1)} + \frac{n_1^2}{n^2(n_2-1)}\right)\right\} E(\tilde{K}^2(y, y')). \end{array}$$

Note that

$$\begin{array}{rl} \tilde{T}_n^3 = & \tilde{S}_{11}^3 + \tilde{S}_{22}^3 - 8\tilde{S}_{12}^3 + 3(\tilde{S}_{11}^2\tilde{S}_{22} + \tilde{S}_{11}\tilde{S}_{22}^2) - 6(\tilde{S}_{11}^2\tilde{S}_{12} + \tilde{S}_{12}\tilde{S}_{22}^2) \\ & +12(\tilde{S}_{11}\tilde{S}_{12}^2 + \tilde{S}_{12}^2\tilde{S}_{22}) - 12\tilde{S}_{11}\tilde{S}_{12}\tilde{S}_{22}. \end{array}$$

First of all, it is easy to see that $E(\tilde{S}_{11}^2\tilde{S}_{22}) = E(\tilde{S}_{11}\tilde{S}_{22}^2) = E(\tilde{S}_{11}^2\tilde{S}_{12}) = E(\tilde{S}_{12}\tilde{S}_{22}^2) = 0$. Further,

$$\begin{aligned}
E(\tilde{S}_{11}\tilde{S}_{12}\tilde{S}_{22}) = &\ \frac{E\Big(\sum_{j\neq k}\tilde{K}(y_{1j},y_{1k})\sum_{\alpha\neq\beta}\tilde{K}(y_{2\alpha},y_{2\beta})\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}\tilde{K}(y_{1r},y_{2s})\Big)}{n_1^2(n_1-1)n_2^2(n_2-1)}\\
= &\ \frac{\sum_{j\neq k}\sum_{\alpha\neq\beta}\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}E(\tilde{K}(y_{1j},y_{1k})\tilde{K}(y_{1r},y_{2s})\tilde{K}(y_{2\alpha},y_{2\beta}))}{n_1^2(n_1-1)n_2^2(n_2-1)}\\
= &\ 0.
\end{aligned}$$

Now let $y, y', y'' \overset{\text{i.i.d.}}{\sim} P$. Then

$$\begin{aligned}
E(\tilde{S}_{11}\tilde{S}_{12}^2) &= \frac{E\Big(\sum_{j\neq k}\tilde{K}(y_{1j},y_{1k})\sum_{\alpha=1}^{n_1}\sum_{\beta=1}^{n_2}\tilde{K}(y_{1\alpha},y_{2\beta})\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}\tilde{K}(y_{1r},y_{2s})\Big)}{n_1^3(n_1-1)n_2^2}\\
&= \frac{\sum_{j\neq k}\sum_{\alpha=1}^{n_1}\sum_{\beta=1}^{n_2}\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}E(\tilde{K}(y_{1j},y_{1k})\tilde{K}(y_{1\alpha},y_{2\beta})\tilde{K}(y_{1r},y_{2s}))}{n_1^3(n_1-1)n_2^2}\\
&= \frac{2}{n_1^3(n_1-1)n_2}\sum_{j\neq k}\sum_{\beta=1}^{n_2}E(\tilde{K}(y_{1j},y_{1k})\tilde{K}(y_{1j},y_{2\beta})\tilde{K}(y_{1k},y_{2\beta}))\\
&= \frac{2}{n_1^2n_2}E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')).
\end{aligned}$$

Similarly, we have $E(\tilde{S}_{12}^2\tilde{S}_{22}) = 2(n_1n_2^2)^{-1}E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y''))$. Moreover

$$\begin{aligned}
E(\tilde{S}_{12}^3) = &\ \frac{1}{n_1^3n_2^3}E\left(\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\tilde{K}(y_{1j},y_{2k})\right)^3\\
= &\ \frac{1}{n_1^3n_2^3}E\left(\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\tilde{K}^3(y_{1j},y_{2k})\right)\\
= &\ \frac{1}{n_1^3n_2^3}(n_1n_2E(\tilde{K}^3(y,y')))\\
= &\ \frac{1}{n_1^2n_2^2}E(\tilde{K}^3(y,y')).
\end{aligned}$$

Finally,

$$\begin{aligned}
E(\tilde{S}_{11}^3) = &\ \frac{8}{n_1^3(n_1-1)^3}E\left(\sum_{j<k}\tilde{K}(y_{1j},y_{1k})\right)^3\\
= &\ \frac{8}{n_1^3(n_1-1)^3}E\Big(\sum_{j<k}\tilde{K}^3(y_{1j},y_{1k}) + 3\sum\nolimits^{*}\tilde{K}^2(y_{1j},y_{1k})\tilde{K}(y_{1\alpha},y_{1\beta})\\
&+6\sum\nolimits^{**}\tilde{K}(y_{1j},y_{1k})\tilde{K}(y_{1\alpha},y_{1\beta})\tilde{K}(y_{1r},y_{1s})\Big)\\
= &\ \frac{8}{n_1^3(n_1-1)^3}\Big\{\frac{n_1(n_1-1)}{2}E(\tilde{K}^3(y,y'))\\
&+6\frac{n_1(n_1-1)(n_1-2)}{3!}E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y''))\Big\}\\
= &\ \frac{8(n_1-2)}{n_1^2(n_1-1)^2}E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')) + \frac{4}{n_1^2(n_1-1)^2}E(\tilde{K}^3(y,y')),
\end{aligned}$$

where $*$ means "$j < k, \alpha < \beta$" and "$(j,k) \neq (\alpha,\beta)$", while $**$ means "$j < k, \alpha < \beta, u < v$" and "$(j,k),(\alpha,\beta),(r,s)$ are not mutually equal to each other.". Similarly,

$$E(\tilde{S}_{22}^3) = \frac{8(n_2-2)}{n_2^2(n_2-1)^2}E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')) + \frac{4}{n_2^2(n_2-1)^2}E(\tilde{K}^3(y,y')).$$

Thus, we have

$$\begin{aligned}
E(\tilde{T}_n^3) = & \tfrac{n_1^3 n_2^3}{n^3}(E(\tilde{S}_{11}^3) + E(\tilde{S}_{22}^3) - 8E(\tilde{S}_{12}^3) + 12E(\tilde{S}_{11}\tilde{S}_{12}^2) + 12E(\tilde{S}_{12}^2\tilde{S}_{22})) \\
= & \tfrac{n_1^3 n_2^3}{n^3}\left\{8\left(\tfrac{n_1-2}{n_1^2(n_1-1)^2} + \tfrac{3}{n_1^2 n_2} + \tfrac{3}{n_1 n_2^2} + \tfrac{n_2-2}{n_2(n_2-1)^2}\right)\right. \\
& \cdot E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')) \\
& \left. +4\left(\tfrac{1}{n_1^2(n_1-1)^2} - \tfrac{2}{n_1^2 n_2^2} + \tfrac{1}{n_2^2(n_2-1)^2}\right) E(\tilde{K}^3(y,y'))\right\} \\
= & 8\left\{1 - \left(\tfrac{n_2^3}{n^3(n_1-1)^2} + \tfrac{n_1^3}{n^3(n_2-1)^2}\right)\right\} E(\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')) \\
& +4\left(\tfrac{n_2^3 n_1}{n^3(n_1-1)^2} - \tfrac{2n_1 n_2}{n^3} + \tfrac{n_1^3 n_2}{n^3(n_2-1)^2}\right) E(\tilde{K}^3(y,y')).
\end{aligned}$$

□

Theorem 4.4. *Assume that $|K(y,y')| \le B_K$ for all $y, y' \in \mathcal{Y}$ for some $B_K < \infty$. Then under Assumptions 1 and 2, as $n \to \infty$, we have $\hat{M}_\ell \xrightarrow{p} M_\ell, \ell = 2, 3$ and*

$$\hat{\beta}_0 \xrightarrow{p} \beta_0, \hat{\beta}_1 \xrightarrow{p} \beta_1, \hat{d} \xrightarrow{p} d.$$

Proof of Theorem 4.4. By (27), under Assumption 2, as $n \to \infty$, we can write

$$\begin{aligned}
\hat{M}_2 = & \tilde{M}_2^* \left[1 + O(n^{-1})\right], \\
\hat{M}_3 = & \tilde{M}_3^* \left[1 - O(n^{-2})\right] + \tilde{M}_{23}^* \left[O(n^{-1})\right],
\end{aligned}$$

where

$$\begin{aligned}
\tilde{M}_2^* = & \tfrac{2}{n(n-1)} \textstyle\sum_{1 \le i < j \le n} (\tilde{K}^*(y_i, y_j))^2, \\
\tilde{M}_3^* = & \tfrac{6}{n(n-1)(n-2)} \textstyle\sum_{1 \le i < j < k \le n} \tilde{K}^*(y_i, y_j)\tilde{K}^*(y_j, y_k)\tilde{K}^*(y_k, y_i), \\
\tilde{M}_{23}^* = & \tfrac{2}{n(n-1)} \textstyle\sum_{1 \le i < j \le n} (\tilde{K}^*(y_i, y_j))^3.
\end{aligned}$$

By Theorem 4.1, we have $\tilde{K}^*(y_i, y_j) = \tilde{K}(y_i, y_j) + O_p(n^{-1/2})$ uniformly for all y_i, y_j's. Since $|\tilde{K}(y,y')| \le 4B_K < \infty$ for all $y, y' \in \mathcal{Y}$, we have

$$\tilde{M}_2^* = \tilde{M}_2 + O_p(n^{-1/2}), \quad \tilde{M}_3^* = \tilde{M}_3 + O_p(n^{-1/2}), \quad \tilde{M}_{23}^* = \tilde{M}_{23} + O_p(n^{-1/2}),$$

where

$$\begin{aligned}
\tilde{M}_2 = & \tfrac{2}{n(n-1)} \textstyle\sum_{1 \le i < j \le n} (\tilde{K}(y_i, y_j))^2, \\
\tilde{M}_3 = & \tfrac{6}{n(n-1)(n-2)} \textstyle\sum_{1 \le i < j < k \le n} \tilde{K}(y_i, y_j)\tilde{K}(y_j, y_k)\tilde{K}(y_k, y_i), \\
\tilde{M}_{23} = & \tfrac{2}{n(n-1)} \textstyle\sum_{1 \le i < j \le n} (\tilde{K}(y_i, y_j))^3.
\end{aligned}$$

Since $\tilde{M}_2, \tilde{M}_3$ and $\tilde{M}_{23}$ are U-statistics for M_2, M_3, and $M_{23} = E\left[\tilde{K}(y,y')\right]^3$ respectively and under the given conditions, we have

$$\begin{aligned}
E\left[\tilde{K}(y,y')\right]^4 \le & \quad (4B_K)^4 < \infty, \\
E\left[\tilde{K}(y,y')\tilde{K}(y,y'')\tilde{K}(y',y'')\right]^2 \le & \quad (4B_K)^6 < \infty, \\
E\left[\tilde{K}(y,y')\right]^6 \le & \quad (4B_K)^6 < \infty,
\end{aligned}$$

then by Lemma A of [19, p. 185], as $n \to \infty$, we have $\tilde{M}_2 \xrightarrow{p} M_2, \tilde{M}_3 \xrightarrow{p} M_3$, and $\tilde{M}_{23} \xrightarrow{p} M_{23}$. It follows that as $n \to \infty$, we have $\tilde{M}_2^* \xrightarrow{p} M_2, \tilde{M}_3^* \xrightarrow{p} M_3$, and $\tilde{M}_{23}^* \xrightarrow{p} M_{23}$. Thus, as $n \to \infty$, we have $\hat{M}_\ell \xrightarrow{p} M_\ell, \ell = 2, 3$. The remaining claims then follow. The theorem is complete. □

Appendix B: Tables of the numerical experimental results

The detailed results of the numerical experiments based on the artificial data sets in Section 5.2 are displayed in Tables 10–15.

Table 10: Empirical sizes (in %) of all tests obtained under Models 1-3. Column M denotes Model, $\mathbf{n} = (n_1, n_2)$, $\mathbf{n}_1 = (20, 30)$, and $\mathbf{n}_2 = (40, 60)$.

M	$\mathbf{n}$	p	ρ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
1	$\mathbf{n}_1$	10	0.2	5.5	5.3	5.4	5.4	3.3	3.7	5.4	5.3
			0.5	5.9	5.8	5.5	5.9	4.7	5.0	5.6	5.8
			0.8	6.0	5.6	6.2	5.5	5.1	5.2	5.5	5.4
		100	0.2	5.8	6.4	5.5	5.3	1.9	2.6	5.7	5.9
			0.5	5.6	5.9	6.0	5.8	4.5	5.0	5.9	5.7
			0.8	4.4	4.1	4.1	4.3	3.8	3.9	4.2	4.0
		500	0.2	6.0	6.0	5.7	5.7	1.9	2.9	5.9	6.2
			0.5	6.4	5.6	5.8	5.4	4.0	4.5	5.2	5.6
			0.8	4.8	4.8	4.8	4.9	4.5	4.4	5.0	4.7
	$\mathbf{n}_2$	10	0.2	4.1	4.0	3.6	4.4	3.2	3.5	3.9	4.2
			0.5	5.1	5.0	5.0	5.5	4.0	4.4	4.7	4.7
			0.8	5.4	5.9	6.1	5.4	5.7	5.8	5.8	5.8
		100	0.2	4.7	4.5	4.8	4.6	2.0	2.8	4.3	4.2
			0.5	4.3	4.1	4.1	4.2	3.1	3.4	3.8	3.9
			0.8	3.9	3.8	3.6	3.8	3.9	3.8	4.2	3.9
		500	0.2	5.6	5.9	5.8	5.5	3.7	4.4	5.6	5.7
			0.5	5.1	6.1	5.1	6.2	5.1	5.3	5.9	5.9
			0.8	5.5	5.3	5.7	5.0	5.0	5.2	5.1	5.4
2	$\mathbf{n}_1$	10	0.2	4.3	4.4	4.1	3.8	1.5	2.6	4.1	3.9
			0.5	5.4	5.4	5.9	6.0	3.5	4.3	5.4	4.9
			0.8	4.6	4.8	4.7	4.9	3.9	4.5	5.0	5.1
		100	0.2	5.0	4.7	4.5	4.3	0.5	2.2	4.3	4.9
			0.5	4.9	4.8	4.8	5.6	2.2	3.9	5.1	4.8
			0.8	5.5	5.9	5.9	5.5	4.9	5.5	5.8	5.9
		500	0.2	4.4	5.0	4.6	5.0	0.2	2.2	4.2	4.6
			0.5	4.3	3.9	4.6	4.0	1.5	3.5	3.8	4.6
			0.8	5.9	4.3	4.6	4.4	3.7	4.9	4.7	5.3
	$\mathbf{n}_2$	10	0.2	5.4	5.0	5.4	4.7	3.4	4.1	4.7	4.9
			0.5	5.3	4.6	4.4	4.6	3.5	3.6	4.6	4.3
			0.8	5.2	5.0	5.4	4.8	4.2	5.4	4.6	5.7
		100	0.2	4.8	5.3	5.5	5.1	0.6	3.3	4.6	4.5
			0.5	3.6	5.2	4.5	4.7	3.1	3.9	4.9	4.2
			0.8	4.7	4.7	4.5	4.6	4.4	4.3	4.6	4.4
		500	0.2	6.0	6.5	5.8	6.2	1.1	3.8	5.8	6.2
			0.5	4.5	4.9	4.8	5.0	3.2	4.2	4.8	4.6
			0.8	3.6	5.1	4.6	5.0	4.3	4.0	4.7	4.2
3	$\mathbf{n}_1$	10	0.2	5.6	5.0	5.8	5.6	3.1	3.9	4.8	5.2
			0.5	5.9	5.7	5.1	5.7	4.3	3.5	5.4	4.8
			0.8	6.6	5.6	5.8	5.1	4.9	5.9	5.4	6.2
		100	0.2	5.1	5.0	5.0	5.2	0.3	2.7	4.5	4.9
			0.5	4.8	4.1	4.2	4.0	1.9	3.8	3.9	4.6
			0.8	5.2	4.3	4.4	4.4	3.7	4.4	4.4	4.7
		500	0.2	5.6	6.3	5.8	6.3	0.3	3.5	5.4	5.7
			0.5	4.4	4.5	4.9	5.0	2.2	4.3	4.5	4.8

Continued on next page

Table 10 – *Continued from previous page*

M	$\mathbf{n}$	p	ρ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
			0.8	5.6	5.4	5.2	5.1	4.3	4.6	5.6	5.2
	$\mathbf{n}_2$	10	0.2	5.3	5.0	4.5	5.1	3.2	4.5	4.8	4.8
			0.5	5.1	6.4	5.7	6.5	5.0	4.8	5.6	4.9
			0.8	5.2	5.8	5.2	5.3	4.8	4.5	5.2	4.9
		100	0.2	5.1	4.8	5.5	5.8	1.5	3.5	4.2	4.8
			0.5	5.0	5.8	5.7	5.6	4.3	4.8	5.6	5.4
			0.8	6.2	5.3	5.3	4.9	4.9	5.0	5.2	5.2
		500	0.2	5.0	5.5	5.2	5.8	1.0	3.2	4.9	5.0
			0.5	4.5	4.8	4.7	4.6	3.2	4.2	4.6	4.8
			0.8	5.7	4.5	5.2	4.9	3.9	5.0	4.3	5.2

Table 11: Empirical powers (in %) of all tests under Model 4 ($\mathbf{n} = (n_1, n_2)$, $\mathbf{n}_1 = (20, 30)$, $\mathbf{n}_2 = (40, 60)$).

$\mathbf{n}$	p	ρ	δ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
$\mathbf{n}_1$	10	0.2	0.2	22.0	19.6	19.4	17.0	15.0	18.0	19.8	21.8
			0.4	63.3	58.5	59.6	53.1	51.4	58.6	58.3	61.5
			0.6	94.6	92.4	92.2	87.6	89.6	92.5	92.0	93.6
		0.5	0.2	14.1	13.3	13.9	11.6	11.1	12.2	13.1	14.0
			0.4	41.7	37.0	40.0	35.2	34.4	37.3	37.3	39.4
			0.6	76.4	68.8	72.0	66.0	66.3	70.4	68.2	72.2
		0.8	0.2	10.1	9.3	9.5	9.2	8.7	8.6	9.6	9.6
			0.4	30.7	27.1	28.8	24.3	25.2	27.1	26.5	27.9
			0.6	57.5	47.9	50.9	43.1	46.3	48.5	47.6	50.1
	100	0.2	0.2	29.5	28.1	28.7	27.4	14.6	20.6	27.2	28.5
			0.4	82.2	80.2	81.1	79.3	67.8	73.8	79.5	81.6
			0.6	99.2	99.3	99.3	99.2	97.6	99.0	99.3	99.3
		0.5	0.2	17.8	17.3	18.1	17.1	13.7	15.6	17.1	18.1
			0.4	45.3	41.4	43.9	40.7	36.5	40.0	40.9	43.6
			0.6	80.0	76.3	79.2	76.2	71.9	75.7	76.0	78.3
		0.8	0.2	13.0	10.2	11.0	9.7	9.4	10.0	10.5	10.9
			0.4	30.9	25.8	27.7	23.4	24.6	25.6	25.5	26.6
			0.6	61.0	54.2	55.9	50.2	52.8	54.8	53.7	56.0
	500	0.2	0.2	29.0	28.9	29.4	29.2	16.0	20.3	28.0	28.7
			0.4	85.7	84.4	85.7	85.3	69.9	77.7	84.5	85.5
			0.6	99.6	99.1	99.4	99.2	97.9	98.8	99.1	99.6
		0.5	0.2	14.6	13.2	13.8	13.0	10.6	11.8	13.5	14.0
			0.4	45.5	39.9	43.5	39.6	35.5	38.6	38.7	42.4
			0.6	81.7	77.0	80.6	77.0	73.3	77.0	77.0	79.2
		0.8	0.2	9.8	9.5	9.9	9.8	9.6	9.4	9.8	9.6
			0.4	32.1	27.1	28.8	24.3	25.9	27.1	26.6	28.0
			0.6	59.1	52.4	54.4	47.9	50.8	52.6	52.2	53.3
$\mathbf{n}_2$	10	0.2	0.2	40.6	36.4	36.3	31.2	33.1	36.4	36.0	40.0
			0.4	91.6	87.9	89.0	83.3	85.8	89.8	88.0	90.9
			0.6	99.9	99.7	99.7	99.7	99.7	99.9	99.7	99.9
		0.5	0.2	27.1	24.1	25.1	22.2	21.5	23.6	23.1	24.5
			0.4	70.7	63.9	68.0	61.8	61.7	65.5	63.1	66.5
			0.6	96.8	94.2	95.7	92.3	93.3	95.0	93.8	95.5
		0.8	0.2	18.1	16.1	16.6	14.6	15.1	15.7	15.7	16.0
			0.4	55.1	47.6	51.0	42.7	46.1	47.5	46.9	48.2
			0.6	88.0	81.8	83.9	78.3	81.1	82.4	81.6	82.6
	100	0.2	0.2	53.8	52.5	53.4	50.8	43.6	48.0	51.5	52.8
			0.4	98.3	98.0	98.3	97.7	96.6	97.7	98.1	98.2
			0.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		0.5	0.2	25.6	22.3	25.2	22.3	20.5	22.1	22.0	23.8
			0.4	77.4	72.3	75.5	72.5	70.4	73.7	72.5	74.6
			0.6	98.3	97.5	98.4	97.7	97.1	98.0	97.5	98.2
		0.8	0.2	17.2	15.2	15.6	13.8	14.1	15.0	14.6	15.5
			0.4	56.6	49.7	52.4	46.5	48.5	50.2	49.2	50.5

Continued on next page

Table 11 – *Continued from previous page*

n	p	ρ	δ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
			0.6	88.7	83.0	84.6	79.2	82.4	83.3	82.7	83.6
	500	0.2	0.2	59.7	57.4	58.7	56.9	45.4	51.4	56.5	58.6
			0.4	99.1	98.8	99.1	99.0	97.8	98.2	98.9	99.1
			0.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		0.5	0.2	24.7	22.8	24.3	22.2	20.8	22.5	22.6	23.7
			0.4	78.5	75.4	77.4	75.2	72.3	75.1	75.3	76.6
			0.6	98.7	98.3	98.7	98.2	97.9	98.3	98.0	98.5
		0.8	0.2	18.9	17.1	17.6	16.1	15.8	16.4	16.7	16.5
			0.4	57.3	50.6	53.3	46.5	49.5	50.8	49.8	51.4
			0.6	90.2	84.2	86.6	80.1	83.2	84.5	83.7	85.5

Table 12: Empirical powers (in %) of all tests under Model 5 ($\mathbf{n} = (n_1, n_2)$).

n	p	ρ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
$(20, 30)$	10	0.3	21.2	70.7	9.4	42.5	61.2	17.4	70.0	20.5
		0.5	11.1	24.8	7.0	21.4	21.6	10.0	24.3	10.9
		0.7	7.6	10.3	7.1	10.6	9.0	6.6	10.0	7.1
	100	0.3	30.6	97.5	9.5	50.3	93.2	17.5	97.8	21.4
		0.5	14.2	35.7	6.9	19.4	28.7	9.7	35.9	11.5
		0.7	7.1	10.9	6.0	9.4	8.9	6.6	10.5	7.0
	500	0.3	32.5	99.3	8.5	49.7	96.2	19.2	99.5	21.8
		0.5	13.8	41.3	7.6	21.0	33.9	10.2	41.4	11.3
		0.7	9.0	12.3	6.4	10.0	10.6	6.8	12.1	7.2
$(40, 60)$	10	0.3	76.7	100.0	9.9	97.9	99.9	50.7	100.0	54.7
		0.5	23.4	71.2	7.9	52.3	65.2	15.7	70.2	16.6
		0.7	9.8	15.6	6.1	15.6	13.6	7.8	14.8	7.9
	100	0.3	98.3	100.0	11.8	100.0	100.0	57.0	100.0	61.0
		0.5	31.9	97.9	5.9	56.2	96.1	15.2	98.1	16.0
		0.7	10.7	19.3	6.4	16.4	17.7	8.0	19.4	8.0
	500	0.3	99.2	100.0	13.2	99.9	100.0	58.6	100.0	62.9
		0.5	34.2	99.2	7.6	56.7	97.6	17.1	99.2	18.2
		0.7	11.3	20.0	6.3	16.8	18	8.3	19.8	8.5

Table 13: Empirical powers (in %) of all tests under Model 6. Column D lists the distributions, $\mathbf{n} = (n_1, n_2)$, $\mathbf{n}_1 = (20, 30)$, and $\mathbf{n}_2 = (40, 60)$.

D	**n**	p	ρ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
$\mathcal{N}$	$\mathbf{n}_1$	10	0.2	5.6	6.1	4.3	4.5	3.5	3.7	6.2	5.6
			0.5	7.1	11.4	5.0	5.1	7.9	6.6	11.3	8.0
			0.8	7.8	10.7	5.1	4.5	9.4	8.3	10.8	8.8
		100	0.2	6.6	13.1	3.4	3.5	1.3	1.6	11.9	6.5
			0.5	14.4	65.0	4.4	4.2	32.1	13.1	64.3	23.6
			0.8	31.8	89.7	3.0	2.8	79.2	35.9	89.3	46.2
		500	0.2	5.1	13.0	2.6	2.6	0.5	0.9	11.6	5.1
			0.5	13.7	87.0	2.2	2.3	36.2	13.5	86.9	26.2
			0.8	54.9	99.9	3.1	2.3	99.8	60.0	99.9	83
	$\mathbf{n}_2$	10	0.2	5.2	7.6	4.4	4.6	6.1	5.0	7.5	6.1
			0.5	9.3	28.0	3.1	3.9	22.2	12.8	25.9	15.4
			0.8	9.5	17.4	4.6	5.2	15.6	11.6	16.6	11.9
		100	0.2	9.6	39.4	3.1	2.8	10.3	4.3	35.0	11.7
			0.5	55.2	100.0	3.1	3.7	100.0	79.4	100.0	89.2
			0.8	99.8	100.0	3.5	3.5	100.0	100.0	100.0	100.0
		500	0.2	9.5	65.8	2.1	2.0	9.9	3.8	62.1	13.1

Continued on next page

Table 13 – *Continued from previous page*

D	$\mathbf{n}$	p	ρ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
			0.5	78.3	100.0	1.8	2.0	100.0	96.8	100.0	99.8
			0.8	100.0	100.0	2.1	1.9	100.0	100.0	100.0	100.0
t_4	$\mathbf{n}_1$	10	0.2	5.8	7.8	5.4	4.7	2.9	3.4	8.0	5.6
			0.5	6.9	15.8	3.9	4.0	9.2	6.0	15.8	8.9
			0.8	7.1	13.3	4.2	3.7	9.7	6.4	12.6	7.3
		100	0.2	6.8	27.0	3.7	4.5	0.1	0.8	22.8	6.8
			0.5	10.6	94.7	3.3	3.4	41.1	8.9	94.9	18.1
			0.8	32.4	99.9	2.7	2.9	99.4	36.6	100.0	46.8
		500	0.2	5.8	48.5	3.3	4.0	0.0	0.5	36.2	6.1
			0.5	13.5	100.0	3.0	4.1	65.8	11.3	100.0	23.7
			0.8	53.6	100.0	2.7	2.9	100.0	60.5	100.0	81.4
	$\mathbf{n}_2$	10	0.2	7.1	10.1	4.3	4.9	5.4	5.3	10.5	7.0
			0.5	11.6	41.2	4.9	4.7	31.6	13.0	39.3	15.1
			0.8	9.4	30.9	5.7	5.4	25.8	11.3	29.4	12.2
		100	0.2	7.9	73.5	4.8	6.4	7.2	4.3	71.1	10.8
			0.5	53.2	100.0	4.4	6.2	100.0	80.0	100.0	90.0
			0.8	99.7	100.0	4.8	4.8	100.0	100.0	100.0	100.0
		500	0.2	9.7	99.2	3.3	4.2	5.9	4.1	99.0	13.9
			0.5	79.6	100.0	5.4	8.2	100.0	97.4	100.0	99.9
			0.8	100.0	100.0	2.9	4.1	100.0	100.0	100.0	100.0

Table 14: Empirical powers (in %) of all tests under Model 7 ($\mathbf{n} = (n_1, n_2)$, $\mathbf{n}_1 = (20, 30)$, $\mathbf{n}_2 = (40, 60)$).

$\mathbf{n}$	p	ρ	δ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
$\mathbf{n}_1$	10	0.2	0.5	4.1	6.7	6.2	7.5	4.1	3.7	6.8	5.2
			0.75	8.2	28.8	19.1	24.1	19.2	8.7	28.5	13.2
			1.0	19.7	80.1	54.0	63.7	67.1	25.3	80.1	33.6
		0.5	0.5	4.8	7.1	5.6	7.2	4.8	4.0	6.8	5.1
			0.75	7.2	18.6	15.4	19.5	14.6	9.4	18.4	10.3
			1.0	11.9	51.0	37.2	48.8	42.8	21.3	50.5	26.4
		0.8	0.5	5.3	6.5	6.1	7.1	5.8	6.2	6.7	6.3
			0.75	5.8	10.7	10.6	12.0	10.3	9.0	10.8	9.4
			1.0	7.2	24.3	20.2	25.9	21.7	15.9	24.3	17.1
	100	0.2	0.5	4.7	10.3	8.5	11.6	2.9	3.0	9.5	5.5
			0.75	9.1	69.0	44.0	63.9	28.5	6.2	67.4	16.3
			1.0	44.0	100.0	96.5	99.4	98.7	38.2	100.0	66.6
		0.5	0.5	5.8	8.7	7.6	9.2	6.0	5.7	9.2	7.2
			0.75	7.0	25.1	18.1	26.8	16.8	9.6	25.0	12.1
			1.0	11.4	70.6	45.0	65.3	55.6	22.0	69.8	27.9
		0.8	0.5	4.5	5.7	5.7	6.3	5.3	5.5	5.8	5.9
			0.75	6.1	9.9	9.4	11.0	9.3	8.2	10.2	8.9
			1.0	8.6	27.0	22.9	28.4	24.9	18.6	27.4	19.9
	500	0.2	0.5	5.9	13.2	9.1	12.6	2.6	2.8	12.4	7.2
			0.75	8.4	78.7	49.4	73.9	27.6	6.4	77.1	16.5
			1.0	47.5	100.0	98.6	99.8	99.2	39.1	100.0	74.3
		0.5	0.5	5.0	7.5	6.1	8.1	4.9	5.3	8.2	6.0
			0.75	7.2	25.1	17.1	26.1	16.8	9.9	24.2	12.6
			1.0	11.7	72.0	48.6	67.3	59.2	25.7	71.8	31.5
		0.8	0.5	4.8	7.1	6.8	7.5	6.6	6.1	7.0	6.7
			0.75	6.8	11.8	11.1	13.5	11.0	9.9	12.3	10.3
			1.0	9.0	26.8	22.4	28.3	25.2	18.3	27.5	20.0
$\mathbf{n}_2$	10	0.2	0.5	8.2	15.3	10.5	12.8	11.4	7.8	14.6	10.0
			0.75	16.7	65.3	43.8	53.7	57.6	23.2	64.4	27.1
			1.0	64.8	99.2	92.1	95.6	98.6	78.9	99.3	81.4
		0.5	0.5	4.5	9.3	7.4	11.2	7.1	5.3	8.5	5.7
			0.75	11.6	39.3	29.1	38.3	35.0	21.0	37.9	22.3
			1.0	31.8	89.4	71.5	84.0	86.2	54.3	88.8	56.0
		0.8	0.5	5.9	7.4	7.2	7.4	6.1	6.1	6.6	6.3

Continued on next page

Table 14 – *Continued from previous page*

n	p	ρ	δ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
			0.75	8.5	19.0	17.3	20.8	17.7	15.8	18.6	16.3
			1.0	19.3	52.9	46.3	55.9	50.5	40.7	51.8	41.5
	100	0.2	0.5	6.6	24.1	15.8	25.5	10.3	5.8	21.8	8.4
			0.75	35.2	99.7	89.8	97.5	95.3	37.0	99.6	53.0
			1.0	99.1	100.0	100.0	100.0	100.0	99.2	100.0	99.6
		0.5	0.5	5.9	11.8	8.8	13.1	8.8	6.7	11.4	7.4
			0.75	12.2	49.9	35.6	51.6	44.9	21.6	50.1	23.5
			1.0	43.6	98.8	86.3	97.3	98.0	68.0	98.7	70.7
		0.8	0.5	4.9	6.9	6.4	7.4	6.4	5.9	6.8	6.1
			0.75	8.4	21.8	19.0	24.6	21.1	16.8	21.9	17.3
			1.0	20.1	56.4	49.5	58.7	54.9	43.7	55.7	44.7
	500	0.2	0.5	6.9	24.6	15.8	27.7	11.0	6.4	23.1	9.3
			0.75	36.6	99.9	94.4	99.2	98.9	38.2	100.0	57.6
			1.0	99.8	100.0	100.0	100.0	100.0	99.8	100.0	100.0
		0.5	0.5	6.4	13.7	9.6	14.2	10.5	7.7	13.2	8.9
			0.75	12.4	56.7	38.7	58.2	49.7	24.3	56.1	26.6
			1.0	42.8	98.9	87.5	98.1	98.6	68.4	98.8	72.0
		0.8	0.5	5.5	7.6	6.7	8.0	6.5	6.6	6.7	7.1
			0.75	7.7	20.3	18.9	22.9	19.5	16.3	20.1	17.0
			1.0	20.9	56.5	48.9	59.6	55.0	44.9	55.8	45.5

Table 15: Empirical powers (in %) of all tests under Model 8 ($\mathbf{n} = (n_1, n_2)$, $\mathbf{n}_1 = (20, 30)$, $\mathbf{n}_2 = (40, 60)$).

n	p	ρ	δ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
$\mathbf{n}_1$	10	0.2	0.5	6.3	21.5	11.2	15.9	12.2	7.1	21.4	10.6
			0.75	20.1	78.4	43.0	53.8	67.7	28.4	78.9	37.4
			1.0	61.8	99.6	85.3	91.2	98.9	78.6	99.6	85.1
		0.5	0.5	6.0	13.9	9.4	12.0	10.8	7.9	13.9	9.0
			0.75	10.0	53.4	30.1	41.6	43.5	21.0	53.0	25.3
			1.0	34.2	93.0	67.5	79.8	89.7	59.5	93.4	64.5
		0.8	0.5	5.0	9.2	8.7	10.1	8.2	7.6	9.3	8.2
			0.75	8.9	28.8	23.0	27.6	25.3	19.7	28.4	21.0
			1.0	18.1	60.6	48.4	54.9	57.3	43.9	60.6	46.3
	100	0.2	0.5	7.2	57.3	21.3	35.3	16.0	6.1	53.5	12.1
			0.75	55.4	100.0	90.4	98.3	99.7	51.6	100.0	79.6
			1.0	99.8	100.0	100.0	100.0	100.0	99.4	100.0	100.0
		0.5	0.5	6.1	19.2	12.2	17.5	12.9	8.4	18.3	10.2
			0.75	15.0	79.8	42.6	62.3	68.9	28.6	79.3	34.9
			1.0	54.1	99.9	86.1	96.2	99.8	75.1	100.0	80.1
		0.8	0.5	4.3	8.6	7.9	8.9	7.8	7.2	9.2	8.1
			0.75	8.6	31.3	24.6	29.8	28.6	21.0	31.2	22.5
			1.0	23.9	72.2	55.2	65.7	68.8	50.7	72.0	53.5
	500	0.2	0.5	7.5	63.7	24.1	43.3	16.0	5.3	60.9	12.2
			0.75	67.9	100.0	94.5	99.5	100.0	60.2	100.0	86.4
			1.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0
		0.5	0.5	4.1	17.9	8.8	16.4	11.2	6.0	17.2	7.9
			0.75	13.6	84.3	47.0	66.1	73.5	31.7	84.4	37.6
			1.0	58.0	100.0	88.8	98.5	100.0	76.8	100.0	83.0
		0.8	0.5	5.9	10.9	9.9	11.3	9.6	8.2	10.2	9.1
			0.75	10.3	33.6	27	32.2	31.5	22.9	34.1	24.3
			1.0	21.7	74.6	55.7	66.5	70.6	50.5	74.3	53.0
$\mathbf{n}_2$	10	0.2	0.5	12.6	46.9	26.2	33.2	40.2	17.0	46.0	20.6
			0.75	66.0	99.1	84.3	91.3	98.8	80.9	99.1	84.9
			1.0	99.7	100.0	99.7	100.0	100.0	99.9	100.0	99.9
		0.5	0.5	10.0	27.9	19.2	25.6	24.3	14.9	27.2	16.3
			0.75	38.5	91.6	65.1	79.8	89.4	60.8	91.0	63.1
			1.0	89.7	100.0	96.3	99.7	100.0	97.2	100.0	97.6
		0.8	0.5	6.9	16.5	14.3	18.3	15.2	13.1	15.6	13.5

Continued on next page

Table 15 – *Continued from previous page*

n	p	ρ	δ	T_{SR}	T_{SG1}	T_{SG2}	T_{SG3}	T_{3c1p}	T_{3c1m}	T_{3c2p}	T_{3c2m}
			0.75	19.5	55.4	46.2	53	53.3	43.8	55.0	44.6
			1.0	55.9	93.0	82.5	90.4	92.0	83.7	92.8	84.2
	100	0.2	0.5	22.0	95.8	57.5	77.8	84.7	25.3	95.1	36.7
			0.75	99.7	100.0	99.8	100.0	100.0	99.7	100.0	99.9
			1.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		0.5	0.5	8.4	38.2	20.9	32.7	32.7	16.5	37.1	18.6
			0.75	54.1	100.0	82.8	95.9	99.9	75.4	100.0	78.3
			1.0	99.0	100.0	99.5	100.0	100.0	99.5	100.0	99.8
		0.8	0.5	7.6	17.7	15.9	18.3	16.8	14.0	17.3	15.5
			0.75	25.5	66.0	53.3	63.1	63.7	50.1	64.9	51.1
			1.0	69.5	97.6	88.7	94.9	97.1	88.9	97.7	90.0
	500	0.2	0.5	28.3	99.1	68.9	89.8	93.2	31.7	98.7	45.7
			0.75	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
			1.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		0.5	0.5	8.8	40.6	22.9	36.5	36.3	17.3	40.2	19.4
			0.75	58.0	99.8	86.5	96.4	99.7	80.0	99.8	82.6
			1.0	99.7	100.0	99.7	100.0	100.0	99.8	100.0	99.8
		0.8	0.5	7.9	17.5	15.1	18.7	16.6	14.3	17.7	15.0
			0.75	25.7	66.2	54.6	63.8	64.6	52.3	65.9	53.5
			1.0	72.3	98.7	90.9	97.1	98.6	91.5	98.7	91.8

Acknowledgments

A part of calculations for simulation studies was made at the Poznań Supercomputing and Networking Center.

Supplementary Material

R code

(doi: 10.1214/22-EJS2033SUPP; .zip). This supplement contains the codes to perform the new tests and to conduct the numerical experiments of Section 5.

References

[1] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed byoligonucleotide arrays. *The Proceedings of the National Academy of Sciences* **96** 6745–6750.

[2] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6** 311–329. MR1399305

[3] Bickel, P. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics* **40** 1–23. MR0256519

[4] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B. and Smola, A. (2006). Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics (ISMB)* **22** e49–e57.

[5] Chen, F., Meintanis, S. G. and Zhu, L. (2019). On some Characterizations and Multidimensional Criteria for Testing Homogeneity, Symmetry and Independence. *Journal of Multivariate Analysis* **173** 125–144. MR3920999

[6] Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics* **38** 808–835. MR2604697

[7] Dong, K., Pang, H., Tong, T. and Genton, M. (2016). Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. *Journal of Multivariate Analysis* **143** 127–142. MR3431423

[8] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.

[9] Friedman, J. and Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* **7** 697–717. MR0532236

[10] Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B. and Smola, A. J. (2007). A kernel approach to comparing distributions. *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)* 1637–1641.

[11] Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B. and Smola, A. J. (2007). A Kernel Method for the Two-Sample-Problem In *Advances in Neural Information Processing Systems 15* 513–520. MIT Press.

[12] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research* **13** 723–773. MR2913716

[13] Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B. K. (2009). A Fast, Consistent Kernel Two-Sample Test. In *Advances in Neural Information Processing Systems 22* 673–681. Curran Associates, Inc.

[14] Hotelling, H. (1951). A generalized t test and measure of multivariate dispersion. *In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* 23–41. MR0044798

[15] Reddi, S., Ramdas, A., Poczos, B., Singh, A. and Wasserman, L. (2015). On the high dimensional power of a linear-time two sample test under mean-shift alternatives. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics* **38** 772–780.

[16] Rizzo, M. and Székely, G. (2019). energy: E-Statistics: Multivariate Inference via the Energy of Data R package version 1.7-6.

[17] Sarkar, S. and Ghosh, A. K. (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat* **7** e187. MR3816902

[18] Schölkopf, B., Tsuda, K. and Vert, J. (2004). *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA.

[19] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

[20] Shawe-Taylor, J., Williams, C., Cristianini, N. and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalisation error of kernel PCA. *IEEE Trans. Inf. Theory* **51** 2510–2522. MR2246374

[21] Smola, A., Gretton, A., Song, L. and Schölkopf, B. (2007). A Hilbert Space Embedding for Distributions. *In Proceedings of the International Conference on Algorithmic Learning Theory* **4754** 13–31.

[22] Székely, G. J. and Rizzo, M. L. (2004). Testing for Equal Distributions in High Dimension. *InterStat* **November**.

[23] R Core Team (2021). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

[24] Zhang, J.-T. (2005). Approximate and Asymptotic Distributions of Chi-Squared-Type Mixtures With Applications. *Journal of the American Statistical Association* **100** 273–285. MR2156837

[25] Zhang, J.-T. (2013). *Analysis of variance for functional data*. CRC Press. MR3185072

[26] Zhang, J.-T., Guo, J., Zhou, B. and Cheng, M.-Y. (2020). A simple two-sample test in high dimensions based on L^2-norm. *Journal of American Statistical Association* **115** 1011–1027. MR4107696