A consistent test of independence between two random vectors of arbitrary dimensions

Soham Sarkar*, Anil K. Ghosh†, and Alok Goswami‡

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India

Abstract

Several methods based on inter-point distances have been proposed in the literature to test for the independence between two random vectors $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ and $\mathbf{Y} \in \mathcal{Y} \subseteq \mathbb{R}^q$. It has been observed that the dependence between $\mathbf{X}$ and $\mathbf{Y}$ often leads to a strong positive or negative association between inter-point distances in $\mathcal{X}$ and $\mathcal{Y}$. Keeping that in mind, we propose and investigate a new test based on these inter-point distances. The proposed test has the large sample consistency under a fairly general class of alternatives. Moreover, it can be conveniently used even when the dimension of the data is larger than the sample size. Several simulated and real data sets are analyzed to demonstrate the usefulness of the proposed test.

Keywords: Inter-point distance; large sample consistency; measure of association; permutation test.

1 Introduction

Test of independence between two sets of measurement variables is an important problem in statistics. In fact, we often test for significance of the relationship between response variables and covariates before building any meaningful statistical model. In general, we assume that $\tilde{\mathbf{z}}_n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ is a sample of n independent observations from the distribution of $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top$, where $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$, $\mathbf{Y} \in \mathcal{Y} \subseteq \mathbb{R}^q$, and we test whether $\mathbf{X}$ and $\mathbf{Y}$ are independent. Several parametric and nonparametric methods are available for this problem. Most notable tests in the parametric regime include Wilk's Λ-test, Roy's largest root test, Hotelling-Lawley trace test and Pillai-Bartlett trace test (see, e.g., Anderson, 2003). These tests are mainly motivated by the normality assumption on the distribution of $\mathbf{Z}$, and they mainly check whether $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated. Hoeffding (1948) and Blum et al. (1961) constructed nonparametric tests based on empirical distribution functions, which can be used when both $\mathbf{X}$ and $\mathbf{Y}$ are one-dimensional. Other nonparametric tests for the univariate case (i.e., $p = q = 1$) include those based on Spearman's ρ, Kendall's τ and quadrant statistics (see, e.g., Blomqvist, 1950; Gibbons and Chakraborti, 2011). Taskinen et al. (2003) and Taskinen et al. (2005) constructed multivariate versions of these three tests based on spatial signs and ranks. Puri and Sen (1971) used coordinate-wise signs and ranks to develop some multivariate tests of independence. Gieser and Randles

*sohamsarkar1991@gmail.com
†akghosh@isical.ac.in
‡alok@isical.ac.in

(1997) extended the quadrant test to higher dimensions using the idea of interdirections. However, most of these above mentioned tests often fail to detect the dependance between $\mathbf{X}$ and $\mathbf{Y}$ when they are uncorrelated. Moreover, they usually yield poor results for moderately high dimensional data, and none of them can be used when the dimension of $\mathbf{Z}$ is larger than the sample size.

Friedman and Rafsky (1983) was the first to propose a test of independence based on graph theoretic approach, which can be conveniently used for data of arbitrary dimensions. Following the same spirit, recently some tests based on inter-point distances have been proposed. Heller et al. (2012) developed some distribution-free tests based on minimal spanning trees. Biswas et al. (2016) identified some limitations of these tests and proposed two modified tests that retain the distribution-free property. Among other tests based on inter-point distances, perhaps the most popular ones are the DCov test based on distance covariance (Székely et al., 2007), the HSIC test based on reproducing kernels (Gretton and Györfi, 2010) and the HHG test based on association of ranks of interpoint distances (Heller et al., 2013). These three tests are known to be consistent under general alternatives. Gretton and Györfi (2010) also constructed consistent tests based on L_1 and Kulback-Leibler distances, but implemented versions of these tests are not quite useful even for moderately high dimensional data, where most of the cells either remain empty or contain very few observations. The DCov test does not have such problems, but it often leads to poor results when the pairwise distances in $\mathcal{X}$ and $\mathcal{Y}$ have a strong negative association (see, e.g., Biswas et al., 2016). We will see that the HSIC test also has a similar problem. However, the HHG test usually works well in such situations (see, e.g., Heller et al., 2013; Biswas et al., 2016).

In the next section, we will see that the test statistic used by the HHG test is obtained by aggregating some measures of association, which do not pay attention to the nature of association between the pairwise distances in $\mathcal{X}$ and $\mathcal{Y}$. If this information is properly used, the resulting test can perform better. Keeping that in mind, we propose a new test, which utilizes this information.

2 The proposed test

Let $d_{\mathbf{X}}$ and $d_{\mathbf{Y}}$ be the distance functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. It is known that $\mathbf{X}$ and $\mathbf{Y}$ are dependent if and only if there exist some $\mathbf{a} \in \mathcal{X}, \mathbf{b} \in \mathcal{Y}$ and $r_x, r_y \in \mathbb{R}^+$ such that $\Pr(d_{\mathbf{X}}(\mathbf{X}, \mathbf{a}) \leq r_x, d_{\mathbf{Y}}(\mathbf{Y}, \mathbf{b}) \leq r_y) \neq \Pr(d_{\mathbf{X}}(\mathbf{X}, \mathbf{a}) \leq r_x)\ \Pr(d_{\mathbf{Y}}(\mathbf{Y}, \mathbf{b}) \leq r_y)$. However, these constants $\mathbf{a}, \mathbf{b}, r_x, r_y$ are not known in practice. Heller et al. (2013) made data driven choices of these constants to construct their test statistic. For each $\mathbf{z}_i = (\mathbf{x}_i^\top, \mathbf{y}_i^\top)^\top$ and $\mathbf{z}_j = (\mathbf{x}_j^\top,\ \mathbf{y}_j^\top)^\top$ $(1 \leq i \neq j \leq n)$, they computed $d_{ij}^{\mathbf{X}} = d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ and $d_{ij}^{\mathbf{Y}} = d_{\mathbf{Y}}(\mathbf{y}_i, \mathbf{y}_j)$. Then, for each $k \neq i, j$, depending on the $\mathbf{x}$-distance and the $\mathbf{y}$-distance between $\mathbf{z}_k$ and $\mathbf{z}_i$ (i.e., $d_{ik}^{\mathbf{X}}$ and $d_{ik}^{\mathbf{Y}}$), they put $\mathbf{z}_k$ in one of the four cells to construct the following 2×2 contingency table.

Table 1: 2×2 table based on inter-point distances for fixed $\mathbf{z}_i, \mathbf{z}_j$

	$d^{\mathbf{X}}_{ik} \leq d^{\mathbf{X}}_{ij}$	$d^{\mathbf{X}}_{ik} > d^{\mathbf{X}}_{ij}$	Total
$d^{\mathbf{Y}}_{ik} \leq d^{\mathbf{Y}}_{ij}$	$n_{11}(i,j)$	$n_{21}(i,j)$	$n_{\circ 1}(i,j)$
$d^{\mathbf{Y}}_{ik} > d^{\mathbf{Y}}_{ij}$	$n_{12}(i,j)$	$n_{22}(i,j)$	$n_{\circ 2}(i,j)$
Total	$n_{1\circ}(i,j)$	$n_{2\circ}(i,j)$	$n-2$

For each $\mathbf{z}_i$ and $\mathbf{z}_j$, they computed

$$S_{ij}(\tilde{\mathbf{z}}_n) = T^2_{ij}(\tilde{\mathbf{z}}_n) = \left[\frac{\{n_{11}(i,j)n_{22}(i,j) - n_{12}(i,j)n_{21}(i,j)\}}{\sqrt{n_{1\circ}(i,j)n_{2\circ}(i,j)n_{\circ 1}(i,j)n_{\circ 2}(i,j)}} \right]^2$$

as a measure of association between $\mathbf{x}$-distances and $\mathbf{y}$-distances and aggregated them to compute the test statistic

$$S_n = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} S_{ij}(\tilde{\mathbf{z}}_n).$$

The resulting HHG test rejects $\mathcal{H}_0$, the null hypothesis of independence, for large values of S_n. Note that for any fixed i and j, while $T_{ij}(\tilde{\mathbf{z}}_n)$ can be used as a one-sided statistic for measuring association between pairwise distances, $S_{ij}(\tilde{\mathbf{z}}_n)$ can be viewed as its two-sided analog, which considers the magnitude of $T_{ij}(\tilde{\mathbf{z}}_n)$ but ignores its sign. In practice, the dependence between $\mathbf{X}$ and $\mathbf{Y}$ is usually reflected by either strong positive or strong negative association between the inter-point distances in $\mathcal{X}$ and $\mathcal{Y}$. So, most of the $T_{ij}(\tilde{\mathbf{z}}_n)$'s are usually of the same sign. But, the HHG test does not use this information. This information can be taken into consideration if we use

$$T_n = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} T_{ij}(\tilde{\mathbf{z}}_n)$$

as the test statistic. Clearly, positive (respectively, negative) value of T_n indicates overall positive (respectively, negative) association among the distances in $\mathcal{X}$ and $\mathcal{Y}$, while the magnitude of T_n indicates the degree of association. So, we use $|T_n|$ as a test statistic and reject $\mathcal{H}_0$ for higher values of $|T_n|$. When most of the $T_{ij}(\tilde{\mathbf{z}}_n)$'s are of the same sign (which is usually the case in practice), T_n, being an aggregate of one-sided statistics, usually leads to better performance than the HHG test based on sum of two sided statistics $S_{ij}(\tilde{\mathbf{z}}_n)$. To see what we mean, first note that we can write

$$T^2_n = \frac{1}{\{n(n-1)\}^2} \left\{ \sum_{1 \leq i \neq j \leq n} T^2_{ij}(\tilde{\mathbf{z}}_n) + \sum_{\substack{i \neq j \\ (i,j) \neq (k,l)}} \sum_{k \neq l} T_{ij}(\tilde{\mathbf{z}}_n) T_{kl}(\tilde{\mathbf{z}}_n) \right\} = \frac{1}{n(n-1)} S_n + \frac{1}{\{n(n-1)\}^2} C_n.$$

The HHG test based on S_n only considers the information contained in the square terms, whereas our new test considers the information contained in the cross-product terms as well. Whenever there is a dominance of positive or negative $T_{ij}(\tilde{\mathbf{z}}_n)$ values, C_n is large

and positive. But, C_n is likely to be close to 0 under $\mathcal{H}_0$. The inclusion of this term usually gives our test an edge and helps it to perform better than the HHG test. Only in the cases, where the dependence between $\mathbf{X}$ and $\mathbf{Y}$ leads to a mixed type of association between the pairwise distances (i.e., there is no dominance of positive or negative signs among the values of $T_{ij}(\tilde{\mathbf{z}}_n)$), the HHG test based on S_n may perform better than the test based on $|T_n|$. However, such examples are not common in practice.

(a) Mixture Normal 1 (b) Mixture Normal 2

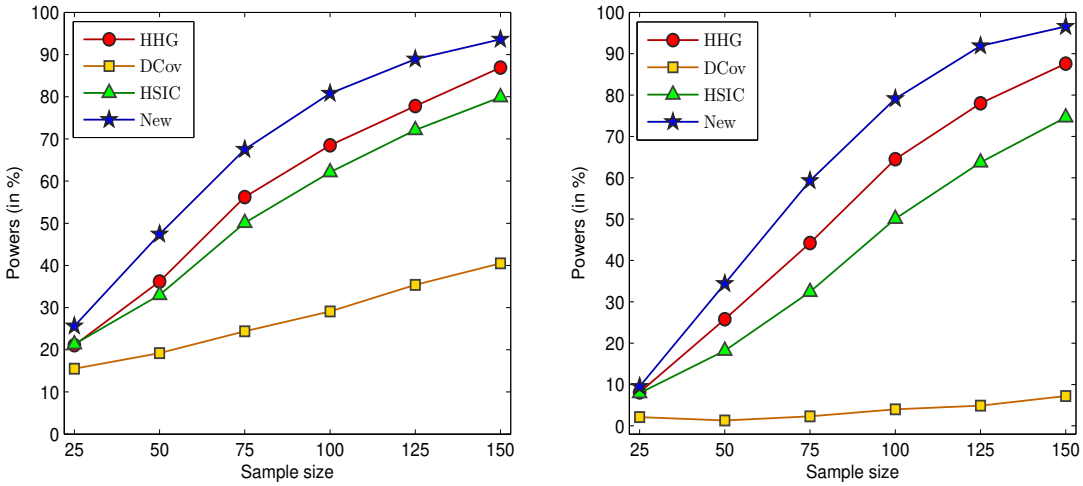

Figure 1: Powers of different tests in Mixture Normal 1 and 2.

To demonstrate the advantage of the proposed test over HHG and other consistent tests, we considered two simple examples. Let F_1 be an equal mixture of $\mathcal{N}_2(0,0,1,1,0)$ and $\mathcal{N}_2(0,0,9,9,0)$, and F_2 be an equal mixture of $\mathcal{N}_2(0,0,1,9,0)$ and $\mathcal{N}_2(0,0,9,1,0)$. Here, $\mathcal{N}_2(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ denotes a bivariate normal distribution with location $(\mu_1,\mu_2)^\top$, marginal variances σ_1^2,σ_2^2, and correlation ρ. We will refer to F_1 as Mixture Normal 1 and to F_2 as Mixture Normal 2. Note that, in Mixture Normal 1, variance of $\mathbf{X}$ is small (respectively, large) when that of $\mathbf{Y}$ is small (respectively, large). So, small (respectively, large) $\mathbf{x}$-distances correspond to small (respectively, large) $\mathbf{y}$-distances. Hence, the pairwise distances are positively correlated. However, the picture is reversed in Mixture Normal 2, where the pairwise distances are negatively associated. We generated 1000 samples consisting of n observations from these distributions, and estimated the powers of different tests by the proportion of times they rejected $\mathcal{H}_0$. These powers are shown in Figure 1 for different values of n. One can see that the new test performed much better than the HHG test on both occasions. We considered two other tests for comparison, namely the DCov test and the HSIC test. Recall that the DCov test (Székely et al., 2007) uses the test statistic

$$D_n = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ij} \Big/ \Big[\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^2\Big)^{1/2} \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}^2\Big)^{1/2} \Big],$$

where $A_{ij} = d^{\mathbf{X}}_{ij} - d^{\mathbf{X}}_{i\circ} - d^{\mathbf{X}}_{\circ j} + d^{\mathbf{X}}_{\circ\circ}$, $d^{\mathbf{X}}_{i\circ} = d^{\mathbf{X}}_{\circ i} = \sum_{j=1}^{n} d^{\mathbf{X}}_{ij}/n$ and $d^{\mathbf{X}}_{\circ\circ} = \sum_{i=1}^{n} d^{\mathbf{X}}_{i\circ}/n$. B_{ij}'s are similarly defined based on the $d^{\mathbf{Y}}_{ij}$s. The HSIC test (Gretton and Györfi, 2010) uses the statistic

$$H_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}\beta_{ij} - \frac{2}{n^3}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\alpha_{ij}\Big)\Big(\sum_{j=1}^{n}\beta_{ij}\Big) + \frac{1}{n^4}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}\Big)\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}\Big),$$

where $\alpha_{ij} = k^{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j), \beta_{ij} = k^{\mathbf{Y}}(\mathbf{y}_i, \mathbf{y}_j)$, for appropriate reproducing kernels $k^{\mathbf{X}}, k^{\mathbf{Y}}$ associated with reproducing kernel Hilbert spaces of functions on $\mathcal{X}$ and $\mathcal{Y}$. As we have mentioned before, both of these tests, especially DCov, often fail to yield satisfactory performance when $\mathbf{x}$-distances and $\mathbf{y}$-distances are negatively associated. We can observe the same in Figure 1(b) for the Mixture Normal 2 example. Even in the Mixture Normal 1 example, powers of these two tests were much lower than the other two tests considered here.

(a) Powers of different tests (b) Boxplot of C_n

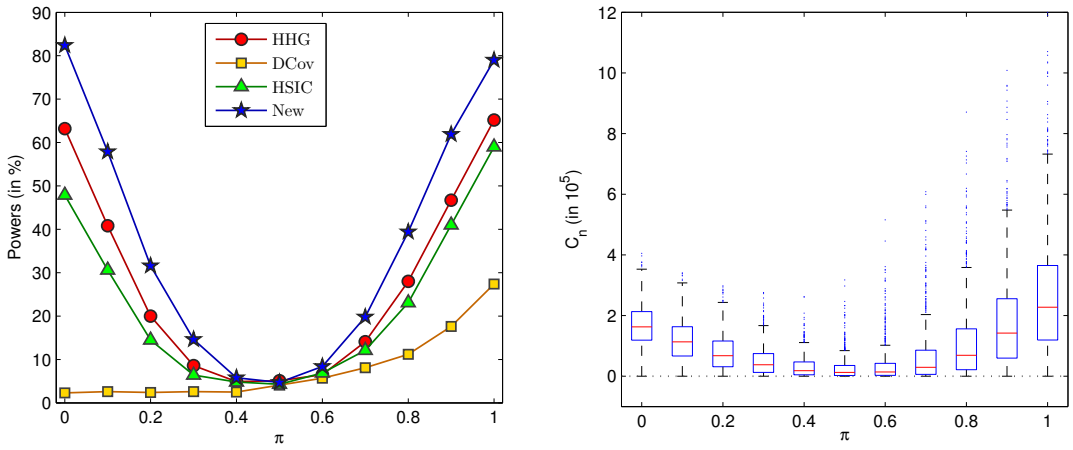

Figure 2: Powers of different tests and boxplots of cross-products for different values of π.

(a) $\pi = 0$ (b) $\pi = 0.5$ (c) $\pi = 1$

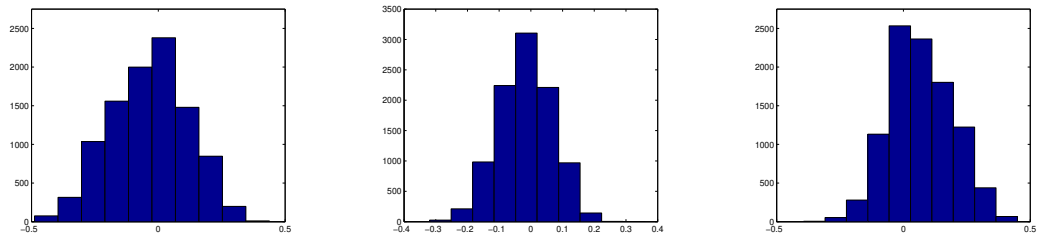

Figure 3: Histogram of T_{ij}'s for different values of π.

We considered another example to further strengthen our assertion that the proposed test benefits from strong positive or negative association between pairwise distances in $\mathcal{X}$ and $\mathcal{Y}$. In this case, for a fixed $\pi \in [0, 1]$, we generated 1000 samples, each consisting of 100 observations from a mixture distribution $\pi F_1 + (1-\pi)F_2$. The powers of different tests were

estimated as before, and they are reported in Figure 2(a) for $\pi = 0, 0.1, \ldots, 1$. Boxplots in Figure 2(b) also show the distribution of the values of C_n for these different choices of π . As we have mentioned before, for values of π close to 1 and 0 (which correspond to Mixture Normal 1 and 2, respectively), the pairwise distances have strong positive and negative association, respectively. Figure 2(b) shows that C_n takes much higher values in such cases. This led to much better performance by the test based on $|T_n|$. In Figure 2(a), we can see that this test outperformed the HHG test. The powers of DCov and HSIC tests were much lower than these two tests, especially for $\pi < 0.5$, where the distances were negatively associated. However, for values of π closer to 0.5, we have mixed types of association. In fact, for $\pi = 0.5$, $\mathbf{X}$ and $\mathbf{Y}$ turn out to be independent. In such cases, C_n takes values close to 0 with high probability. So, the advantage over the HHG test and other tests was not that evident. Figure 3 shows the histograms of $T_{ij}(\tilde{\mathbf{z}}_n)$ values for $\pi = 0, 0.5$ and 1. One can see that for $\pi = 0.5$, the distribution of the $T_{ij}(\tilde{\mathbf{z}}_n)$'s is almost symmetric around 0. But, for $\pi = 0$ (respectively, 1), they are negatively (respectively, positively) skewed, where we have a dominance of negative (respectively, positive) values. This dominance is usually observed when there is a dependence between $\mathbf{X}$ and $\mathbf{Y}$, and that leads to better performance by the test based on $|T_n|$ compared to that based on S_n.

3 Large sample consistency of the proposed test

For any fixed $\mathbf{z}_i = (\mathbf{x}_i^\top, \mathbf{y}_i^\top)^\top$, let us define two random variables $D_{\mathbf{X}}(\mathbf{z}_i) = d_{\mathbf{X}}(\mathbf{X}, \mathbf{x}_i)$ and $D_{\mathbf{Y}}(\mathbf{z}_i) = d_{\mathbf{Y}}(\mathbf{Y}, \mathbf{y}_i)$. Now, for any fixed $\mathbf{z}_j = (\mathbf{x}_j^\top, \mathbf{y}_j^\top)^\top$ $(j \neq i)$, we define

$$\pi_{11}(\mathbf{z}_i, \mathbf{z}_j) = \Pr\left(D_{\mathbf{X}}(\mathbf{z}_i) \leq d_{ij}^{\mathbf{X}}, D_{\mathbf{Y}}(\mathbf{z}_i) \leq d_{ij}^{\mathbf{Y}}\right), \quad \pi_{12}(\mathbf{z}_i, \mathbf{z}_j) = \Pr\left(D_{\mathbf{X}}(\mathbf{z}_i) \leq d_{ij}^{\mathbf{X}}, D_{\mathbf{Y}}(\mathbf{z}_i) > d_{ij}^{\mathbf{Y}}\right)$$
$$\pi_{21}(\mathbf{z}_i, \mathbf{z}_j) = \Pr\left(D_{\mathbf{X}}(\mathbf{z}_i) > d_{ij}^{\mathbf{X}}, D_{\mathbf{Y}}(\mathbf{z}_i) \leq d_{ij}^{\mathbf{Y}}\right) \quad \pi_{22}(\mathbf{z}_i, \mathbf{z}_j) = \Pr\left(D_{\mathbf{X}}(\mathbf{z}_i) > d_{ij}^{\mathbf{X}}, D_{\mathbf{Y}}(\mathbf{z}_i) > d_{ij}^{\mathbf{Y}}\right).$$

Clearly, $\pi_{kl}(\mathbf{z}_i, \mathbf{z}_j)$ is the population analog of $n_{kl}(i,j)/n$ for $k, l = 1, 2$. Therefore, the population counterpart of $T_{ij}(\tilde{\mathbf{z}}_n)$ is given by

$$g(\mathbf{z}_i, \mathbf{z}_j) = \frac{\pi_{11}(\mathbf{z}_i, \mathbf{z}_j)\pi_{22}(\mathbf{z}_i, \mathbf{z}_j) - \pi_{12}(\mathbf{z}_i, \mathbf{z}_j)\pi_{21}(\mathbf{z}_i, \mathbf{z}_j)}{\sqrt{\pi_{1\circ}(\mathbf{z}_i, \mathbf{z}_j)\pi_{2\circ}(\mathbf{z}_i, \mathbf{z}_j)\pi_{\circ 1}(\mathbf{z}_i, \mathbf{z}_j)\pi_{\circ 2}(\mathbf{z}_i, \mathbf{z}_j)}};$$

where $\pi_{k\circ}(\mathbf{z}_i, \mathbf{z}_j) = \pi_{k1}(\mathbf{z}_i, \mathbf{z}_j) + \pi_{k2}(\mathbf{z}_i, \mathbf{z}_j)$ and $\pi_{\circ l}(\mathbf{z}_i, \mathbf{z}_j) = \pi_{1l}(\mathbf{z}_i, \mathbf{z}_j) + \pi_{2l}(\mathbf{z}_i, \mathbf{z}_j)$. Note that conditionally given $\mathbf{z}_i$ and $\mathbf{z}_j$, using the almost sure convergence for sums of independent and identically distributed random variables and the continuous mapping theorem, one can show that $T_{ij}(\tilde{\mathbf{z}}_n) \xrightarrow{a.s.} g(\mathbf{z}_i, \mathbf{z}_j)$. This almost sure convergence also holds unconditionally (see Lemma 1 in the Appendix). Using this result we can prove the convergence of T_n, which is given by the following theorem.

Theorem 1. *If $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent and identically distributed, then T_n converges to $g = E\{g(\mathbf{Z}_1, \mathbf{Z}_2)\}$ in probability as n tends to infinity.*

Note that for fixed $\mathbf{z}_i$ and $\mathbf{z}_j$, if we define $\zeta_{ij}(\mathbf{X}) = \mathbb{I}(D_{\mathbf{X}}(\mathbf{z}_i) \leq d_{ij}^{\mathbf{X}})$ and $\zeta_{ij}(\mathbf{Y}) = \mathbb{I}(D_{\mathbf{Y}}(\mathbf{z}_i) \leq d_{ij}^{\mathbf{Y}})$, then $g(\mathbf{z}_i, \mathbf{z}_j) = \text{corr}\{\zeta_{ij}(\mathbf{X}), \zeta_{ij}(\mathbf{Y})\}$. So, $g = \text{E}[\text{corr}\{\zeta_{ij}(\mathbf{X}), \zeta_{ij}(\mathbf{Y})\}]$ can

be seen as a measure of overall association between the distances in $\mathcal{X}$ and $\mathcal{Y}$. Under $\mathcal{H}_0$, we have $g(\mathbf{z}_i, \mathbf{z}_j) = 0$ for each fixed $\mathbf{z}_i$ and $\mathbf{z}_j$. So, g turns out to be 0 as well. But, if $\mathbf{X}$ and $\mathbf{Y}$ are dependent, g is expected to take either positive or negative value depending on the nature of association between the pairwise distances. The test based on $|T_n|$ turns out to be consistent in such situations. This result is stated below.

Theorem 2. *If $g = E\{g(\mathbf{Z}_1, \mathbf{Z}_2)\} \neq 0$, the power of the proposed test based on $|T_n|$ converges to unity as the sample size n tends to infinity.*

Note that our proposed test can be conveniently used for high dimension, low sample size data and even for functional data taking values in an infinite dimensional Banach space. In fact, the large sample consistency of the test holds whenever $\mathbf{Z}$ takes values in a separable metric space. Conditions for the consistency as well as the proofs remain the same as in Theorems 1 and 2.

4 Simulation study

We used five simulated examples to compare the level and the power properties of our proposed test with HHG (Heller et al., 2013), DCov (Székely et al., 2007) and HSIC (Gretton and Györfi, 2010) tests. For each example, we used data sets with dimensions $p = q = 1$ and $p = q = 5$. In each case, we generated 1000 random samples of different sizes to estimate the powers (sizes under $\mathcal{H}_0$) of different tests as before. For HHG and DCov tests, we used the codes available in `R` packages `HHG` and `energy`, respectively. For the HSIC test, we used the codes available at http://www.gatsby.ucl.ac.uk/~gretton/indepTestFiles/indep.htm. Note that none of these tests are distribution-free. So, we used conditional tests based on 1000 random permutations. We also considered the multivariate tests based on coordinate-wise ranks and spatial ranks discussed in Section 1. But their overall performance was much inferior to the tests considered here. So, we decided not to report those results in this article. Because of the same reason, here we do not report the results for the tests proposed by Friedman and Rafsky (1983), Heller et al. (2012) and Biswas et al. (2016). Throughout this article, all tests are considered to have 5% nominal level.

We started by examining the level properties of different tests. For this purpose, we generated observations on $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top$ from the $(p+q)$-variate standard normal distribution. So, in this example, $\mathbf{X}$ and $\mathbf{Y}$ were independent. Observed sizes of different tests are reported in Table 2 for $n = 15, 30$ and 45. This table shows that all tests had observed sizes close to their nominal levels.

To study the power properties of different tests, in the next four examples, we generated observations having various dependence structures. Observed powers of different tests for varying sample sizes are shown in Figures 4 and 5 for $p = q = 1$ and $p = q = 5$, respectively. In the first of these example, we generated observations from the standard multivariate t distribution with 5 degrees of freedom. So, $\mathbf{X}$ and $\mathbf{Y}$ were uncorrelated but

Table 2: Observed sizes of different tests in Normal example (with 5% nominal level).

	$p = q = 1$				$p = q = 5$			
n	HHG	DCov	HSIC	New	HHG	DCov	HSIC	New
15	5·0	5·0	5·4	4·8	4·2	4·9	4·7	5·4
30	4·3	4·3	5·2	3·5	4·2	5·6	5·2	4·4
45	4·8	5·3	4·5	4·5	3·5	4·2	4·6	4·1

not independent. Here, our proposed test outperformed all its competitors, though the HHG test also had competitive performance for $p = q = 5$.

In the next example, observations were generated from the uniform distribution over the $(p + q)$ dimensional unit sphere. In this example also, $\mathbf{X}$ and $\mathbf{Y}$ were uncorrelated but not independent. Again, our proposed test outperformed its competitors for both $p = q = 1$ and $p = q = 5$. The DCov test had poor performance in both cases. For $p = q = 1$, HSIC and HHG tests had similar powers, while HHG had a slight edge for larger sample sizes. However, the HSIC test had poor performance for $p = q = 5$. Like the DCov test, in this case, it had powers smaller than the nominal level.

In the last two examples, we considered additive and multiplicative error models. In both cases, we generated each coordinate of $\mathbf{X}$ independently from $\mathcal{U}(-1, 1)$. In the '$\mathbf{y} = \mathbf{x} + \epsilon$' example, each coordinate of $\mathbf{Y}$ was generated by adding an independent standard Gaussian noise to the corresponding coordinate of $\mathbf{X}$. In this example also, the proposed test performed much better than the HHG test. The DCov test had the best performance in this example. The proposed test had the second highest powers for $p = q = 1$, but for $p = q = 5$, it was outperformed by the HSIC test. In the '$\mathbf{y} = \mathbf{x}\epsilon$' example, we generated each coordinate of $\mathbf{Y}$ by multiplying an independent standard Gaussian noise with the corresponding coordinate of $\mathbf{X}$. In this example, the proposed test had substantially higher powers than its competitors. The DCov test performed poorly both for $p = q = 1$ and $p = q = 5$, while the HHG test also had much lower power in the latter case.

5 Real data analysis

We analyzed two real data sets, the `Canadian Weather` data and the `Gait` data, for further evaluation of the proposed test. These data sets are available in the `R` package `fda`, and their detailed descriptions can be found in Ramsay and Silverman (2005). To have a meaningful comparison among different tests, following the idea of Biswas et al. (2016), we generated random samples of different sizes from these data sets. For each sample size, the random generation was repeated 1000 times, and the observed powers of different tests are shown in Figure 6.

The `Canadian Weather` data set contains daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994. Here, we were interested to check for the existence of a relationship between the time series of temperature and precipitation so that precipitation can be predicted based on temperature. In this example,

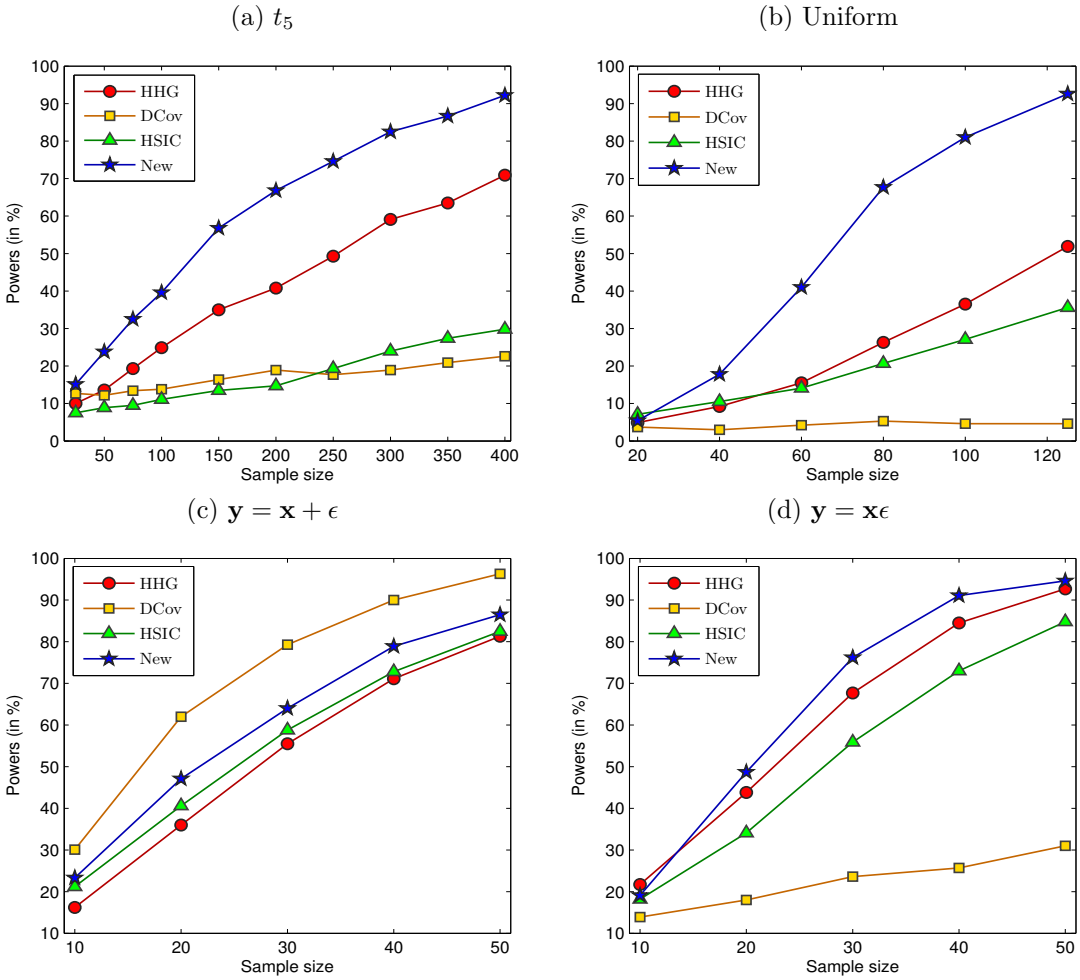

Figure 4: Powers of different tests in simulated data sets ($p = q = 1$).

all competing tests performed well, but the proposed test had an edge over its competitors.

The `Gait` data set consists of angular rotations in the sagittal plane of the hip and the knee of 39 normal 5 year old children. This data set was collected at the Motion Analysis Laboratory at the Children's Hospital, San Diego. The observations were taken at 20 different points over a gait cycle consisting of one double step taken by each child. Leurgans et al. (1993) used this data set in the context of functional canonical correlation analysis. Here it is of interest to know whether there is any significant relationship between the knee angles and the hip angles, which can be modeled to understand the walking mechanism of children. Figure 6 clearly shows that the powers of the proposed test were substantially higher compared to the other three tests in this example

6 Concluding remarks

In this article, we have proposed a test of independence between two random vectors of arbitrary dimensions and proved its large sample consistency under a fairly general class of alternatives. This test can be conveniently used for high dimension, low sample size

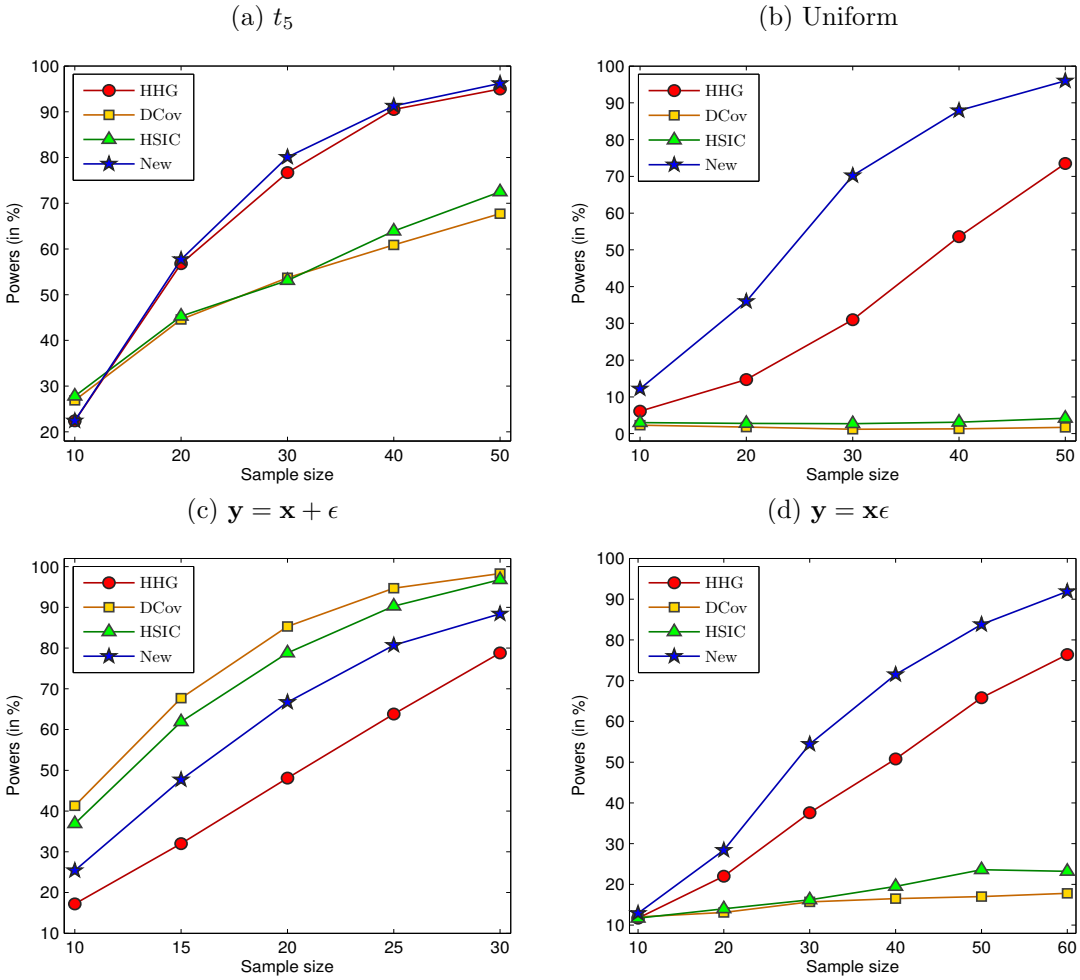

Figure 5: Powers of different tests in simulated data sets ($p = q = 5$).

data and even for data taking values in a separable metric space. Unlike DCov and HSIC tests, this test does not get affected by the presence of negative correlation among pairwise distances, and it usually performs better than the HHG test in practice. Analyzing several simulated and real data sets in this article, we have amply demonstrated these important features of the proposed test. Throughout this article, for any fixed (i, j), we have used $(\mathbf{X}_i, \mathbf{Y}_i)$ as centers and $(\|\mathbf{X}_i - \mathbf{X}_j\|, \|\mathbf{Y}_i - \mathbf{Y}_j\|)$ as radii to construct balls in $\mathcal{X}$ and $\mathcal{Y}$ and hence to compute $T_{ij}(\tilde{\mathbf{z}}_n)$. However, for fixed (i, j, k), one can also use $(\mathbf{X}_i, \mathbf{Y}_i)$ as centers and $(\|\mathbf{X}_j - \mathbf{X}_k\|, \|\mathbf{Y}_j - \mathbf{Y}_k\|)$ as radii, and repeat it for different choices of i, j and k to compute the test statistic and to perform the test. This construction requires more computing time than the proposed method, but our empirical experience suggests that it does not lead to any substantial improvement in the performance of the resulting test.

Appendix. Proofs and Mathematical details

Lemma 1. *Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ be independent d-dimensional random vectors defined on probability spaces $(\Omega_i, \mathcal{A}_i, \Pr_i)_{i \geq 1}$ and $(\Omega, \mathcal{A}, \Pr)$ be the product probability space. Let*

(a) Canadian Weather ($p = q = 365$) (b) Gait ($p = q = 20$)

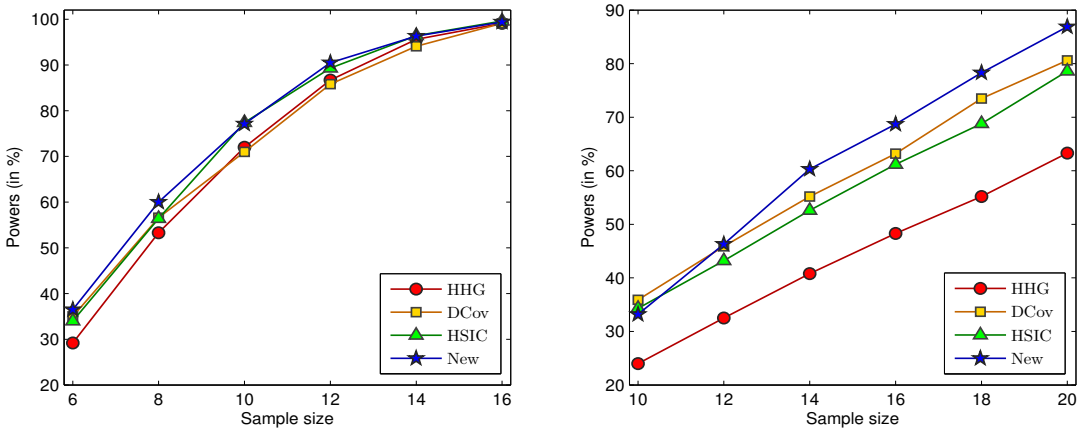

Figure 6: Powers of different tests in real data sets.

$f_n : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \to \mathbb{R}, n = 1, 2, \ldots$ and $f : \mathbb{R}^d \to \mathbb{R}$ be functions such that for every fixed $\omega_1 \in \Omega_1$, $f_n(\mathbf{Z}_1(\omega_1), \mathbf{Z}_2, \ldots, \mathbf{Z}_n) \overset{a.s.}{\to} f(\mathbf{Z}_1(\omega_1))$ as $n \to \infty$. Then $f_n(\mathbf{Z}_1, \ldots, \mathbf{Z}_n) \overset{a.s.}{\to} f(\mathbf{Z}_1)$ as $n \to \infty$.

Proof: Let $A = \{(\omega_1, \omega_2, \ldots) \in \Omega : f_n(\mathbf{Z}_1(\omega_1), \ldots, \mathbf{Z}_n(\omega_n)) \to f(\mathbf{Z}_1(\omega_1)) \text{ as } n \to \infty\}$. Let $A(\cdot) = \{(\cdot, \omega_2, \ldots) \in \Omega : f_n(\mathbf{Z}_1(\cdot), \ldots, \mathbf{Z}_n(\omega_n)) \to f(\mathbf{Z}_1(\cdot)) \text{ as } n \to \infty\}$. Then from our assumption, $\Pr(A(\omega_1)) = 1$ for every $\omega_1 \in \Omega_1$. Hence $\Pr(A) = \int \Pr(A(\omega_1)) dP_1(\omega_1) = 1$, which implies that the event A occurs almost surely. This proves the lemma. □

As a consequence of Lemma 1 we get the following lemma, which is helpful in our context.

Lemma 2. *If $\tilde{\mathbf{Z}}_n = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ denotes a collection of n independent copies of $\mathbf{Z}$, $T_{ij}(\tilde{\mathbf{Z}}_n) \to g(\mathbf{Z}_i, \mathbf{Z}_j)$ almost surely as the sample size n diverges to infinity.*

Proof of Theorem 1: In order to prove the theorem, we shall show that

$$(i) \quad \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \left\{ T_{ij}(\tilde{\mathbf{Z}}_n) - g(\mathbf{Z}_i, \mathbf{Z}_j) \right\} \to 0 \text{ in probability as } n \to \infty \text{ and}$$

$$(ii) \quad \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} g(\mathbf{Z}_i, \mathbf{Z}_j) \to g \text{ in probability as } n \to \infty$$

Now, for part (i), using Markov's inequality, we get

$$\Pr\left[\left| \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \left\{ T_{ij}(\tilde{\mathbf{Z}}_n) - g(\mathbf{Z}_i, \mathbf{Z}_j) \right\} \right| > \epsilon \right] \le \frac{\mathrm{E}\left[\sum_{i \ne j} \{ T_{ij}(\tilde{\mathbf{Z}}_n) - g(\mathbf{Z}_i, \mathbf{Z}_j) \} \right]^2}{n^2 (n-1)^2 \epsilon^2}.$$

$$
\begin{aligned}
&\text{Now, } \mathrm{E}\Big[\sum_{i\neq j}\big\{T_{ij}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_i,\mathbf{Z}_j)\big\}\Big]^2\\
&\quad=\sum_{i\neq j}\mathrm{E}\big\{T_{ij}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_i,\mathbf{Z}_j)\big\}^2+\sum_{\substack{i\neq j\\(i,j)\neq(k,l)}}\sum_{k\neq l}\mathrm{E}\big[\big\{T_{ij}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_i,\mathbf{Z}_j)\big\}\big\{T_{kl}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_k,\mathbf{Z}_l)\big\}\big]\\
&\quad=n(n-1)\mathrm{E}\big\{T_{12}(\tilde{\mathbf{z}}_n)-g(\mathbf{Z}_1,\mathbf{Z}_2)\big\}^2\\
&\qquad+2n(n-1)(n-2)\mathrm{E}\big\{T_{12}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_1,\mathbf{Z}_2)\big\}\mathrm{E}\big\{T_{13}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_1,\mathbf{Z}_3)\big\}\\
&\qquad+n(n-1)(n-2)(n-3)\mathrm{E}\big\{T_{12}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_1,\mathbf{Z}_2)\big\}\mathrm{E}\big\{T_{34}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_3,\mathbf{Z}_4)\big\}
\end{aligned}
$$

For every $1\leq i\neq j\leq n$, $|T_{ij}(\tilde{\mathbf{Z}}_n)|, |g(\mathbf{Z}_i,\mathbf{Z}_j)|\leq 1$ with probability 1. So, the first two terms on the right-hand side are $\mathbf{o}_p(n^4)$. Since $T_{ij}(\tilde{\mathbf{Z}}_n)\overset{a.s}{\to} g(\mathbf{Z}_i,\mathbf{Z}_j)$ as $n\to\infty$ (see Lemma 2), using the Dominated Convergence theorem, we get $\mathrm{E}\big\{T_{ij}(\tilde{\mathbf{Z}}_n)-g(\mathbf{Z}_i,\mathbf{Z}_j)\big\}\to 0$ as $n\to\infty$. This proves the first part.

For part (ii), first note that

$$
\mathrm{E}\Big[\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}g(\mathbf{Z}_i,\mathbf{Z}_j)\Big]=\mathrm{E}\{g(\mathbf{Z}_1,\mathbf{Z}_2)\}=g.
$$

Now, $g(\mathbf{Z}_i,\mathbf{Z}_j)$ and $g(\mathbf{Z}_k,\mathbf{Z}_l)$ are independent if (i,j) and (k,l) are disjoint. Therefore,

$$
\begin{aligned}
&\mathrm{var}\Big[\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}g(\mathbf{Z}_i,\mathbf{Z}_j)\Big]\\
&=\frac{1}{n(n-1)}\mathrm{var}\{g(\mathbf{Z}_1,\mathbf{Z}_2)\}+\frac{n(n-1)(n-2)}{n^2(n-1)^2}\big[\mathrm{cov}\{g(\mathbf{Z}_1,\mathbf{Z}_2),g(\mathbf{Z}_1,\mathbf{Z}_3)\}+\mathrm{cov}\{g(\mathbf{Z}_1,\mathbf{Z}_3),g(\mathbf{Z}_2,\mathbf{Z}_3)\}\big]\\
&\leq 4\left\{\frac{1}{n(n-1)}+\frac{2(n-2)}{n(n-1)}\right\}\to 0 \text{ as } n\to\infty.
\end{aligned}
$$

This proves the second part. Now, combining part (i) and part (ii), we get the proof of the theorem. □

Proof of Theorem 2: Under the null hypothesis, $|T_n|\to 0$ in probability. So, for any fixed $\alpha\in(0,1)$, the cut-off of the proposed test, say $c_n(\alpha)$, also converges to zero in probability. Now, under the alternative, $|T_n|\to|g|>0$ in probability. Therefore, the power of the test $\Pr(|T_n|>c_n(\alpha))\to 1$ as n tends to infinity. □

References

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Biswas, M., Sarkar, S. and Ghosh, A. K. (2016) On some exact distribution-free tests of independence between two random vectors of arbitrary dimensions. *J. Statist. Plann. Inference*, **175**, 78–86.

Blomqvist, N. (1950) On a measure of dependence between two random variables. *Ann. Math.*

Statist., **21**, 593–600.

Blum, J. R., Kiefer, J. and Rosenblatt, M. (1961) Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.*, **32**, 485–498.

Friedman, J. H. and Rafsky, L. C. (1983) Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.*, **11**, 377–391.

Gibbons, J. D. and Chakraborti, S. (2011) *Nonparametric Statistical Inference*. CRC Press, Boca Raton, Florida.

Gieser, P. W. and Randles, R. H. (1997) A nonparametric test of independence between two vectors. *J. Amer. Statist. Assoc.*, **92**, 561–567.

Gretton, A. and Györfi, L. (2010) Consistent nonparametric tests of independence. *J. Mach. Learn. Res.*, **11**, 1391–1423.

Heller, R., Gorfine, M. and Heller, Y. (2012) A class of multivariate distribution-free tests of independence based on graphs. *J. Statist. Plann. Inference*, **142**, 3097–3106.

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

Hoeffding, W. (1948) A non-parametric test of independence. *Ann. Math. Statist.*, **19**, 546–557.

Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993) Canonical correlation analysis when the data are curves. *J. Royal Statist. Soc. Ser. B*, **55**, 725–740.

Puri, M. and Sen, P. K. (1971) *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis*. Springer, New York.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

Taskinen, S., Kankainen, A. and Oja, H. (2003) Sign test of independence between two random vectors. *Statist. Probab. Letters*, **62**, 9–21.

Taskinen, S., Oja, H. and Randles, R. H. (2005) Multivariate nonparametric tests of independence. *J. Amer. Statist. Assoc.*, **100**, 916–925.