

On some graph-based two-sample tests for high dimension, low sample size data

Soham Sarkar¹ · Rahul Biswas² · Anil K. Ghosh³

Received: 18 March 2019 / Revised: 11 July 2019 / Accepted: 22 October 2019 /

Published online: 13 November 2019

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2019

Abstract

Testing for equality of two high-dimensional distributions is a challenging problem, and this becomes even more challenging when the sample size is small. Over the last few decades, several graph-based two-sample tests have been proposed in the literature, which can be used for data of arbitrary dimensions. Most of these test statistics are computed using pairwise Euclidean distances among the observations. But, due to concentration of pairwise Euclidean distances, these tests have poor performance in many high-dimensional problems. Some of them can have powers even below the nominal level when the scale-difference between two distributions dominates the location-difference. To overcome these limitations, we introduce some new dissimilarity indices and use them to modify some popular graph-based tests. These modified tests use the distance concentration phenomenon to their advantage, and as a result, they outperform the corresponding tests based on the Euclidean distance in a wide variety of examples. We establish the high-dimensional consistency of these modified tests under fairly general conditions. Analyzing several simulated as well as real data sets, we demonstrate their usefulness in high dimension, low sample size situations.

Keywords Distance concentration \cdot High-dimensional consistency \cdot Minimum spanning tree \cdot Nearest neighbor \cdot Non-bipartite matching \cdot Permutation test \cdot Shortest Hamiltonian path

Editor: Karsten Borgwardt.

Soham Sarkar soham.sarkar@epfl.ch

Rahul Biswas rbiswas1@uw.edu

Anil K. Ghosh akghosh@isical.ac.in

- Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland
- Department of Statistics, University of Washington, Seattle, WA 98195, USA
- Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India



1 Introduction

Let $\mathcal{X}_m = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ and $\mathcal{Y}_n = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ be two sets of independent observations from d-dimensional continuous distributions F and G, respectively. In the two-sample problem, we use these observations to test the null hypothesis $\mathcal{H}_0 : F = G$ against the alternative hypothesis $\mathcal{H}_A : F \neq G$. This problem is well-investigated, and several tests are available for it. Interestingly, many of these tests are based on pairwise Euclidean distances. Under some mild conditions, Maa et al. (1996) showed that for $\mathbf{X}_1, \mathbf{X}_2 \overset{i.i.d.}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{i.i.d.}{\sim} G$, $\|\mathbf{X}_1 - \mathbf{X}_2\|$, $\|\mathbf{Y}_1 - \mathbf{Y}_2\|$ and $\|\mathbf{X}_1 - \mathbf{Y}_1\|$ have the same distribution if and only if F and G are identical. So, pairwise Euclidean distances contain useful information about the difference between two distributions. Also, these distances can be easily computed in any dimension. Because of these reasons, pairwise Euclidean distances have been extensively used for the construction of two-sample tests, which are applicable to high dimension, low sample size (HDLSS) data.

Existing tests based on pairwise Euclidean distances can be broadly categorized into two groups: (i) tests based on averages of three types (**XX**, **XY** and **YY**) of pairwise distances and (ii) tests based on graphs. Tests based on averages of pairwise distances include Baringhaus and Franz (2004, 2010); Székely and Rizzo (2004, 2013); Aslan and Zech (2005); Gretton et al. (2012); Biswas and Ghosh (2014); Sarkar and Ghosh (2018); Tsukada (2019).

Almost all graph-based tests consider an edge-weighted complete graph $\mathcal G$ on the vertex set $\mathcal{Z}_N = \mathcal{X}_m \cup \mathcal{Y}_n$ (here N = m + n is the total sample size), where the Euclidean distance between two vertices is taken to be the weight associated with the edge connecting them. Different tests consider different sub-graphs of \mathcal{G} and look at their topologies. The deviation of the topology of a sub-graph from the one expected under \mathcal{H}_0 is used to construct the test statistic. Friedman and Rafsky (1979) used the minimum spanning tree (MST) of \mathcal{G} as the sub-graph to construct the Kolmogorov-Smirnov test and the Wald-Wolfowitz run test for multivariate data. Biswas et al. (2014) used the shortest Hamiltonian path (SHP) on \mathcal{G} to develop another multivariate run test. Rosenbaum (2005) constructed the cross-match test using $\lfloor N/2 \rfloor$ disconnected edges of \mathcal{G} (here $\lfloor t \rfloor$ is the largest integer $\leq t$) for which the total edge weight is minimum. Liu and Modarres (2011) considered all cliques of size 3 to construct their test statistic. The tests based on nearest-neighbor type coincidences (see, e.g., Schilling 1986; Henze 1988; Mondal et al. 2015) can be viewed as tests based on directed subgraphs of \mathcal{G} , whereas the test by Hall and Tajvidi (2002) can be viewed as an aggregation of tests based on sub-graphs for different levels of neighborhood. Recently, Chen and Friedman (2017) also constructed some two-sample tests using graph-theoretic ideas.

In this article, we will mainly consider four tests, namely, the test based on nearest-neighbors (Schilling 1986; Henze 1988), the multivariate run test based on MST (Friedman and Rafsky 1979), the multivariate run test based on SHP (Biswas et al. 2014) and the cross-match test based on optimal non-bipartite matching (Rosenbaum 2005) for our investigation. We will refer to these four tests as the NN test, the MST-run test, the SHP-run test and the NBP test, respectively. Brief descriptions of these tests are given below.

NN test (Schilling 1986; Henze 1988): Consider the edge-weighted complete graph \mathcal{G} on the vertex set \mathcal{Z}_N , as discussed above. Assume that an undirected edge (\mathbf{u}, \mathbf{v}) in \mathcal{G} corresponds to two directed edges $(\overline{\mathbf{u}}, \overline{\mathbf{v}})$ and $(\overline{\mathbf{v}}, \overline{\mathbf{u}})$. For a fixed k < N, consider the sub-graph \mathcal{T}_k , which contains an edge $(\overline{\mathbf{u}}, \overline{\mathbf{v}})$ if and only if \mathbf{v} is among the first k nearest-neighbors (in terms of the Euclidean distance) of \mathbf{u} . So, \mathcal{T}_k contains Nk directed edges. The NN test uses the test statistic $T_{NN} = \frac{1}{Nk} \sum_{(\overline{\mathbf{u}}, \overline{\mathbf{v}}) \in \mathcal{T}_k} \mathbb{I}(\mathbf{u}, \overline{\mathbf{v}})$, where $\mathbb{I}(\mathbf{u}, \overline{\mathbf{v}})$ is an indicator variable that takes the value 1 if \mathbf{u} and \mathbf{v} are from the same distribution. It rejects \mathcal{H}_0 for large values of T_{NN} . A more familiar



expression of this test statistic is $T_{NN} = \frac{1}{Nk} \left[\sum_{i=1}^{m} \sum_{r=1}^{k} \mathbb{I}_r(\mathbf{x}_i) + \sum_{i=1}^{n} \sum_{r=1}^{k} \mathbb{I}_r(\mathbf{y}_i) \right]$, where $\mathbb{I}_r(\mathbf{z})$ is an indicator variable that takes the value 1 if \mathbf{z} and its r-th nearest-neighbor (in terms of the Euclidean distance) come from the same distribution.

MST-run test (Friedman and Rafsky 1979): Unlike the NN test, this test is based on an undirected sub-graph of \mathcal{G} . Let \mathcal{M} be the minimum spanning tree (MST) of \mathcal{G} . The MST-run test uses the test statistic $T_{MST} = 1 + \sum_{i=1}^{N-1} \lambda_i^{\mathcal{M}}$, where $\lambda_i^{\mathcal{M}}$ is an indicator variable that takes the value 1 if and only if the *i*-th edge (i = 1, ..., N-1) of \mathcal{M} connects two observations from different distributions. The null hypothesis \mathcal{H}_0 is rejected for small values of T_{MST} .

SHP-run test (Biswas et al. 2014): Instead of MST, it uses the shortest Hamiltonian path (SHP). Let S be the SHP on G. The number of runs along S is computed as $T_{SHP} = 1 + \sum_{i=1}^{N-1} \lambda_i^S$, where the indicator λ_i^S takes the value 1 if and only if the i-th edge of S connects two observations from different distributions. The SHP-run test rejects \mathcal{H}_0 for small values of T_{SHP} .

NBP test (Rosenbaum 2005): It uses the optimal non-bipartite matching algorithm (see, e.g., Lu et al. 2011) to find $\lfloor N/2 \rfloor$ disconnected edges (i.e., no two edges share a common vertex) in \mathcal{G} such that the total weight of the edges is minimum. Let $\mathcal{C} = \{(\mathbf{u}_i, \mathbf{v}_i) : i = 1, \ldots, \lfloor N/2 \rfloor \}$ be the collection of these edges. The NBP test rejects \mathcal{H}_0 for small values of the test statistic $T_{NBP} = \sum_{i=1}^{\lfloor N/2 \rfloor} \lambda_i^{\mathcal{C}}$, where $\lambda_i^{\mathcal{C}}$ is an indicator variable that takes the value 1 if and only if \mathbf{u}_i and \mathbf{v}_i are from two different distributions.

These tests based on pairwise Euclidean distances are known to work well in many highdimensional problems even when the dimension is much larger than the sample size. But they also have some limitations. To demonstrate this, we consider the following examples.

Example 1 F and G are Gaussian with the same mean $\mathbf{0}_d = (0, \dots, 0)^{\top}$ and diagonal dispersion matrices $\Lambda_{1,d}$ and $\Lambda_{2,d}$, respectively. The first $\lfloor d/2 \rfloor$ diagonal elements of $\Lambda_{1,d}$ are 1 and the rest are 2, whereas for $\Lambda_{2,d}$, the first $\lfloor d/2 \rfloor$ diagonal elements are 2 and the rest are 1.

Example 2 For $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})^{\top} \sim F$, $X^{(1)}, \dots, X^{(d)}$ are i.i.d. $\mathcal{N}(0, 5)$, while for $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(d)})^{\top} \sim G$, $Y^{(1)}, \dots, Y^{(d)}$ are i.i.d. $t_5(0, 3)$. Here $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 , and $t_{\nu}(\mu, \sigma^2)$ denotes the Student's t-distribution with ν degrees of freedom, location μ and scale σ .

For each of these examples, we carried out our experiment with $d=2^i$ for $i=1,\ldots,10$. For different values of d, we generated 20 observations from each distribution and used them to test $\mathcal{H}_0: F=G$ against $\mathcal{H}_A: F\neq G$. We repeated each experiment 500 times and estimated the power of a test by the proportion of times it rejected \mathcal{H}_0 . These estimated powers are reported in Fig. 1. The SHP-run test and the NBP test are distribution-free. For the other two tests, we used conditional tests based on 1000 random permutations. For the NN test, we used k=3 for all numerical work since it has been reported to perform well in the literature (see, e.g., Schilling 1986). Throughout this article, all tests are considered to have 5% nominal level.

Figure 1 clearly shows that all these tests had poor performance in Examples 1 and 2. Note that in both of these examples, each measurement variable has different distributions under F and G. So, each of them carries a signal against \mathcal{H}_0 . Therefore, the power of a reasonable test is expected to increase to one as the dimension increases. But we did not observe that for these tests based on pairwise Euclidean distances.

The reason behind the poor performance by these tests can be explained using the *distance concentration* phenomenon in high dimensions. To understand it, let us consider four



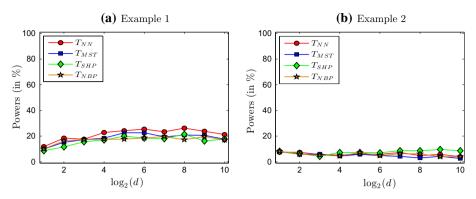


Fig. 1 Powers of NN, MST-run, SHP-run and NBP tests in Examples 1 and 2

independent random vectors $\mathbf{X}_1, \mathbf{X}_2 \sim F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \sim G$, where F and G have mean vectors $\boldsymbol{\mu}_F, \boldsymbol{\mu}_G$ and dispersion matrices $\boldsymbol{\Sigma}_F, \boldsymbol{\Sigma}_G$, respectively. We also make the following assumptions.

Assumption 1 For $X \sim F$ and $Y \sim G$, fourth moments of the component variables of X and Y are uniformly bounded.

Assumption 2 For $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2$, $\mathbf{Y}_1 - \mathbf{Y}_2$ and $\mathbf{X}_1 - \mathbf{Y}_1$, $\sum_{1 \le q \ne q' \le d} \operatorname{corr} \{ (W^{(q)})^2, (W^{(q')})^2 \}$ is of the order $\mathbf{o}(d^2)$.

Assumption 3 There exist non-negative constants v^2 , σ_F^2 and σ_G^2 such that $d^{-1}\|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|^2 \to v^2$, $d^{-1}trace(\boldsymbol{\Sigma}_F) \to \sigma_F^2$ and $d^{-1}trace(\boldsymbol{\Sigma}_G) \to \sigma_G^2$ as $d \to \infty$.

These assumptions are quite common in the HDLSS asymptotic literature (see, e.g., Hall et al. 2005; Jung and Marron 2009; Biswas et al. 2014; Dutta et al. 2016). Under Assumptions 1 and 2, the weak law of large numbers (see, e.g., Billingsley 1995) holds for the sequence of possibly dependent and non-identically distributed random variables $\{(W^{(q)})^2: q \geq 1\}$, i.e., $|d^{-1}\sum_{i=1}^d (W^{(q)})^2 - d^{-1}\sum_{i=1}^d E((W^{(q)})^2)| \stackrel{Pr}{\to} 0$ as $d \to \infty$. Assumption 3 gives the limiting value of $d^{-1}\sum_{i=1}^d E((W^{(q)})^2)$. These three assumptions hold in Examples 1 and 2, and under these assumptions, we have the following result on the high-dimensional behavior of pairwise Euclidean distances. The proof follows from the discussions in Sections 3.1 and 3.2 of Hall et al. (2005).

Lemma 1 (Hall et al. 2005) Suppose that $\mathbf{X}_1, \mathbf{X}_2 \sim F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \sim G$ are independent random vectors. If F and G satisfy Assumptions 1–3, then $d^{-1/2} \|\mathbf{X}_1 - \mathbf{X}_2\| \stackrel{Pr}{\to} \sigma_F \sqrt{2}$, $d^{-1/2} \|\mathbf{Y}_1 - \mathbf{Y}_2\| \stackrel{Pr}{\to} \sigma_G \sqrt{2}$ and $d^{-1/2} \|\mathbf{X}_1 - \mathbf{Y}_1\| \stackrel{Pr}{\to} \sqrt{\sigma_F^2 + \sigma_G^2 + v^2}$ as d tends to infinity.

Under Assumptions 1–3, using Lemma 1, Biswas et al. (2014) proved the high-dimensional consistency (i.e., the convergence of power to 1 as d tends to infinity) of the SHP-run test when $v^2 > 0$ or $\sigma_F^2 \neq \sigma_G^2$. Under the same condition, one can show this consistency for the NBP test as well (follows using arguments similar to those used in the proof of part (b) of Theorem 1). When $v^2 > |\sigma_F^2 - \sigma_G^2|$, such high-dimensional consistency can also be proved for the NN test (follows using arguments similar to those used in the proof of part (a) of Theorem 2) and the MST-run test (see Biswas et al. 2014). However, in Examples 1 and 2, we have $v^2 = 0$ and $\sigma_F^2 = \sigma_G^2$. So, $d^{-1/2} ||\mathbf{X}_1 - \mathbf{X}_2||$, $d^{-1/2} ||\mathbf{Y}_1 - \mathbf{Y}_2||$ and $d^{-1/2} ||\mathbf{X}_1 - \mathbf{Y}_1||$



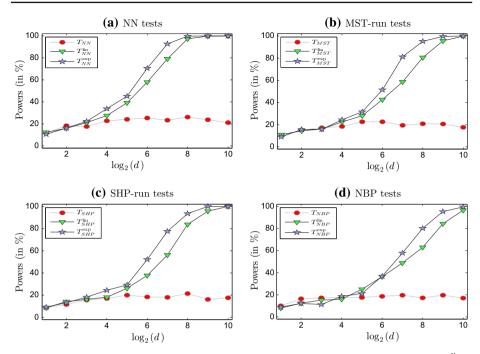


Fig. 2 Powers of different graph-based tests and their modified versions based on $\varphi_{h,\psi}$ in Example 1. T^{lin} & T^{exp} correspond to tests based on $\psi_1(t) = t$ and $\psi_2(t) = 1 - \exp(-t)$, respectively

all converge to the same value (follows from Lemma 1). Therefore, tests based on pairwise Euclidean distances fail to capture the difference between two underlying distributions and have poor performance.

Recently, Sarkar and Ghosh (2018) identified this problem for tests based on averages of pairwise Euclidean distances. To cope with such situations, instead of the Euclidean distance, they suggested to use distance functions of the form $\varphi_{h,\psi}(\mathbf{u},\mathbf{v}) = h\{d^{-1}\sum_{q=1}^d \psi(|u^{(q)}-v^{(q)}|)\}$ for suitably chosen strictly increasing functions $h,\psi:[0,\infty)\to [0,\infty)$ with $h(0)=\psi(0)=0$. Now, one may be interested to know what happens to the graph-based tests if they are constructed using $\varphi_{h,\psi}$ (i.e., the edge-weights in $\mathcal G$ are defined using $\varphi_{h,\psi}$). For this investigation, here we consider two choices of ψ , namely, $\psi_1(t)=t$ and $\psi_2(t)=1-\exp(-t)$, with h(t)=t in both cases. Figures 2 and 3 show powers of these modified tests in Examples 1 and 2 (curves corresponding to T^{lin} and T^{exp} represent the tests based on ψ_1 and ψ_2 , respectively). These tests had excellent performance in Example 1. Their powers converged to one as the dimension increased. Modified versions of SHP-run test had similar behavior in Example 2 as well. In this example, powers of modified NBP tests also increased with the dimension, but those of modified NN and MST-run tests dropped down to zero as the dimension increased. In the next section, we investigate the reasons behind such contrasting behavior of these tests based on $\varphi_{h,\psi}$.



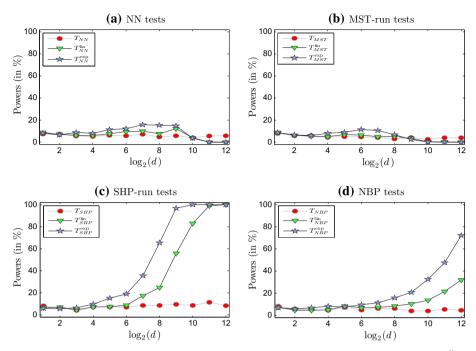


Fig. 3 Powers of different graph-based tests and their modified versions based on $\varphi_{h,\psi}$ in Example 2. T^{lin} & T^{exp} correspond to tests based on $\psi_1(t) = t$ and $\psi_2(t) = 1 - \exp(-t)$, respectively

2 High-dimensional behavior of the tests based on $\phi_{h,\psi}$

In this section, we carry out a theoretical investigation on the high-dimensional behavior of the modified tests based on $\varphi_{h,\psi}$. For this investigation, we make the following assumption.

Assumption 4 Let $\mathbf{X}_1, \mathbf{X}_2 \sim F, \mathbf{Y}_1, \mathbf{Y}_2 \sim G$ be independent random vectors. The function $\psi: [0, \infty) \to [0, \infty)$ satisfies $\psi(0) = 0$, and for $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2, \mathbf{Y}_1 - \mathbf{Y}_2$ and $\mathbf{X}_1 - \mathbf{Y}_1, d^{-1} \sum_{g=1}^d \left\{ \psi(|W^{(g)}|) - \mathbb{E}\psi(|W^{(g)}|) \right\} \stackrel{\Pr}{\to} 0$ as $d \to \infty$.

Recall that Assumptions 1 and 2 lead to probability convergence of $d^{-1}\sum_{d=1}\{(W^{(q)})^2 - E((W^{(q)})^2)\}$. So, Assumption 4 can be viewed as a generalization of Assumptions 1 and 2 with $(W^{(q)})^2$ replaced by $\psi(|W^{(q)}|)$. Note that Assumption 1 holds if and only if the second moments of $\{(W^{(q)})^2: q \geq 1\}$ are uniformly bounded. So, if the random variables $\{\psi(|W^{(q)}|): q \geq 1\}$ have uniformly bounded second moments (which are trivial if ψ is bounded), and they are weakly dependent in the sense that $\sum_{1\leq q\neq q'\leq d} \mathrm{corr}\{\psi(|W^{(q)}|),\psi(|W^{(q')}|)\} = \mathbf{o}(d^2)$, Assumption 4 holds. In particular, it holds if the sequence $\{\psi(|W^{(q)}|): q \geq 1\}$ is m-dependent (see, e.g., Billingsley 1995, page 90) or ρ -mixing (see, e.g., Lin and Lu 1996, page 4). However, Assumption 4 holds in many other situations. For instance, some sufficient conditions for mixingale sequence of random variables have been derived by Andrews (1988) and de Jong (1995). Dutta et al. (2016, Section 7) also made some assumptions for deriving the week law for the sequence $\{(W^{(q)})^2: q \geq 1\}$. Similar assumptions with appropriate modifications for ψ also lead to Assumption 4.



If h is uniformly continuous (which we will assume throughout this article, unless otherwise mentioned), under Assumption 4, $\{\varphi_{h,\psi}(\mathbf{X}_1,\mathbf{X}_2) - \varphi_{h,\psi}^*(F,F)\}$, $\{\varphi_{h,\psi}(\mathbf{Y}_1,\mathbf{Y}_2) - \varphi_{h,\psi}^*(G,G)\}$ and $\{\varphi_{h,\psi}(\mathbf{X}_1,\mathbf{Y}_1) - \varphi_{h,\psi}^*(F,G)\}$ converge in probability to 0 as d tends to infinity, where

$$\begin{split} \varphi_{h,\psi}^*(F,F) &= h \Big\{ d^{-1} \sum_{q=1}^d \mathrm{E} \psi(|X_1^{(q)} - X_2^{(q)}|) \Big\}, \\ \varphi_{h,\psi}^*(G,G) &= h \Big\{ d^{-1} \sum_{q=1}^d \mathrm{E} \psi(|Y_1^{(q)} - Y_2^{(q)}|) \Big\} \quad \text{and} \\ \varphi_{h,\psi}^*(F,G) &= h \Big\{ d^{-1} \sum_{q=1}^d \mathrm{E} \psi(|X_1^{(q)} - Y_1^{(q)}|) \Big\}. \end{split} \tag{1}$$

An interesting lemma involving these three quantities is stated below.

Lemma 2 (Sarkar and Ghosh 2018, Lemma 1) Suppose that h is a strictly increasing, concave function and $\psi'(t)/t$ is a non-constant, monotone function. Then, for any fixed $d \geq 1$, $e_{h,\psi}(F,G) = 2\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,F) - \varphi_{h,\psi}^*(G,G) \geq 0$, where the equality holds if and only if F and G have the same univariate marginal distributions.

The quantity $e_{h,\psi}(F,G)$ can be viewed as an energy distance between the distributions F and G (see, e.g., Székely and Rizzo 2004; Aslan and Zech 2005), and it serves as a measure of separation between the two distributions. Lemma 2 shows that for every $d \geq 1$, $e_{h,\psi}(F,G)$ is positive unless F and G have the same univariate marginals. For deriving high-dimensional results, we assume that $\widetilde{e}_{h,\psi}(F,G) = \liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$, which ensures that the energy distance between the two populations is asymptotically non-negligible. The following theorem shows the high-dimensional consistency of SHP-run and NBP tests based on $\varphi_{h,\psi}$ under this assumption.

Theorem 1 Let $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim G$ (where m+n=N) be independent random vectors, where F and G satisfy Assumption 4 with $\widetilde{e}_{h,\psi}(F,G) = \liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$.

- (a) If $N/\binom{N}{m} < \alpha$, then the power of the SHP-run test (of level α) based on $\varphi_{h,\psi}$ converges to 1 as d tends to infinity.
- (b) If $c(m, n) < \alpha$, then the power of the NBP test (of level α) based on $\varphi_{h,\psi}$ converges to 1 as d tends to infinity. Here, c(m, n) is given by

$$c(m,n) = \begin{cases} \frac{(N/2)!}{\binom{N}{m}(m/2)!(n/2)!}, & \text{if both } m, n \text{ are even} \\ \\ \frac{2(N/2)!}{\binom{N}{m}((m-1)/2)!((n-1)/2)!}, & \text{if both } m, n \text{ are odd} \\ \\ \frac{((N-1)/2)!}{\binom{N-1}{m}(m/2)!((n-1)/2)!}, & \text{if } m \text{ is even and } n \text{ is odd} \\ \\ \frac{((N-1)/2)!}{\binom{N-1}{m-1}((m-1)/2)!(n/2)!}, & \text{if } m \text{ is odd and } n \text{ is even} \end{cases}$$

Theorem 1 shows that if $\tilde{e}_{h,\psi}(F,G) > 0$, then SHP-run and NBP tests based on $\varphi_{h,\psi}$ have the high-dimensional consistency if the sample sizes are not too small. In view of



Lemma 2, for our two choices of h and ψ , we have $\widetilde{e}_{h,\psi}(F,G)>0$ in Examples 1 and 2. This was the reason behind the excellent performance by these tests. However, for the tests based on the Euclidean distance (i.e., when $h(t)=\sqrt{t}$ and $\psi(t)=t^2$), we have $\widetilde{e}_{h,\psi}(F,G)=2\sqrt{\sigma_F^2+\sigma_G^2+\nu^2}-\sigma_F\sqrt{2}-\sigma_G\sqrt{2}$ (follows from Lemma 1), which is positive if and only if $\nu^2>0$ or $\sigma_F^2\neq\sigma_G^2$. This was not the case in Examples 1 and 2, where all tests based on the Euclidean distance had poor performance.

For the high-dimensional consistency of NN and MST-run tests based on $\varphi_{h,\psi}$, we need some additional conditions, as shown by the following theorem.

Theorem 2 Let $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim G$ (where m+n=N) be independent random vectors, where F and G satisfy Assumption 4. Also assume that both $\liminf_{d\to\infty} \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,F)\}$ and $\liminf_{d\to\infty} \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(G,G)\}$ are positive.

- (a) Define $N_0 = \lceil N/(k+1) \rceil$ and $m_0 = \lceil \min\{m, n\}/(k+1) \rceil$ (here $\lceil t \rceil$ denotes the smallest integer larger than or equal to t). If $k < \min\{m, n\}$ and $\binom{N_0}{m_0} < \alpha \binom{N}{m}$, then the power of the NN test (of level α) based on $\varphi_{h,\psi}$ converges to 1 as d tends to infinity.
- (b) If $\max\{\lfloor N/m \rfloor, \lfloor N/n \rfloor\} < \alpha \binom{N}{m}$, then the power of the MST-run test (of level α) based on $\varphi_{h,\psi}$ converges to 1 as d tends to infinity.

In Theorems 1 and 2, for the high-dimensional consistency of different two-sample tests, we need some conditions on sample sizes. These conditions are satisfied if m and n are not too small. For instance, for $\alpha = 0.05$, we have the consistency for NN and MST-run tests when $m, n \geq 4$. For the SHP-run test and the NBP test, we need $m, n \geq 5$ and $m, n \geq 8$, respectively. For smaller values of α , we need larger sample sizes. To have a better understanding on the requirement of sample sizes, for different choices of m and n, we have computed the minimum value of α for which the resulting test turns out to be consistent. These minimum values are plotted in Fig. 4. It shows that for NN, MST-run and SHP-run tests, the minimum value of α drops monotonically as the sample sizes increase. For the NBP test, we see two different type of results corresponding to even and odd sample sizes. However, for $m, n \geq 10$, the minimum value of α is very close to zero for all of these tests.

The conditions $\liminf_{d\to\infty} \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,F)\} > 0$ and $\liminf_{d\to\infty} \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(G,G)\} > 0$ assumed in Theorem 2 ensure that the neighborhood structure (in terms of $\varphi_{h,\psi}$) is preserved in high dimensions, i.e., an observation and its nearest neighbor (in terms of $\varphi_{h,\psi}$) come from the same distribution with high probability. For the Euclidean distance (i.e., for $h(t) = \sqrt{t}$ and $\psi(t) = t^2$), these conditions are equivalent to $\nu^2 > |\sigma_F^2 - \sigma_G^2|$ (follows from Lemma 1), i.e., the location-difference between two distributions dominates the overall scale-difference. In Example 1, from the descriptions of F and G, it is easy to check that $\lim_{d\to\infty} \varphi_{h,\psi}^*(F,F) = \lim_{d\to\infty} \varphi_{h,\psi}^*(G,G)$. So, in view of Lemma 2, we have $\lim_{d\to\infty} \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,F)\} > 0$ and $\lim_{d\to\infty} \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,G)\} > 0$. But in Example 2, one can show that $\varphi_{h,\psi}^*(F,G)$ lies between $\varphi_{h,\psi}^*(F,F)$ and $\varphi_{h,\psi}^*(G,G)$. This violation of neighborhood structure led to poor performance by NN and MST-run tests based on $\varphi_{h,\psi}$. The following theorem shows that in such situations, powers of these tests may even drop down to zero.

Theorem 3 Let $X_1, ..., X_m \sim F$ and $Y_1, ..., Y_n \sim G$ be independent random vectors, where F and G satisfy Assumption 4. Also assume that $\limsup_{d\to\infty} \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,F)\} < 0$ (interchange F and G if required, and in that case, interchange F and F accordingly).

(a) If $k < \min\{m, n\}$ and $(m - 1)/n > (1 + \alpha)/(1 - \alpha)$, then the power of the NN test (of level α) based on $\varphi_{h,\psi}$ converges to 0 as d tends to infinity.



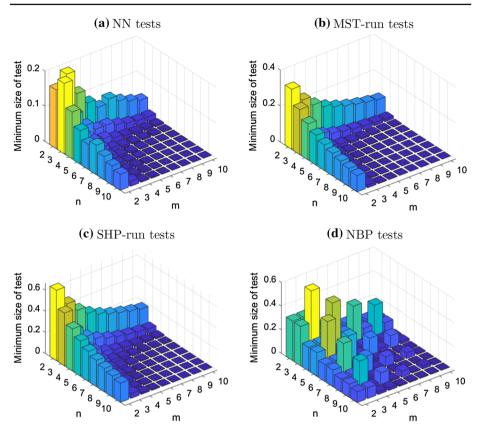


Fig. 4 Minimum size of a consistent test for different sample sizes

(b) If $m/n > (1+\alpha)/(1-\alpha)$, then the power of the MST-run test (of level α) based on $\varphi_{h,\psi}$ converges to 0 as d tends to infinity.

The conditions involving m and n in Theorem 3 are only sufficient. Note that these conditions do not hold in Example 2, but NN and MST-run tests based on $\varphi_{h,\psi}$ had powers close to 0. To overcome the limitations of NN and MST-run tests, in the next section, we introduce a new class of dissimilarity measures called MADD and use it for further modification of the tests.

3 Further modification of tests based on MADD

From the discussion in the previous section, it is clear that in order to have good performance by NN and MST-run tests in high dimensions, we need to use a distance function or dissimilarity measure which preserves the neighborhood structure. To achieve this, we define the dissimilarity between two observations \mathbf{x} and \mathbf{y} in \mathbb{Z}_N as

$$\rho_{h,\psi}(\mathbf{x}, \mathbf{y}) = \frac{1}{N - 2} \sum_{\mathbf{z} \in \mathcal{Z}_N \setminus \{\mathbf{x}, \mathbf{y}\}} \left| \varphi_{h,\psi}(\mathbf{x}, \mathbf{z}) - \varphi_{h,\psi}(\mathbf{y}, \mathbf{z}) \right|, \tag{2}$$



where $\varphi_{h,\psi}$ is as defined in Sect. 1. Since this dissimilarity index is based on the Mean of Absolute Differences of pairwise Distances, we call it MADD (see also Sarkar and Ghosh 2019). Using $h(t) = \sqrt{t}$ and $\psi(t) = t^2$, we get MADD based on the Euclidean distance, which is defined as

$$\rho_0(\mathbf{x}, \mathbf{y}) = \frac{1}{N - 2} \sum_{\mathbf{z} \in \mathcal{Z}_N \setminus \{\mathbf{x}, \mathbf{y}\}} d^{-1/2} \big| \|\mathbf{x} - \mathbf{z}\| - \|\mathbf{y} - \mathbf{z}\| \big|.$$
(3)

MADD has several desirable properties. One such property is mentioned below.

Lemma 3 For $N \geq 3$, the dissimilarity index $\rho_{h,\psi}$ is a semi-metric on \mathcal{Z}_N .

The index $\rho_{h,\psi}$ is not a metric since $\rho_{h,\psi}(\mathbf{x},\mathbf{y}) = 0$ does not necessarily imply $\mathbf{x} = \mathbf{y}$. However, if F and G are absolutely continuous distributions, then for any $\mathbf{x} \neq \mathbf{y}$, $\rho_{h,\psi}(\mathbf{x},\mathbf{y})$ is strictly positive with probability 1. So, in practice, $\rho_{h,\psi}$ behaves like a metric. When $\varphi_{h,\psi}$ is a metric, using the triangle inequality, we also get $\rho_{h,\psi}(\mathbf{x},\mathbf{y}) \leq \varphi_{h,\psi}(\mathbf{x},\mathbf{y})$. So, closeness in terms of $\varphi_{h,\psi}$ indicates closeness in terms of $\rho_{h,\psi}$, but not the other way around. Note that if h is uniformly continuous, under Assumption 4, we have the probability convergence of $\left|\varphi_{h,\psi}(\mathbf{X}_1,\mathbf{X}_2) - \varphi_{h,\psi}^*(F,F)\right|, \left|\varphi_{h,\psi}(\mathbf{Y}_1,\mathbf{Y}_2) - \varphi_{h,\psi}^*(G,G)\right|$ and $\left|\varphi_{h,\psi}(\mathbf{X}_1,\mathbf{Y}_1) - \varphi_{h,\psi}^*(F,G)\right|$ to 0 as d tends to infinity. This leads to the probability convergence of $\rho_{h,\psi}(\mathbf{X}_1,\mathbf{X}_2)$, $\rho_{h,\psi}(\mathbf{Y}_1,\mathbf{Y}_2)$ and $\left\{\rho_{h,\psi}(\mathbf{X}_1,\mathbf{Y}_1) - \rho_{h,\psi}^*(F,G)\right\}$ to 0, where

$$\rho_{h,\psi}^*(F,G) = \frac{1}{N-2} \left\{ (m-1) \left| \varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,F) \right| + (n-1) \left| \varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(G,G) \right| \right\}.$$

So, in the case of high-dimensional data, unlike $\varphi_{h,\psi}$, the dissimilarity index $\rho_{h,\psi}$ usually takes small values for observations from the same distribution. The quantity $\rho_{h,\psi}^*(F,G)$ is non-negative. But, in order to preserve the neighborhood structure (in terms of $\rho_{h,\psi}$) in high dimensions, we need to choose the functions h and ψ such that $\rho_{h,\psi}^*(F,G)$ is strictly positive. The following lemma provides some guidance in this regard.

Lemma 4 Let $h, \psi : [0, \infty) \to [0, \infty)$ be strictly increasing functions such that $h(0) = \psi(0) = 0$ and $\psi'(t)/t$ is a non-constant, completely monotone function. Then, for every $d \ge 1$, $\rho_{h,h}^*(F,G)$ is positive unless F and G have the same univariate marginal distributions.

Figures 5 and 6 show the performance of NN and MST-run tests in Examples 1 and 2, when they were constructed based on $\rho_{h,\psi}$ for three different choices of h and ψ : (i) $h_0(t) = \sqrt{t}$, $\psi_0(t) = t^2$ (i.e., the Euclidean distance), (ii) $h_1(t) = t$, $\psi_1(t) = t$ and (iii) $h_2(t) = t$, $\psi_2(t) = 1 - \exp(-t)$. We denote the corresponding dissimilarity indices by ρ_0 , ρ_1 and ρ_2 , respectively. Though NN and MST-run tests based on ρ_0 could not perform well (see the curves corresponding to $\widetilde{T}_{NN}^{\rho_0}$ and $\widetilde{T}_{NST}^{\rho_2}$), those based on ρ_1 and ρ_2 had excellent performance (see the curves corresponding to $T_{NN}^{\rho_1}$, $\widetilde{T}_{NN}^{\rho_2}$ and $\widetilde{T}_{MST}^{\rho_1}$, respectively). Note that ψ_1 and ψ_2 satisfy the conditions stated in Lemma 4. So, for these two choices, we have $\rho_{h,\psi}^*(F,G) > 0$, and hence the neighborhood structure is preserved. This was the reason behind the excellent performance by these tests. The function ψ_0 does not satisfy the conditions of Lemma 4. For ρ_0 , the neighborhood structure is preserved when $v^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$. But, that was not the case in Examples 1 and 2.

Figures 5 and 6 also show the results for NBP and SHP-run tests based on $\rho_{h,\psi}$. Again, the test based on ρ_0 had poor performance, but those based on ρ_1 and ρ_2 performed well. In Example 2, NBP and SHP-run tests based on $\rho_{h,\psi}$ outperformed those



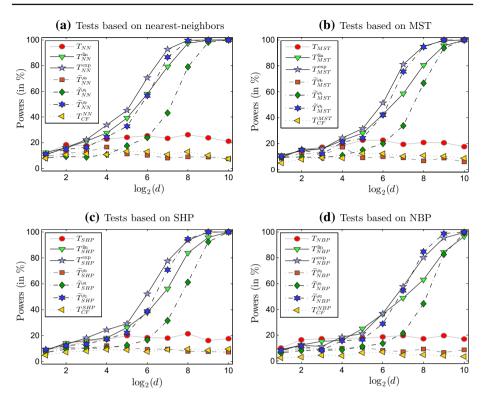


Fig. 5 Powers of different tests in Example 1

based on $\varphi_{h,\psi}$, but in Example 1, the tests based on $\varphi_{h,\psi}$ had an edge. Recall that for the high-dimensional consistency of NBP and SHP-run tests based on $\varphi_{h,\psi}$, we need $\widetilde{e}_{h,\psi}(F,G) = \liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$ (see Theorem 1). Note that $e_{h,\psi}(F,G)$ can be expressed as $e_{h,\psi}(F,G) = \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(F,F)\} + \{\varphi_{h,\psi}^*(F,G) - \varphi_{h,\psi}^*(G,G)\}$. Our empirical experience suggests that when both of these terms are positive, tests based on $\varphi_{h,\psi}$ usually have higher powers than those based on $\varphi_{h,\psi}$. But if they are of opposite signs, the tests based on $\varphi_{h,\psi}$ usually perform better. In Examples 1 and 2, we observed it not only for NBP and SHP-run tests, but for NN and MST-run tests as well.

Recently, Chen and Friedman (2017) developed a general framework to construct graph-based two-sample tests for multivariate data, where one counts the numbers of **XX**-type and **YY**-type edges (S_{xx} and S_{yy} , say) in a sub-graph of \mathcal{G} and computes their deviations from their expected values under \mathcal{H}_0 . The test rejects \mathcal{H}_0 for higher values of the statistic $T_{CF} = (\mathbf{S} - \mathbf{E}_{\mathcal{H}_0}(\mathbf{S}))^{\top} [\mathbf{var}_{\mathcal{H}_0}(\mathbf{S})]^{-1} (\mathbf{S} - \mathbf{E}_{\mathcal{H}_0}(\mathbf{S}))$, where $\mathbf{S} = (S_{xx}, S_{yy})^{\top}$. Chen and Friedman (2017) used k-nearest-neighbor graph (k-NN graph) and MST of \mathcal{G} for all numerical work. The k-NN graph is an undirected sub-graph of \mathcal{G} , which contains the edge (\mathbf{u} , \mathbf{v}) if either \mathbf{v} is among the k nearest-neighbors of \mathbf{u} or \mathbf{u} is among the k nearest neighbors of \mathbf{v} . We computed this statistic for the shortest Hamiltonian path and the optimal non-bipartite sub-graph of \mathcal{G} as well. Powers of these tests (henceforth, referred to as CF-NN, CF-MST, CF-SHP and CF-NBP tests, respectively) are also reported in Figs. 5 and 6. Just like the tests based on ρ_0 , these CF tests based on the Euclidean distance also had poor performance in these examples.



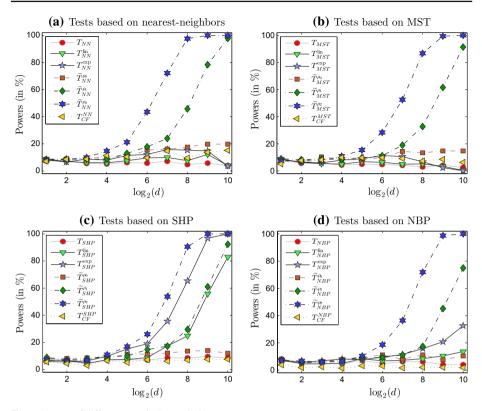


Fig. 6 Powers of different tests in Example 2

Now, to carry out a theoretical study on the high-dimensional behavior of the tests based on $\rho_{h,\psi}$, in view of Lemma 4, we make the following assumption.

Assumption 5
$$\widetilde{\rho}_{h,\psi}(F,G) = \liminf_{d\to\infty} \rho_{h,\psi}^*(F,G) > 0.$$

From the proof of Lemma 4, one can see that for any fixed d, $\rho_{h,\psi}^*(F,G)=0$ if and only if $e_{F,G}^{(q)}=2\mathrm{E}\psi(|X_1^{(q)}-Y_1^{(q)}|)-\mathrm{E}\psi(|X_1^{(q)}-X_2^{(q)}|)-\mathrm{E}\psi(|Y_1^{(q)}-Y_2^{(q)}|)=0$ for $q=1,\ldots,d$. This quantity $e_{F,G}^{(q)}$ is an energy distance between the q-th univariate marginals of F and G (see, e.g., Székely and Rizzo 2013) that contains signal against \mathcal{H}_0 . Now, $\widetilde{\rho}_{h,\psi}(F,G)$ becomes 0 only when $\lim\inf_{d\to\infty}d^{-1}\sum_{q=1}^d e_{F,G}^{(q)}=0$. So, Assumption 5 asserts that the average signal is asymptotically non-negligible. In classical asymptotic regime, we consider d to be fixed and expect to get more information as m and n increase. But, in the HDLSS asymptotic regime, where we consider m and n to be fixed, we expect to get more information as d increases. This is ensured by Assumptions 4 and 5. The following theorem shows the high-dimensional consistency of modified NN and MST-run tests based on $\rho_{h,\psi}$ under these assumptions.

Theorem 4 Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim G$ are independent random vectors, and $\rho_{h,\psi}$ is used to construct the test statistics, where h and ψ satisfy the conditions of Lemma 4. Then, under Assumptions 4 and 5, we get the following results.



- (a) Let N_0 and m_0 be defined as in Theorem 2. If $k < \min\{m, n\}$ and $\binom{N_0}{m_0} < \alpha \binom{N}{m}$, then the power of the NN test (of level α) based on $\rho_{h,\psi}$ converges to 1 as d tends to infinity.
- (b) If $\max\{\lfloor N/m \rfloor, \lfloor N/n \rfloor\} < \alpha \binom{N}{m}$, then the power of the MST-run test (of level α) based on $\rho_{h,\psi}$ converges to 1 as d tends to infinity.

This theorem shows that if the sample sizes are not too small (i.e., m and n satisfy the conditions stated in Theorem 2), NN and MST-run tests based on $\rho_{h,\psi}$ have the high-dimensional consistency. Under Assumptions 4 and 5, SHP-run and NBP tests based on $\rho_{h,\psi}$ also have this high-dimensional consistency when m and n satisfy the conditions stated in Theorem 1. We state this result in the following theorem. The proof is similar to the proof of Theorem 1. So, we skip the proof.

Theorem 5 Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim G$ are independent random vectors, and $\rho_{h,\psi}$ is used to construct the test statistics, where h and ψ satisfy the conditions of Lemma 4. Then, under Assumptions 4 and 5, we get the following results.

- (a) If $N/\binom{N}{m} < \alpha$, then the power of the SHP-run test (of level α) based on $\rho_{h,\psi}$ converges to 1 as d tends to infinity.
- (b) Let c(m, n) be as defined in Theorem 1. If $c(m, n) < \alpha$, then the power of the NBP test (of level α) based on $\rho_{h,\psi}$ converges to 1 as d tends to infinity.

3.1 Performance under weak signal

In Theorems 4 and 5, we have established the consistency of the tests based on $\rho_{h,\psi}$ when $\widetilde{\rho}_{h,\psi}(F,G)>0$, or equivalently, $\liminf_{d\to\infty}d^{-1}\sum_{q=1}^d e_{F,G}^{(q)}>0$. So, we need $\sum_{q=1}^d e_{F,G}^{(q)}$, the total signal against \mathcal{H}_0 , to increase at least at the rate of d. But, if only a few of the measurement variables carry information against \mathcal{H}_0 , we may have $\liminf_{d\to\infty}d^{-1}\sum_{q=1}^d e_{F,G}^{(q)}=0$. Next, we investigate the high-dimensional behavior of the tests based on $\rho_{h,\psi}$ in such situations. For two independent random vectors \mathbf{U} , \mathbf{V} from F or G, let us assume that $\mathrm{var}\{\sum_{q=1}^d \psi(|U^{(q)}-V^{(q)}|)\}=\mathbf{O}(\vartheta^2(d))$. If the measurement variables are highly correlated, we usually have $\vartheta^2(d)=\mathbf{O}(d^2)$. But weak dependence among the measurement variables leads to $\vartheta^2(d)=\mathbf{o}(d^2)$. For instance, when they are m-dependent, one gets $\vartheta^2(d)=d$. Now, for our investigation, we make the following assumption, which is weaker than Assumption 5.

Assumption 6 As d tends to infinity, $\rho_{h,\psi}^*(F,G) \, d/\vartheta(d)$ diverges to infinity.

In Assumption 6, we allow $\rho_{h,\psi}^*(F,G)$ to converge to 0, but at a rate slower than that of $\vartheta(d)/d$. For instance, for an m-dependent sequence, we allow $\rho_{h,\psi}^*(F,G)$ to converge to 0 at a rate slower than $d^{-1/2}$. Even when the measurement variables are not m-dependent, under certain weak dependence assumptions on the underlying marginal distributions, we have $\vartheta^2(d) = dL(d)$, where L is a slowly varying function (see, e.g., Lin and Lu 1996, Chap. 2). In that case, we allow $\rho_{h,\psi}^*(F,G)$ to converge to 0 at a rate slower than $d^{-1/2}L^{1/2}(d)$. Under Assumption 6, $\rho_{h,\psi}$ preserves the neighborhood structure in high dimensions when h is Lipschitz continuous, and this ensures the high-dimensional consistency of the resulting tests. The result is stated below.

Theorem 6 Suppose that $X_1, ..., X_m \sim F$ and $Y_1, ..., Y_n \sim G$ are independent random vectors, where F and G satisfy Assumptions 4 and 6. If h is Lipschitz continuous, then

$$Pr\bigg[\min_{i,j}\rho_{h,\psi}(\mathbf{X}_i,\mathbf{Y}_j) > \max\big\{\max_{i\neq j}\rho_{h,\psi}(\mathbf{X}_i,\mathbf{X}_j), \max_{i\neq j}\rho_{h,\psi}(\mathbf{Y}_i,\mathbf{Y}_j)\big\}\bigg] \to 1 \ as \ d \to \infty.$$



Consequently, if m and n satisfy the conditions of Theorem 4, then the powers of NN and MST-run tests (of level α) based on $\rho_{h,\psi}$ converge to 1 as d tends to infinity. Similarly, if m and n satisfy the conditions of Theorem 5, then the powers of NBP and SHP-run tests (of level α) based on $\rho_{h,\psi}$ also converge to 1 as d tends to infinity.

For independent random vectors $\mathbf{X} \sim F$ and $\mathbf{Y} \sim G$, under the assumptions of Theorem 6, we get $\rho_{h,\psi}(\mathbf{X},\mathbf{Y}) = \rho_{h,\psi}^*(F,G) + \mathbf{O}_P(\vartheta(d)/d)$ (see the proof of Theorem 6). While $\rho_{h,\psi}^*(F,G)$ is the signal against \mathcal{H}_0 , the quantity $\vartheta(d)/d = \sqrt{\operatorname{var}\left\{d^{-1}\sum_{q=1}^d \psi(|X^{(q)}-Y^{(q)}|)\right\}}$ represents the stochastic variation or noise. Theorem 6 shows the high-dimensional consistency of the tests based on $\rho_{h,\psi}$ when h is Lipschitz continuous and the signal-to-noise ratio diverges with d. Similar results can be obtained even when h is not Lipschitz. For instance, in the case of ρ_0 , $h(t) = \sqrt{t}$ is not Lipschitz continuous, but we have the following result.

Theorem 7 Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim G$ are independent random vectors, where F and G have mean vectors $\boldsymbol{\mu}_F, \boldsymbol{\mu}_G$ and dispersion matrices $\boldsymbol{\Sigma}_F, \boldsymbol{\Sigma}_G$, respectively. Further assume that $\liminf_{d\to\infty} \min\{trace(\boldsymbol{\Sigma}_F), trace(\boldsymbol{\Sigma}_G)\}/\vartheta(d) > 0$, where $\vartheta^2(d)$ is the order of $var(\|\mathbf{W}\|^2)$ for $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2, \mathbf{Y}_1 - \mathbf{Y}_2$ and $\mathbf{X}_1 - \mathbf{Y}_1$ (i.e., $var(\|\mathbf{W}\|^2) = \mathbf{O}(\vartheta^2(d))$). If $\|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|^2/\vartheta(d)$ or $|trace(\boldsymbol{\Sigma}_F) - trace(\boldsymbol{\Sigma}_G)|/\vartheta(d)$ diverges to infinity as d increases, then

$$Pr\Big[\min_{i,j}\rho_0(\mathbf{X}_i,\mathbf{Y}_j) > \max\big\{\max_{i\neq j}\rho_0(\mathbf{X}_i,\mathbf{X}_j), \max_{i\neq j}\rho_0(\mathbf{Y}_i,\mathbf{Y}_j)\big\}\Big] \to 1 \ as \ d \to \infty.$$

Consequently, if m and n satisfy the conditions of Theorem 4, then the powers of NN and MST-run tests (of level α) based on ρ_0 converge to 1 as d tends to infinity. Similarly, if m and n satisfy the conditions of Theorem 5, then the powers of NBP and SHP-run tests (of level α) based on ρ_0 also converge to 1 as d tends to infinity.

Thus, when the measurement variables are m-dependent, for the consistency of the tests based on ρ_0 , we need either $d^{-1/2}\|\mu_F - \mu_G\|^2$ or $d^{-1/2}|trace(\Sigma_F) - trace(\Sigma_G)|$ to diverge to infinity as d increases. This condition is much weaker than the condition ' $d^{-1}\|\mu_F - \mu_G\|^2 > 0$ or $d^{-1}|trace(\Sigma_F) - trace(\Sigma_G)| > 0$ ' (i.e., $v^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$) assumed for the tests based on the Euclidean distance.

3.2 Computational issues

Computation of MADD between two data points has an associated cost of the order O(dn) compared to O(d) needed for the Euclidean distance or $\varphi_{h,\psi}$. But in the HDLSS set up, where d is much larger than n, these are of the same asymptotic order. Moreover, after computing all pairwise distances, the steps used for obtaining the test statistics are the same in all cases. Therefore, for HDLSS data, though the tests based on MADD require more time compared to the corresponding tests based on the Euclidean distance or $\varphi_{h,\psi}$, the time difference is very small. This is quite evident from Table 1, which shows average computing times required by different tests based on the Euclidean distance and ρ_0 for various dimensions and sample sizes. We used MATLAB codes for the NN test, the MST-run test and the SHP-run test, while the codes for the NBP test were written in R. All the codes were run on a computer with 16 GB RAM, having Intel Core i5 CPU with a clock speed of 1.60 GHz.



Sample size	d = 200				d = 500				d = 1000				
(m = n)	index	NN	MST	SHP	NBP	NN	MST	SHP	NBP	NN	MST	SHP	NBP
20	Euclidean	5.97	5.51	3.87	0.31	11.25	10.83	9.15	0.50	20.02	19.77	17.78	0.73
	MADD (ρ_0)	6.23	5.71	4.03	0.72	11.30	10.89	9.21	0.83	20.09	19.82	17.83	0.96
40	Euclidean	10.94	11.47	8.44	0.59	21.65	22.14	18.82	1.14	39.32	40.11	36.67	2.34
	MADD (ρ_0)	11.27	11.96	8.43	2.05	21.88	22.59	19.00	2.62	39.96	40.30	36.75	3.79

 Table 1 Average run times (of 100 trials) for different tests (in seconds)

4 Results from the analysis of simulated and real data sets

In this section, we analyze a few more simulated data sets and two real data sets for further evaluation of the tests based on $\varphi_{h,\psi}$ and $\rho_{h,\psi}$. For each of these examples, we repeated the experiment 500 times to compute the powers of different tests, which are shown in Figs. 7, 8, 9, 10, 11, 12 and 13. We also report the results for the usual tests based on the Euclidean distance and those proposed by Chen and Friedman (2017) to facilitate comparison. For all simulated data sets, we used m = n = 20 as before.

4.1 Analysis of simulated data sets

We start with two examples (Examples 3 and 4), where both populations have i.i.d. measurement variables, and they differ only in their locations and scales. In **Example 3** (respectively, **Example 4**), these measurement variables have normal (respectively, Cauchy) distribution. While they have location parameter 0 and scale parameter 1 for the first population, those are 0.1 and 1.1 for the second population. Powers of different tests in these two examples are shown in Fig. 7. One can see that in the presence of normal distribution (which is light-tailed) in Example 3, the tests based on $\psi_1(t) = t$ performed better than the ones based on $\psi_2(t) = 1 - \exp(-t)$. But, we observed a completely different picture in Example 4. In the presence of Cauchy distribution (which is heavy-tailed), all tests based on ψ_1 had powers close to the nominal level, but those based on ψ_2 had excellent performance. This shows the robustness of our proposed tests based on bounded ψ functions against heavy-tailed distributions. CF tests based on the Euclidean distance performed well in Example 3, but just like other tests based on the Euclidean distance, due to lack of robustness, they failed in Example 4.

Figure 7 also shows that in both examples, modified NN and MST-run tests based on $\varphi_{h,\psi}$ performed very poorly. Note that in both of these examples, the scale-difference between two distributions dominates the location-difference, and for all three choices of h and ψ , we have $\varphi_{h,\psi}^*(F,F) < \varphi_{h,\psi}^*(F,G) < \varphi_{h,\psi}^*(G,G)$. This violation of neighborhood structure was the reason behind the failure of these tests. As it is expected from our earlier discussion, in these examples, modified versions of SHP-run and NBP tests based on $\rho_{h,\psi}$ performed much better than those based on $\varphi_{h,\psi}$.

Our next two examples involve alternatives with sparse signals, where only a fraction of the measurement variables contain information against \mathcal{H}_0 , and that fraction shrinks to 0 as the dimension increases. So, Assumption 5 does not hold in these examples.

Example 5 We consider two normal distributions $\mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $\mathcal{N}_d(\boldsymbol{\mu}_d, \boldsymbol{\Lambda}_d)$ as F and G, respectively, where $\boldsymbol{\mu}_d = (\mu_1, \dots, \mu_d)^{\top}$ with $\mu_i = \sqrt{0.01 \log(d)}$ for $i = 1, \dots, \lfloor d^{1/2} \rfloor$



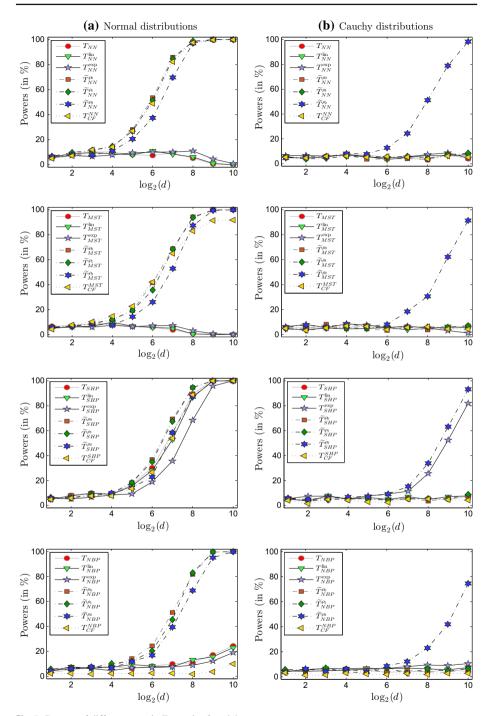


Fig. 7 Powers of different tests in Examples 3 and 4



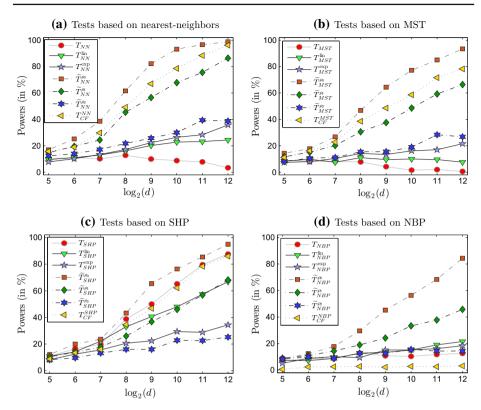


Fig. 8 Powers of different tests in Example 5

and 0 otherwise. The diagonal matrix Λ_d has the first $\lfloor d^{1/2} \rfloor$ elements equal to 0.5 $\log(d)$ and the rest equal to 1.

In this example, modified tests based on ρ_0 outperformed all other tests considered here (see Fig. 8). This is quite expected because here the two distributions have light tails, and they differ only in their locations and scales. Note that for ρ_0 , $\sum_{q=1}^d e_{F,G}^{(q)}$ is of the order $\mathbf{O}(d^{1/2}\log(d))$, while for ρ_2 , it is of the order $\mathbf{O}(d^{1/2})$ since the function ψ_2 is bounded. In this example, ν^2 have smaller order than $|\sigma_F^2 - \sigma_G^2|$ (here ν^2 , σ_F^2 and σ_G^2 have the same meaning as in Assumption 3). So, NN and MST-run tests based on the Euclidean distance had powers close to zero. But, the SHP-run test based on the Euclidean distance performed much better. This is consistent with our theoretical results.

Example 6 Both F and G have independent measurement variables. For F, they are i.i.d. $\mathcal{N}(0, 1)$. For G, the first $\lfloor d^{2/3} \rfloor$ variables are $t_3(0, 1/3)$, and the rest are $\mathcal{N}(0, 1)$. So, F and G have the same location and dispersion structure, while the first $\lfloor d^{2/3} \rfloor$ marginal distributions differ in their shapes.

In this example, we observed a different picture (see Fig. 9) than what we observed in Example 5. Modified tests based on ρ_1 and ρ_2 , particularly those based on ρ_2 performed much better than all other tests considered here. For the SHP-run test, modified versions based on $\varphi_{h,\psi}$ with ψ_1 and ψ_2 had powers comparable to those based on ρ_1 and ρ_2 , respectively. Note that in this example, two distributions have the same location and scale, but they differ in



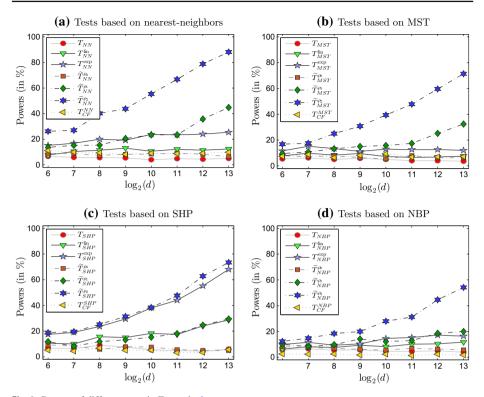


Fig. 9 Powers of different tests in Example 6

their univariate marginal distributions. In such a case, pairwise Euclidean distances fail to extract the signal against \mathcal{H}_0 . So, CF tests, the tests based on the Euclidean distance and those based on ρ_0 , all had powers close to the nominal level.

In our last two examples, we consider some situations where the assumptions used for our theoretical investigations do not hold. These examples are used to investigate how our proposed tests perform when the assumptions are violated.

Example 7 We take F to be an equal mixture of $\mathcal{N}_d(0.3\mathbf{1}_d, \mathbf{I}_d)$ and $\mathcal{N}_d(-0.3\mathbf{1}_d, 4\mathbf{I}_d)$, while G is taken to be an equal mixture of $\mathcal{N}_d(0.3\boldsymbol{\alpha}_d, \mathbf{I}_d)$ and $\mathcal{N}_d(-0.3\boldsymbol{\alpha}_d, 4\mathbf{I}_d)$, where $\boldsymbol{\alpha}_d = (1, -1, \dots, (-1)^{d+1})^{\top}$.

In this example, NN and MST-run tests based on the Euclidean distance performed poorly (see Fig. 10), and the same was observed for CF-NN and CF-MST tests as well. However, SHP-run and NBP tests based on the Euclidean distance and the corresponding CF tests (CF-SHP and CF-NBP) performed well. Tests based on $\rho_{h,\psi}$, especially the ones based on ρ_0 had much superior performance in this example. Note that in this example, Assumptions 1–4 do not hold for the mixture distributions, but they hold for each component distribution. If we consider them as separate distributions, using the distance concentration phenomenon, we can explain the reasons behind the poor performance of the NN and MST-run tests based on the Euclidean distance and the superiority of their modifications based on MADD. Since the component distributions are light-tailed, the performance of the tests based on ρ_0 and ρ_1 was much better than the tests based on ρ_2 . The same was observed for the corresponding tests based on $\varphi_{h,\psi}$ as well.



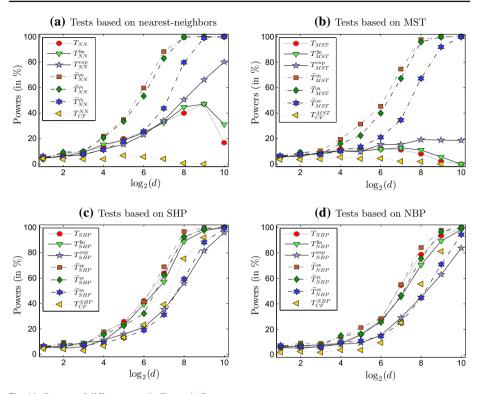


Fig. 10 Powers of different tests in Example 7

Example 8 We take F to be the $\mathcal{N}_d(\mathbf{0}_d, 3\mathbf{I}_d)$ distribution, while G is taken to be the standard multivariate t distribution with 3 degrees of freedom.

Here two populations have the same mean vector and the same dispersion matrix, while they differ only in their shapes. The results reported in Fig. 11 shows that the modified tests based on MADD performed much better in this example. The tests based on the Euclidean distance and $\varphi_{h,\psi}$ failed for NN and MST-run tests, but they performed well for SHP-run and NBP tests. CF-SHP and CF-NBP tests also performed well, but CF-NN and CF-MST tests showed some strange behaviors. The powers of these two tests initially increased with the dimension, but then they dropped down to zero as the dimension increased further.

4.2 Analysis of benchmark data sets

We analyzed two benchmark data sets, taken from the *UCR Time Series Classification Archive* (http://www.cs.ucr.edu/~eamonn/time_series_data/). These data sets, namely the GUN- POINT data and the LIGHTING- 2 data, have been extensively used in the literature of supervised classification. It is well-known that there is a reasonable separation between the two distributions in both of these data sets. So, assuming \mathcal{H}_0 to be false, we compared different tests based on their powers. These data sets consist of separate training and test samples. For our analysis, we merged these sets and following Biswas et al. (2014), we used random sub-samples of different sizes from the whole data set keeping the proportions of observations from different distributions as close as they are in the original data set. Each experiment was



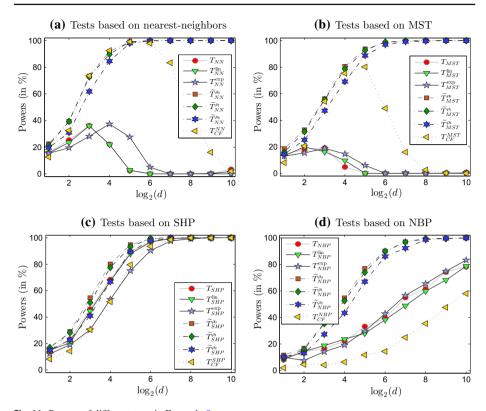


Fig. 11 Powers of different tests in Example 8

repeated 500 times to compute the powers of different tests, and they are shown in Figs. 12 and 13.

GUN-POINT data set comes from the video surveillance domain. It contains 100 observations from each of two classes: *Gun-Draw* and *Point*. For *Gun-Draw*, an actor draws a gun from a hip-mounted holster, points it at a target for approximately one second, and then returns the gun to the holster. For *Point*, the actor does the same move, but instead of the gun, points the index finger to the target for approximately one second, and then returns to the initial position. For each class, an observation consists of 150 measurements corresponding to the X co-ordinate of the centroid of the actor's right hand during one movement. In this data set, modified tests based on $\varphi_{h,\psi}$ performed better than the tests based on the Euclidean distance, MADD and CF tests (see Fig. 12). Among these modified tests, the one based on ψ_2 had a slight edge. The performance of the tests based on the Euclidean distance and MADD were quite similar, but the CF tests had inferior performance in all cases.

LIGHTNING- 2 data set consists of observations from two classes: *Cloud-to-Ground* lightning and *Intra-Cloud* lightning. Each observation corresponds to transient electromagnetic events detected by FORTE satellite. Every input went through a Fourier transform to get a spectrogram, which was then collapsed in frequency to produce a power density time series. These time series were smoothed to produce 637-dimensional observations. The data set consists of 48 and 73 observations from the two classes. Figure 13 shows that the tests based on $\varphi_{h,\psi}$ and $\varphi_{h,\psi}$ had superior performance in this example when ψ_1 or ψ_2 was used. Among



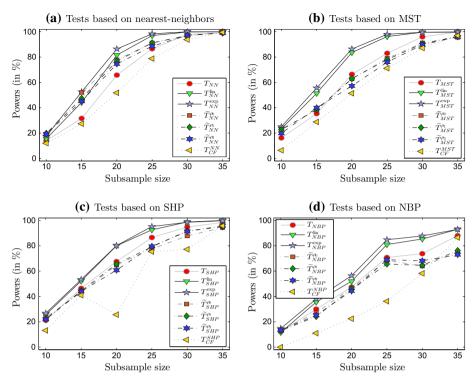


Fig. 12 Powers of different tests in GUN-POINT data

these tests, the ones based on $\psi_2(t) = 1 - \exp(-t)$ performed better. Tests based on ρ_0 and those based on the Euclidean distance had almost similar performance. CF tests did not have satisfactory performance in this example. Powers of theses tests were much lower than all other tests considered here.

5 Concluding remarks

In this article, we have used a new class of distance functions, $\varphi_{h,\psi}$, and associated MADD indices $\rho_{h,\psi}$ to modify some popular graph-based two-sample tests constructed using pairwise distances. In high dimension, while usual tests based on pairwise Euclidean distances are mainly useful for populations differing in their locations and overall scales, our modified tests can perform well for a larger variety of alternatives. There are some relative merits and demerits of these two types of modifications. For SHP-run and NBP tests, the first type of modification based on $\varphi_{h,\psi}$ is sufficient for the high-dimensional consistency. But this is not the case for NN and MST-run tests, where the modification based on $\varphi_{h,\psi}$ can produce poor results. In such cases, the use of MADD improves the performance of the tests significantly. When the neighborhood structure is retained by $\varphi_{h,\psi}$, the modified tests based on it perform better than the corresponding version based on MADD. But, when the neighborhood structure is violated, the use of MADD is preferred. For both $\varphi_{h,\psi}$ and $\rho_{h,\psi}$, the use of bounded ψ functions (e.g., $\psi(t) = 1 - \exp(-t)$) make the resulting tests robust against outliers gen-



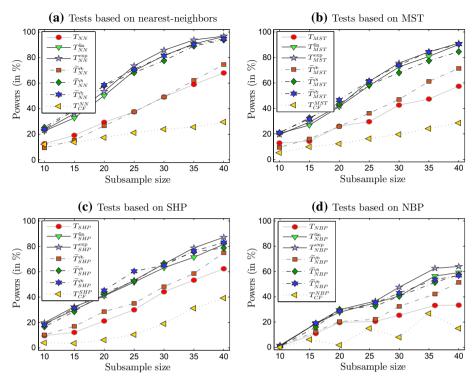


Fig. 13 Powers of different tests in LIGHTNING- 2 data

erated from heavy-tailed distributions. But, for light-tailed distributions, the use of linear or square functions (i.e., $\psi(t) = t$ or $\psi(t) = t^2$) produce better results.

Our general recipe based on MADD can also be used to improve the high-dimensional performance of many other two-sample tests. For instance, we can modify the tests based on averages of pairwise distances (see, e.g., Baringhaus and Franz 2004, 2010; Biswas and Ghosh 2014; Tsukada 2019). CF tests (Chen and Friedman 2017) can also be modified using $\varphi_{h,\psi}$ or $\rho_{h,\psi}$. High-dimensional consistency of the resulting tests can be proved using arguments similar to those used in this article. These modified average-distance-based and graph-based tests have relative merits and demerits. While average-distance-based tests (see, e.g., Sarkar and Ghosh 2018) put more emphasis on the magnitude of the observations, graph-based tests rely mainly on different types of edge-counts. When the underlying distributions are light-tailed, average-distance-based tests usually perform better than graphbased tests. On the other hand, graph-based tests are more robust against contamination and outliers generated from heavy-tailed distributions. As a result, they often outperform average-distance-based tests when the underlying distributions have heavy tails. Our empirical experience also suggests that there are several other cases when graph-based tests are preferred. For instance, when one or more of the underlying distributions are mixtures of several distributions, average-distance-based tests may perform poorly while graph-based tests can produce excellent results. However, it needs an extensive investigation, which is beyond the scope of this article. Using our ideas based on $\varphi_{h,\psi}$ or $\rho_{h,\psi}$, several multi-sample tests can also be modified to achieve better performance in high dimensions. For the NN test and its modified versions, throughout this article, we have reported all numerical results for



k = 3 only. However, our findings remained almost the same for other values of k as well. This is expected in view of the theoretical results stated in this article.

For the construction of the general version of MADD [see Eq. (2)], we have used transformation on each of the measurement variables. This type of construction helps us to detect difference between the univariate marginals. However, this method can fail in situations where the distributions have same marginals, but differ in their joint structures. In such situations, following Sarkar and Ghosh (2018, Section 3), one can partition the measurement vector \mathbf{x} into K non-overlapping blocks $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_K$ of sizes d_1, \dots, d_K ($\sum_{i=1}^K d_i = d$), respectively, and define MADD using blocked distance functions of the form $\varphi_{h,\psi}^K(\mathbf{x}, \mathbf{y}) = 0$ $h\{K^{-1}\sum_{q=1}^K \psi(\|\tilde{\mathbf{x}}_q - \tilde{\mathbf{y}}_q\|)\}$. As long as the block sizes are uniformly bounded, and the joint distributions of variables in the blocks are different under F and G, consistency of the resulting tests based on MADD can be proved under conditions similar to Assumptions 4–6. One can also see that similar results can be proved for the corresponding tests based on $\varphi_{h,\psi}^B$ as well. This type of blocking can reveal more minute differences between two distributions. For instance, using blocks of size 2, one can distinguish between two distributions having the same univariate marginals but different correlation structures (see, e.g., Sarkar and Ghosh 2018). In that case, ideally, one would like to put highly correlated variables in the same block. In general, we would like to find blocks which are nearly independent, but the variables inside a block have significant dependence among themselves. But, at this moment, it is not yet clear how to develop an algorithm for finding such optimal blocks from the data. This can be considered as an interesting problem for future research.

6 Codes

Codes for implementation of different tests proposed in the article can be found at https://drive.google.com/open?id=1Ym8r6reNDfZrQ79LP4MJ93ybNc8-A2b1.

Acknowledgements This research was partially supported by Keysight Technologies, Inc., USA.

Proofs and mathematical details

Throughout this section, we use \Pr^* to denote conditional probability given \mathcal{Z}_N . So, we use $\Pr^*(A)$ to denote $\Pr(A \mid \mathcal{Z}_N)$ for an event A. For NN, MST-run, SHP-run and NBP tests based on $\varphi_{h,\psi}$, the tests statistics are denoted by $T_{NN}^{h,\psi}$, $T_{MST}^{h,\psi}$, $T_{SHP}^{h,\psi}$ and $T_{NBP}^{h,\psi}$, respectively.

Proof of Theorem 1 (a) Since h is uniformly continuous, under Assumption 4, we have

$$\varphi_{h,\psi}(\mathbf{X}_i, \mathbf{X}_j) - a_d \stackrel{\Pr}{\to} 0, \varphi_{h,\psi}(\mathbf{Y}_i, \mathbf{Y}_j) - b_d \stackrel{\Pr}{\to} 0 \text{ and } \varphi_{h,\psi}(\mathbf{X}_i, \mathbf{Y}_j) - c_d \stackrel{\Pr}{\to} 0 \text{ as } d \to \infty,$$
(4)

where $a_d = \varphi_{h,\psi}^*(F,F)$, $b_d = \varphi_{h,\psi}^*(G,G)$ and $c_d = \varphi_{h,\psi}^*(F,G)$ (see the discussion before Lemma 2). Since $\liminf_{d\to\infty}(2c_d-a_d-b_d)=\liminf_{d\to\infty}e_{h,\psi}(F,G)>0$, following the proof of Theorem 1 in Biswas et al. (2014), it is easy to show that $\Pr(T_{SHP}^{h,\psi}\leq 3)\to 1$ as $d\to\infty$. Under \mathcal{H}_0 , $T_{SHP}^{h,\psi}$ is distribution-free, and $\Pr_{\mathcal{H}_0}(T_{SHP}^{h,\psi}\leq 3)=m!n!/(m+n-1)!=N/\binom{N}{m}<\alpha$ implies that the cut-off is larger than 3. This completes the proof.

(b) For the NBP test, first assume that N is even. In that case, either (i) both m and n are even or (ii) both m and n are odd. In case (i), the test statistic $T_{NBP}^{h,\psi}$ can take only even values, say



2k. So, there are 2k pairs of the \mathbf{XY} -type, (m-2k)/2 pairs of the \mathbf{XX} -type and (n-2k)/2 pairs of the \mathbf{YY} -type. If $\Delta_{2k,d}^{m,n} = \sum_{i=1}^{N/2} \varphi_{h,\psi}(\mathbf{Z}_{i1},\mathbf{Z}_{i2})$ denotes the corresponding total weight, then $\Delta_{2k,d}^{m,n} - C_{2k,d} \stackrel{\text{Pr}}{\to} 0$ as $d \to \infty$, where $C_{2k,d} = (m-2k)a_d/2 + (n-2k)b_d/2 + 2kc_d = (2c_d - a_d - b_d)k + ma_d/2 + nb_d/2$ [see Eq. (4)]. Since $\lim\inf_{d\to\infty}(2c_d - a_d - b_d) > 0$, for all large d, this value is minimized for k=0. So, $T_{NBP}^{h,\psi} \stackrel{\text{Pr}}{\to} 0$ as $d\to\infty$. In case (ii), $T_{NBP}^{h,\psi}$ can take only odd values, say 2k-1. Here also, one can check that $\Delta_{2k-1,d}^{m,n} - C_{2k-1,d} \stackrel{\text{Pr}}{\to} 0$ as $d\to\infty$, where $C_{2k-1,d} = (k-1)(2c_d - a_d - b_d) + ma_d/2 + nb_d/2$. For all large d, $C_{2k-1,d}$ is minimized for k=1. So, $T_{NBP}^{h,\psi} \stackrel{\text{Pr}}{\to} 1$ as $d\to\infty$. Under \mathcal{H}_0 , $T_{NBP}^{h,\psi}$ is distribution-free, and following Rosenbaum (2005), one can show that under the conditions on m and n, both $\Pr_{\mathcal{H}_0}\left(T_{NBP}^{h,\psi} \le 0\right)$ in case (i) and $\Pr_{\mathcal{H}_0}\left(T_{NBP}^{h,\psi} \le 1\right)$ in case (ii) are less than α . This completes the proof.

Now consider the case when N is odd. Without loss of generality, let m be odd and n be even. Since N is odd, one observation remains unpaired, and it is removed from the data. There are two possibilities, (i) m is reduced to m-1 and (ii) n is reduced to n-1. If case (i) happens, since both m-1 and n are even, following our discussion in the previous paragraph, for all large d, the total weight is minimized for k=0 and $\Delta_{0,d}^{m-1,n}-A_{0,d}\overset{\Pr}{\to} 0$ as $d\to\infty$, where $A_{0,d}=(m-1)a_d/2+nb_d/2$. Similarly, if case (ii) happens, since both m and n-1 are odd, for all large d, the total weight is minimized for k=1 and $\Delta_{1,d}^{m,n-1}-A_{1,d}\overset{\Pr}{\to} 0$ as $d\to\infty$, where $A_{1,d}=A_{0,d}+c_d$. Note that $2c_d\geq e_{h,\psi}(F,G)$, and hence $\lim\inf_{d\to\infty}c_d>0$. So, $A_{0,d}$ is strictly smaller than $A_{1,d}$ for all large d. So, case (i) happens with probability tending to one, and hence $T_{NBP}^{h,\psi}\overset{\Pr}{\to} 0$ as $d\to\infty$. Now, under the condition on m and n, $\Pr_{\mathcal{H}_0}\left(T_{NBP}^{h,\psi}\leq 0\right)=c(m,n)<\alpha$. This completes the proof.

Proof of Theorem 2 (a) Note that the NN test is based on $T_{NN}^{h,\psi} = \frac{1}{Nk} \Big[\sum_{i=1}^m \sum_{t=1}^k \mathbb{I}_t^{h,\psi}(\mathbf{X}_i) + \sum_{j=1}^n \sum_{t=1}^k \mathbb{I}_t^{h,\psi}(\mathbf{Y}_j) \Big]$, where $\mathbb{I}_t^{h,\psi}(\mathbf{Z})$ is an indicator variable that takes the value 1 if \mathbf{Z} and its t-th nearest-neighbor in terms of $\varphi_{h,\psi}$ are from the same distribution. Recall that since h is uniformly continuous, under Assumption 4, $\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{X}_j) - a_d \overset{\Pr}{\to} 0$, $\varphi_{h,\psi}(\mathbf{Y}_i,\mathbf{Y}_j) - b_d \overset{\Pr}{\to} 0$ and $\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{Y}_j) - c_d \overset{\Pr}{\to} 0$ as $d \to \infty$, where $a_d = \varphi_{h,\psi}^*(F,F)$, $b_d = \varphi_{h,\psi}^*(G,G)$ and $c_d = \varphi_{h,\psi}^*(F,G)$ [see Eq. (4)]. Since $\lim\inf_{d\to\infty}(c_d-a_d)>0$ and $\lim\inf_{d\to\infty}(c_d-b_d)>0$, it follows that

$$\Pr\left\{\max_{i\neq j}\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{X}_j) < \min_{i,j}\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{Y}_j)\right\} \to 1 \text{ and}$$

$$\Pr\left\{\max_{i\neq j}\varphi_{h,\psi}(\mathbf{Y}_i,\mathbf{Y}_j) < \min_{i,j}\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{Y}_j)\right\} \to 1 \text{ as } d \to \infty. \tag{5}$$

So, for every $t \leq k$, $\mathbb{I}_t^{h,\psi}(\mathbf{X}_i) \stackrel{\Pr}{\to} 1$ for $i=1,\ldots,m$ and $\mathbb{I}_t^{h,\psi}(\mathbf{Y}_j) \stackrel{\Pr}{\to} 1$ for $j=1,\ldots,n$ as $d\to\infty$. Thus, $T_{NN}^{h,\psi}$ converges in probability to its maximum value 1. Now, to prove the consistency of the test based on $T_{NN}^{h,\psi}$, we shall show that $\Pr^*(T_{NN}^{h,\psi}=1) < \alpha$ for almost every \mathcal{Z}_N .

Call $S \subseteq \mathcal{Z}_N$ to be a neighbor-complete set if for any $\mathbf{z} \in S$, all of its k nearest-neighbors based on $\varphi_{h,\psi}$ also belong to S, and no proper subset of S has this property. Clearly, $k+1 \le |S| \le N$, where |S| denotes the cardinality of S. Let \mathcal{Z}_N be partitioned into r such neighbor-complete sets, i.e., $\mathcal{Z}_N = S_1 \cup \ldots \cup S_r$, where $r \le \lfloor N/(k+1) \rfloor$. Note that $T_{NN}^{h,\psi} = 1$ if and only if, for each $i = 1, \ldots, r$, all observations in S_i have the same label. If $r_1(< r)$ of



these S_i 's are labelled F and the rest are labelled G, then the sum of cardinalities of these r_1 sets should be m. Let $c_0(m, n)$ be the number of ways in which this can be done. Clearly, $\Pr^*(T_{NN}^{h,\psi} = 1) = c_0(m,n)/\binom{N}{m}$. So, it is enough to show that $c_0(m,n) \leq \binom{N_0}{m_0}$.

First observe that we cannot have $T_{NN}^{h,\psi}=1$ if N<2(k+1). If N=2(k+1), it is possible only when m=n=k+1 and $\mathcal{Z}_N=S_1\cup S_2$, with $|S_1|=|S_2|=k+1$. So, in that case, all observations either in S_1 or in S_2 must be labelled as F. This leads to $c_0(m,n)=2$, and the result holds for N=2(k+1).

Now, we shall prove the result using the method of mathematical induction on N. First assume that the result holds for all N with $2(k+1) \le N \le M$. Without loss of generality, let us also assume that $m \le n$. For N = M+1, first note that observations in S_1 may or may not be labelled as F. Therefore, if $|S_1| = k_1$, we have $c_0(m, n) = c_0(m - k_1, n) + c_0(m, n - k_1)$. So, using the result for $N - k_1$, we get

$$c_0(m,n) \leq \left(\left\lceil \frac{N-k_1}{k+1} \right\rceil \atop \left\lceil \frac{m-k_1}{k+1} \right\rceil \right) + \left(\left\lceil \frac{N-k_1}{k+1} \right\rceil \atop \left\lceil \frac{m}{k+1} \right\rceil \right).$$

Here $\lceil (N-k_1)/(k+1) \rceil \leq \lceil N/(k+1) \rceil - \lfloor k_1/(k+1) \rfloor \leq N_0 - 1$ and $\lceil m/(k+1) \rceil = m_0$. So,

$$c_0(m,n) \leq \binom{N_0-1}{\left\lceil \frac{m-k_1}{k+1} \right\rceil} + \binom{N_0-1}{m_0}.$$

Also, observe that $\lceil (m-k_1)/(k+1) \rceil \le \lceil m/(k+1) \rceil - 1 = m_0 - 1$ and $m_0 - 1 \le (N_0 - 1)/2$. Thus,

$$c_0(m,n) \le \binom{N_0-1}{m_0-1} + \binom{N_0-1}{m_0} = \binom{N_0}{m_0}.$$

(b) Note that the MST-run test based on $\varphi_{h,\psi}$ uses the test statistic $T_{MST}^{h,\psi}=1+\sum_{i=1}^{n-1}\lambda_i$, where λ_i is an indicator variable that takes the value 1 if the *i*-th edge of the MST on the complete graph with edge weights defined using $\varphi_{h,\psi}$ connects two observations from different distributions. From Eq. (5), it follows that for sufficiently large d, the MST on the vertex set \mathcal{Z}_N has a sub-tree \mathcal{T}_1 on vertices corresponding to m observations from G. These two sub-trees are connected by an edge of the **XY**-type (see Biswas et al. 2014). As a result, $T_{MST}^{h,\psi}$ converges in probability to its minimum value 2. From the proof of Theorem 2 in Biswas et al. (2014), it follows that for sufficiently large d, $\Pr^*(T_{MST}^{h,\psi} \leq 2) \leq \max\{\lfloor N/m \rfloor, \lfloor N/n \rfloor\}/\binom{N}{m} < \alpha$ for almost every \mathcal{Z}_N . This proves the result.

Proof of Theorem 3 (a) Recall that since h is uniformly continuous, under Assumption 4, we have

$$\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{X}_j) - a_d \overset{\Pr}{\to} 0, \ \ \varphi_{h,\psi}(\mathbf{Y}_i,\mathbf{Y}_j) - b_d \overset{\Pr}{\to} 0 \ \ \text{and} \ \ \varphi_{h,\psi}(\mathbf{X}_i,\mathbf{Y}_j) - c_d \overset{\Pr}{\to} 0,$$

where $a_d = \varphi_{h,\psi}^*(F,F)$, $b_d = \varphi_{h,\psi}^*(G,G)$ and $c_d = \varphi_{h,\psi}^*(F,G)$ [see Eq. (4)]. Since $2c_d - a_d - b_d \ge 0$ and $\limsup_{d \to \infty} (c_d - a_d) < 0$, it follows that $\liminf_{d \to \infty} (c_d - b_d) > 0$, and hence

$$\Pr\left\{\max_{i\neq j}\varphi_{h,\psi}(\mathbf{Y}_i,\mathbf{Y}_j) < \min_{i,j}\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{Y}_j)\right\} \to 1 \text{ and}$$

$$\Pr\left\{\max_{i,j}\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{Y}_j) < \min_{i\neq j}\varphi_{h,\psi}(\mathbf{X}_i,\mathbf{X}_j)\right\} \to 1 \text{ as } d \to \infty. \tag{6}$$

As a result, for every $t \leq k$, $\mathbb{I}_t^{h,\psi}(\mathbf{X}_i) \stackrel{\Pr}{\to} 0$ for $i = 1, \ldots, m$ and $\mathbb{I}_t^{h,\psi}(\mathbf{Y}_j) \stackrel{\Pr}{\to} 1$ for $j = 1, \ldots, n$ as $d \to \infty$. Thus, $T_{NN}^{h,\psi} \stackrel{\Pr}{\to} n/N$ as $d \to \infty$. Now, from the proof of Theorem 3.2(b) in Biswas and Ghosh (2014), it follows that when $(m-1)/n > (1+\alpha)/(1-\alpha)$, $\Pr^*(T_{NN}^{h,\psi} \geq n/N) < \alpha$ for almost every \mathcal{Z}_N . This proves part (a) of the theorem.

(b) Equation (6) implies that $T_{MST}^{h,\psi} \stackrel{\Pr}{\to} m+1$ as $d\to\infty$ (see Biswas and Ghosh 2014). Under the condition $m/n > (1+\alpha)/(1-\alpha)$, from the proof of Theorem 2(ii) in Biswas et al. (2014), it also follows that $\Pr^*(T_{MST}^{h,\psi} \le m+1) \ge (m-n)/N > \alpha$ for almost every \mathcal{Z}_N . Thus, the cut-off obtained using the permutation principle turns out to be strictly smaller than the observed value with probability converging to one as the dimension increases. This completes the proof.

Proof of Lemma 3 Symmetry and non-negativity of $\rho_{h,\psi}$ are obvious. So, we shall prove the triangle inequality for $\rho_{h,\psi}$. First observe that

$$\begin{aligned} & \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_3) - \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_3) \right| \\ &= \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_2) - \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_3) - \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_2) + \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_3) \right| \\ &\leq \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_2) - \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_2) \right| + \left| \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_1) - \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_1) \right|. \end{aligned}$$

This proves the result for N = 3. If N > 4, for any \mathbf{z}_k with k > 4,

$$\begin{aligned} \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_k) \right| \\ &= \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_k) + \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_k) \right| \\ &\leq \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_k) \right| + \left| \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_k) \right|. \end{aligned}$$

Combining these above-mentioned inequalities, we get

$$\begin{split} \sum_{k \neq 1,2} \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_k) \right| &\leq \sum_{k \neq 1,3} \left| \varphi_{h,\psi}(\mathbf{z}_1, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_k) \right| \\ &+ \sum_{k \neq 2,3} \left| \varphi_{h,\psi}(\mathbf{z}_2, \mathbf{z}_k) - \varphi_{h,\psi}(\mathbf{z}_3, \mathbf{z}_k) \right|. \end{split}$$

This implies $\rho_{h,\psi}(\mathbf{z}_1,\mathbf{z}_2) \leq \rho_{h,\psi}(\mathbf{z}_1,\mathbf{z}_3) + \rho_{h,\psi}(\mathbf{z}_2,\mathbf{z}_3)$.

Proof of Lemma 4 Since h is strictly increasing, $\rho_{h,\psi}^*(F,G) = 0$ implies $\sum_{q=1}^d \mathrm{E}\psi(|X_1^{(q)} - X_2^{(q)}|) = \sum_{q=1}^d \mathrm{E}\psi(|Y_1^{(q)} - X_2^{(q)}|)$ and $\sum_{q=1}^d \mathrm{E}\psi(|X_1^{(q)} - Y_2^{(q)}|) = \sum_{q=1}^d \mathrm{E}\psi(|Y_1^{(q)} - Y_2^{(q)}|)$. So,

$$\sum_{q=1}^d e_{F,G}^{(q)} = \sum_{q=1}^d \left\{ 2 \mathrm{E} \psi(|X_1^{(q)} - Y_1^{(q)}|) - \mathrm{E} \psi(|X_1^{(q)} - X_2^{(q)}|) - \mathrm{E} \psi(|Y_1^{(q)} - Y_2^{(q)}|) \right\} = 0.$$

Since $\psi'(t)/t$ is a non-constant, completely monotone function, for each $q=1,\ldots,d$, $e_{F,G}^{(q)}$ is non-negative and it takes the value 0 if and only if the q-th marginal distributions of F and G are the same (see Baringhaus and Franz 2010; Biswas et al. 2015). Thus, $\rho_{h,\psi}^*(F,G)=0$ implies that F and G have the same univariate marginal distributions. On the other hand, when F and G have the same univariate marginal distributions, it follows trivially that $\varphi_{h,\psi}^*(F,F)=\varphi_{h,\psi}^*(G,G)=\varphi_{h,\psi}^*(F,G)$, and hence $\rho_{h,\psi}^*(F,G)=0$.

Proof of Theorem 4 The proof is similar to the proof of Theorem 2 with the use of Assumption 5. Hence we skip the details of the proof.



Proof of Theorem 5 The proof is similar to the proof of Theorem 1 with the use of Assumption 5. Hence we skip the details of the proof.

Proof of Theorem 6 Consider independent random vectors $\mathbf{X}_1, \mathbf{X}_2 \sim F, \mathbf{Y}_1, \mathbf{Y}_2 \sim G$ and $\mathbf{Z} \sim H$, where H = F or G. Define $S_d = d^{-1} \sum_{q=1}^d \psi(|X_1^{(q)} - Y_1^{(q)}|)$. Since $(S_d - \mathrm{E}(S_d))/\sqrt{\mathrm{var}(S_d)} = \mathbf{O}_P(1)$, we get $S_d - \mathrm{E}(S_d) = \mathbf{O}_P(\vartheta(d)/d)$. As h is Lipschitz continuous,

$$|\varphi_{h,\psi}(\mathbf{X}_1,\mathbf{Y}_1) - \varphi_{h,\psi}^*(F,G)| = |h(S_d) - h\{E(S_d)\}| \le C_0|S_d - E(S_d)| = \mathbf{O}_P(\vartheta(d)/d).$$

Similarly, $|\varphi_{h,\psi}(\mathbf{X}_1, \mathbf{Z}) - \varphi_{h,\psi}^*(F, H)|$ and $|\varphi_{h,\psi}(\mathbf{Y}_1, \mathbf{Z}) - \varphi_{h,\psi}^*(G, H)|$ are also of the order $\mathbf{O}_P(\vartheta(d)/d)$. So, $|\varphi_{h,\psi}(\mathbf{X}_1, \mathbf{Z}) - \varphi_{h,\psi}(\mathbf{Y}_1, \mathbf{Z})| = |\varphi_{h,\psi}^*(F, H) - \varphi_{h,\psi}^*(G, H)| + \mathbf{O}_P(\vartheta(d)/d)$. Since m and n are finite, this implies $\rho_{h,\psi}(\mathbf{X}_1, \mathbf{Y}_1) = \rho_{h,\psi}^*(F, G) + \mathbf{O}_P(\vartheta(d)/d)$. Similarly, we get $\rho_{h,\psi}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{O}_P(\vartheta(d)/d)$ and $\rho_{h,\psi}(\mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{O}_P(\vartheta(d)/d)$. Under Assumption 6, $\rho_{h,\psi}^*(F, G)$ has asymptotic order higher than that of $\vartheta(d)/d$. So,

$$\Pr[\rho_{h,\psi}(\mathbf{X}_1,\mathbf{Y}_1) > \max\{\rho_{h,\psi}(\mathbf{X}_1,\mathbf{X}_2), \rho_{h,\psi}(\mathbf{Y}_1,\mathbf{Y}_2)\}] \to 1 \text{ as } d \to \infty.$$

This proves the first part of the theorem. The consistency of the tests now follows using arguments similar to those used in the proofs of Theorems 1 and 2. \Box

Proof of Theorem 7 For ρ_0 , we use $h(t) = \sqrt{t}$ and $\psi(t) = t^2$. So, for $\mathbf{X}_1 \sim F$ and $\mathbf{Y}_1 \sim G$, taking $S_d = d^{-1} \sum_{q=1}^d (X_1^{(q)} - Y_1^{(q)})^2$, we get

$$\varphi_{h,\psi}(\mathbf{X}_1, \mathbf{Y}_1) - \varphi_{h,\psi}^*(F, G) = \sqrt{S_d} - \sqrt{E(S_d)} = \frac{S_d - E(S_d)}{\sqrt{S_d} + \sqrt{E(S_d)}}.$$

Here $\mathrm{E}(S_d) = d^{-1}\{\|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|^2 + tr(\boldsymbol{\Sigma}_F + \boldsymbol{\Sigma}_G)\} \ge d^{-1}tr(\boldsymbol{\Sigma}_F)$. So, $\sqrt{d\mathrm{E}(S_d)/\vartheta(d)}$ remains bounded away from 0, and hence $\sqrt{\vartheta(d)}/(\sqrt{dS_d} + \sqrt{d\mathrm{E}(S_d)})$ remains bounded as d goes to infinity. Now, $(S_d - \mathrm{E}(S_d))/\sqrt{\mathrm{var}(S_d)} = \mathbf{O}_P(1)$ implies $S_d - \mathrm{E}(S_d) = \mathbf{O}_P(\vartheta(d)/d)$. Again, $1/(\sqrt{S_d} + \sqrt{\mathrm{E}(S_d)}) = \mathbf{O}_P(\sqrt{d/\vartheta(d)})$. So, $\varphi_{h,\psi}(\mathbf{X}_1,\mathbf{Y}_1) = \varphi_{h,\psi}^*(F,G) + \mathbf{O}_P(\sqrt{\vartheta(d)/d})$. Thus, as in the proof of Theorem 6, we get $\rho_0(\mathbf{X}_1,\mathbf{Y}_1) = \rho_0^*(F,G) + \mathbf{O}_P(\sqrt{\vartheta(d)/d})$, where $\rho_0^*(F,G)$ is $\rho_{h,\psi}^*(F,G)$ with $h(t) = \sqrt{t}$ and $\psi(t) = t^2$. Similarly, we have $\rho_0(\mathbf{X}_1,\mathbf{X}_2) = \mathbf{O}_P(\sqrt{\vartheta(d)/d})$ and $\rho_0(\mathbf{Y}_1,\mathbf{Y}_2) = \mathbf{O}_P(\sqrt{\vartheta(d)/d})$. Now, it is easy to check that when $\|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|^2/\vartheta(d)$ or $|tr(\boldsymbol{\Sigma}_F) - tr(\boldsymbol{\Sigma}_G)|/\vartheta(d)$ diverge to infinity, $\rho_0^*(F,G)$ is of higher order than $\sqrt{\vartheta(d)/d}$. The rest of the proof is similar to the proof of Theorem 6.

References

Andrews, D. W. K. (1988). Laws of large numbers for dependent nonidentically distributed random variables. Econometric Theory, 4, 458–467.

Aslan, B., & Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation*, 75, 109–119.

Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88, 190–206.

Baringhaus, L., & Franz, C. (2010). Rigid motion invariant two-sample tests. *Statistica Sinica*, 20, 1333–1361. Billingsley, P. (1995). *Probability and measure*. New York: Wiley.

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

Biswas, M., Mukhopadhyay, M., & Ghosh, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, 101, 913–926.



- Biswas, M., Mukhopadhyay, M., & Ghosh, A. K. (2015). On some exact distribution-free one-sample tests for high dimension low sample size data. *Statistica Sinica*, 25, 1421–1435.
- Chen, H., & Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112, 397–409.
- de Jong, R. M. (1995). Laws of large numbers for dependent heterogeneous processes. *Econometric Theory*, 11, 347–358.
- Dutta, S., Sarkar, S., & Ghosh, A. K. (2016). Multi-scale classification using localized spatial depth. *Journal of Machine Learning Research*, 17(217), 1–30.
- Friedman, J. H., & Rafsky, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. The Annals of Statistics, 7, 697–717.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. Journal of Machine Learning Research, 13, 723–773.
- Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B*, 67, 427–444.
- Hall, P., & Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. Biometrika, 89, 359–374.
- Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. The Annals of Statistics, 16, 772–783.
- Jung, S., & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. The Annals of Statistics, 37, 4104–4130.
- Lin, Z., & Lu, C. (1996). Limit theory for mixing dependent random variables. Dordrecht: Kluwer Academic Publishers.
- Liu, Z., & Modarres, R. (2011). A triangle test for equality of distribution functions in high dimensions. Journal of Nonparametric Statistics, 23, 605–615.
- Lu, B., Greevy, R., Xu, X., & Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. The American Statistician, 65, 21–30.
- Maa, J. F., Pearl, D. K., & Bartoszyński, R. (1996). Reducing multidimensional two-sample data to onedimensional interpoint comparisons. *The Annals of Statistics*, 24, 1069–1074.
- Mondal, P. K., Biswas, M., & Ghosh, A. K. (2015). On high dimensional two-sample tests based on nearest neighbors. *Journal of Multivariate Analysis*, 141, 168–178.
- Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society, Series B*, 67, 515–530.
- Sarkar, S., & Ghosh, A. K. (2018). On some high dimensional two-sample tests based on averages of inter-point distances. Stat. 7, e187.
- Sarkar, S., & Ghosh, A. K. (2019). On perfect clustering of high dimension, low sample size data. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2019.2912599.
- Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81, 799–806.
- Székely, G. J., & Rizzo, M. L. (2004). Testing for equal distributions in high dimension. InterStat, 5.
- Székely, G. J., & Rizzo, M. L. (2013). Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143, 1249–1272.
- Tsukada, S.-I. (2019). High dimensional two-sample test based on the inter-point distance. *Computational Statistics*, 34, 599–615.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

