# Interpoint-Ranking Sign Covariance for Test of Independence

BY HAEUN MOON AND KEHUI CHEN

*Department of Statistics, University of Pittsburgh,*
*Pittsburgh, Pennsylvania 15213, U.S.A*
ham98@pitt.edu khchen@pitt.edu

## SUMMARY

We generalize the sign covariance introduced by Bergsma & Dassios (2014) to multivariate random variables and beyond. The new interpoint-ranking sign covariance is applicable to general types of random objects as long as a meaningful similarity measure can be defined, and it is shown to be zero if and only if the two random variables are independent. The test statistic is a $U$-statistic, whose large sample behavior guarantees that the proposed test is consistent against general types of alternatives. Numerical experiments and data analyses demonstrate the great empirical performance of the proposed method.

*Some key words*: Consistent; Independence Test; Interpoint distance; Nonparametric; Sign Covariance.

## 1. INTRODUCTION

Let $X$ and $Y$ be random variables with marginal distributions $P_X$ on $\mathcal{X}$ and $P_Y$ on $\mathcal{Y}$, respectively, and joint distribution $P_{XY}$ on $\mathcal{X} \times \mathcal{Y}$. In this paper, we aim at testing

$$H_0 : P_{XY} = P_X P_Y \text{ versus } H_1 : P_{XY} \neq P_X P_Y$$

based on samples $(x_1, y_1)$,...,$(x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ of size $n$ drawn independently and identically from $P_{XY}$. This fundamental statistical question has received much attention with a wide range of applications.

For a simple case, $(\mathcal{X}, \mathcal{Y}) = (\mathbb{R}, \mathbb{R})$, there exists classical measures of association such as Pearson correlation (Pearson, 1895), Kendall's $\tau$ (Kendall, 1938), and Spearman's $\rho$ (Spearman, 1904). However, it has been shown that they could be zero even in the presence of dependence between random variables, i.e., these tests are not consistent against general types of alternatives. Two alternative tests based on Hoeffding's $D$ coefficient (Hoeffding, 1948) and Blum-Kiefer-Rosenblatt's $R$ coefficient (Blum et al., 1961) were developed with better consistency guarantees. In recent years, there have been active attempts to develop or understand tests of independence that are consistent against general types of alternatives, for example, the distance covariance test or its extensions (Székely et al., 2007; Székely & Rizzo, 2013; Lyons et al., 2013; Shen et al., 2019; Zhu et al., 2020), the Hilbert-Schmidt independence criterion (Gretton et al., 2005, 2008; Sejdinovic et al., 2013), the mutual information test (Berrett & Samworth, 2019). We draw particular attention to a rank-based nonparametric test, the sign covariance test, which denoted by $\tau^*$, was introduced as a modification of Kendall's $\tau$ by Bergsma & Dassios (2014). Based on the concordance and discordance of four points rather than two points as $\tau$ does, the sign covariance test is consistent against general types of alternatives and meanwhile, it enjoys robustness, simplicity and interpretability due to its rank-based nature (Bergsma & Dassios, 2014; Nandy et al., 2016; Dhar et al., 2016; Weihs et al., 2018). However, the original Bergsma & Dassios (2014) paper only developed tests for bivariate distributions, and they particularly noted that a naive generalization to the multivariate setting by replacing the absolute value $|\cdot|$ with a general metric will lose the consistency property. In this paper, we introduce an interpoint-ranking sign covariance and develop a nonparametric test of independence based on it. Heuristically, for each point $(x, y)$ in the domain $\mathcal{X} \times \mathcal{Y}$, we rank all the other $X$ data points according to the distance or similarity to this particular point $x$, do the same for $Y$ and then compute the sign covariance with regard to the fixed point $(x, y)$ using the interpoint ranks as data in the original formula. The final statistic is an average over all point $(x, y)$ in $\mathcal{X} \times \mathcal{Y}$, and the test statistic has a $U$-statistic. The proposed test of independence is applicable to general types of random objects as long as a meaningful similarity measure can be defined, and it is consistent against general types of alternatives.

Generalizing rank-based tests from univariate data to multivariate data is an active research area. A straightforward extension was explored in Leung et al. (2018) and Drton et al. (2020), which derived a family of test statistics for testing mutual independence by collecting all pairwise dependent signals, where $\tau^*$ can be readily computed for each pair of one-dimensional variables. Another appealing idea is to use the projection approach. Kim et al. (2020) illustrated that using a projection-averaging approach, the sign covariance independence test can be generalized to multivariate data through integrations of the projected univariate $\tau^*$ over the unit sphere. A similar projection idea has been suggested by Zhu et al. (2017) to generalize Hoeffding's $D$ (Hoeffding, 1948) to multivariate cases. Instead of projecting to a one-dimensional space, there are also works trying to define multivariate ranks directly. Two recent papers, Deb & Sen (2019) and Shi et al. (2020), developed tests of independence for multivariate variables based on a recent breakthrough in Hallin's multivariate rank (del Barrio et al., 2018), where one first discretizes the unit ball $S_d$ to $n$ grid points with a well-defined ordering and obtain multivariate rank through an optimal coupling between the observed data points and the grid points. These above-mentioned tests have appealing properties in $\mathbb{R}^d$ settings, but in general are not extendable to more general types of data, such as spherical surfaces, planar graph, symmetric positive matrices equipped with Riemannian geometry and manifold-valued functional data. Such complex data increasingly arise in practice (Free et al., 2001; Zheng, 2015; Dai et al., 2018; Adriaenssens et al., 2011; Masucci et al., 2009). There are earlier works using interpoint distances to rank multivariate data or build graphical tests, for example, Mantel (1967); Friedman et al. (1983); Biswas et al. (2016); Sarkar & Ghosh (2018); Guo & Modarres (2020). However, these tests are not consistent against general types of alternatives, and we focus on developing a consistent nonparametric test in this paper. During the revision of the manuscript, we found a paper Heller & Heller (2016a) that proposed to transform multivariate $K$-sample and independence tests to univariate tests by comparing the univariate distributions of the distances from a selected center point, and also discussed several ways to combine the results from multiple center points. Their numerical experiments were only for two-sample tests in $\mathbb{R}^d$. Finally, two related alternatives are Heller et al. (2012) and Pan et al. (2019). The "HHG" method (Heller et al., 2012) transforms the original problem into many aggregated 2 $\times$2 contingency tables and use the Pearson's $\chi^2$ test of independence. The ball covariance method (Pan et al., 2019) defines a class of Ball covariance measures by integrating the Hoeffding's dependence measure on the coordinate of radius over poles and is applicable to Banach spaces. Interestingly, the "HHG" method, the ball covariance with their recommended weight functions and the proposed method all depend on the ranking of interpoint distances, although they are derived from three different perspectives.

## 2. INTERPOINT-RANKING SIGN COVARIANCE

Bergsma & Dassios (2014) proposed a sign covariance as a new rank-based measure of independence between two random variables. When $X, Y \in \mathbb{R}$ are random variables and $(X^1, Y^1), \ldots, (X^4, Y^4)$ are independent and identically distributed copies of $(X, Y)$, the population version of sign covariance is defined as

$$\tau^*(X, Y) = E\{a(X^1, X^2, X^3, X^4)a(Y^1, Y^2, Y^3, Y^4)\},$$

where

$$a(z_1, z_2, z_3, z_4) = sign(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|) \tag{1}$$

for $z_1, z_2, z_3, z_4 \in \mathbb{R}$.

While Kendall's $\tau$, defined as $E\{sign(X^1 - X^2)(Y^1 - Y^2)\}$, is based on the probabilities of ordinal structure between two points, $\tau^*$ is based on the probabilities of ordinal structure between four points so it can be viewed as a modified Kendall's $\tau$. The main theorem in Bergsma & Dassios (2014) states that when $(X, Y)$ has a bivariate discrete or continuous distribution, or a mixture of the two, $\tau^*(X, Y) \geq 0$ with equality if and only if $X$ and $Y$ are independent. Moreover, the authors of that paper conjectured that this property holds in general without continuous or discrete conditions. Actually, the conditions have been eased later, though not completely, to the extent that $X$ and $Y$ have both continuous marginal distributions (Drton et al., 2020). Our simulation results support the conjecture of the authors.

Regarding a generalization to multivariate variables, the authors mentioned the definition of $\tau^*$ can be straightforwardly extended to variables in an arbitrary metric space, by defining $a_d(z_1, z_2, z_3, z_4) = sign\{d(z_1, z_2) + d(z_3, z_4) - d(z_1, z_3) - d(z_2, z_4)\}$. But in this case, $\tau^*$ might be smaller than zero and the consistency of the $\tau^*$ test does not hold in general. In the following, we extend sign covariance to a Banach space while preserving a general consistency.

Let $(\mathcal{X}, \rho)$, $(\mathcal{Y}, \zeta)$ be two separable Banach spaces, where $\rho$ and $\zeta$ also represent distances induced by norms. Let $\theta$ be a Borel probability measure on $\mathcal{X} \times \mathcal{Y}$ with marginals $\mu$ on $\mathcal{X}$ and $\nu$ on $\mathcal{Y}$. Let $(X, Y)$ be a pair of random variables where $(X, Y) \sim \theta$, $X \sim \mu$ and $Y \sim \nu$. Let $(X^0, Y^0)$,…,$(X^4, Y^4)$ be independent and identically distributed copies of $(X, Y)$.

DEFINITION 1. *The interpoint-ranking sign covariance, or IPR-$\tau^*$, is defined as*

$$\textit{IPR-}\tau^*(X, Y) = E[a\{\rho(X^0, X^1), \rho(X^0, X^2), \rho(X^0, X^3), \rho(X^0, X^4)\} \\ a\{\zeta(Y^0, Y^1), \zeta(Y^0, Y^2), \zeta(Y^0, Y^3), \zeta(Y^0, Y^4)\}],$$

*where $a(z_1, z_2, z_3, z_4)$ is defined in Eq. (1).*

In Definition 1, for a fixed reference point $(X^0, Y^0)$, we can view $\rho(X^0, X^i)$ and $\zeta(Y^0, Y^i)$ as $X$- and $Y$-interpoint distances arising from $(X^0, Y^0)$. The univariate $\tau^*$ can be applied to $\rho(X^0, X^i)$ and $\zeta(Y^0, Y^i)$. Then we view the reference point $(X^0, Y^0)$ as an extra independent copy in the definition. The interpoint-ranking sign covariance collects the signal of dependency between $X$- and $Y$-interpoint distances arising from all anchor points $(X^0, Y^0)$ in $\mathcal{X} \times \mathcal{Y}$. Here IPR-$\tau^*$ is defined using five independent copies of $(X, Y)$ rather than four copies as used in the original sign covariance, which can be considered as the expense of the extended domain. IPR-$\tau^*$ remain invariant under the monotone transformation of distances since $a$ represents a coordination structure of interpoint distances.

Empirically, if there are $n$ copies of the data $X$, one can compute an $n \times n$ distance matrix $[\rho_{ij}]$ for $X$, with $n(n-1)/2$ distinct values. For each fixed data point $i$, ranking all other points based on the order of $\rho_{ij}$ gives the interpoint ranking $R_{i(j)}$. Interpoint distance has been used in various ways to characterize the distribution and geometry of multivariate data, including some independence tests. However, most of them directly built a dependence measure on the two vectors containing $n(n-1)/2$ $X$-interpoint distances and $n(n-1)/2$ $Y$-interpoint distances, which did not produce consistent tests in general.

Theorem 1 states that this new coefficient is nonnegative and becomes zero if and only if $X$ and $Y$ are independent, provided that the joint probability distribution is discrete or continuous, or a mixture of the two.

THEOREM 1. *Let $(\mathcal{X}, \rho)$, $(\mathcal{Y}, \zeta)$ be two separable Banach spaces and $\theta$ be a Borel probability measure on $\mathcal{X} \times \mathcal{Y}$ with marginals $\mu$ on $\mathcal{X}$ and $\nu$ on $\mathcal{Y}$. Assume $\theta$ is discrete or continuous, or a mixture of the two, that is, there exists a probability mass function $P_{XY}$ and a density function $h$ such that*

$$\theta(A \times B) = \sum_{x_i \in A, y_i \in B} P_{XY}(x_i, y_i) + \int_{A \times B} h(x, y) G(dx) G(dy),$$

*where $A \subset \mathcal{X}$, $B \subset \mathcal{Y}$ are any two open sets and $G$ is the Abstract Wiener measure on $\mathcal{X}$ and $\mathcal{Y}$. In addition, assume $h(x, y)$ is continuous on any continuous point of $\theta$, that is, it is continuous on every point at which the probability measure is zero. Then, IPR-$\tau^*(X, Y) \geq 0$ with equality if and only if X and Y are independent.*

*Remark* 1. The abstract Wiener measure is a standardized multivariate Gaussian measure extendable to separable Banach space. It is defined on a Borel $\sigma$-algebra generated by open subsets and has a positive measure for any open subset. Due to the absence of the Lebesgue measure in infinite dimensional spaces, we use the Abstract Wiener measure to define the notion of continuous distributions, and it is easy to see that it is equivalent to the conventional definition when restricted to a finite dimensional space.

*Remark* 2. If we consider $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}^q$, with $l^d$ metric, for $p, q, d \geq 3$, the spaces are not of strong negative type (Lyons et al., 2013), and the distance covariance test is not consistent. The proposed method can cover these cases.

*Remark* 3. We consider a separable Banach space, instead of a more general metric space, as the Borel $\sigma$-algebra on the product space is the $\sigma$-algebra generated by the product of the open subsets. Also the Abstract Wiener measure is well-defined on a separable Banach space with desired properties to define the continuous distribution. The results may be pushed to a separable metric space with a more mathematically sophisticated definition of continuous distributions in the absence of the Lebesgue measure. The assumptions regarding continuous and discrete distributions are inherited from the original sign covariance paper, which may be relaxed as the authors of the sign covariance paper conjectured. Our numerical experiments show that the method works under various settings within and beyond these requirements. Moreover, well behaved metric spaces are isometric to subspaces of Banach Space; See for example Kuratowski embedding for bounded metric spaces (Kuratowski, 1935), the generalized Banach-Mazur theorem for separable metric spaces (Kleiber & Pervin, 1969), and Nash embedding (Nash, 1956) for Riemannian manifolds. Therefore the independence-zero equivalence property in most applications can be studied in Banach spaces, with a restricted measure support.

## 3. Test of Independence

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent and identically distributed sample realizations of $(X, Y)$ from a joint distribution $\theta$.

Definition 2. *We propose an empirical IPR-$\tau^*$ in the form of a U-Statistic of order 5,*

$$H_n(X,Y) = \frac{1}{\binom{n}{5}} \sum_{1 \le i_1 < \ldots < i_5 \le n} \phi\{(x_{i_1}, y_{i_1}), \ldots, (x_{i_5}, y_{i_5})\} \tag{2}$$

*with the kernel $\phi\{(x_1, y_1), \ldots, (x_5, y_5)\}$ defined as*

$$\frac{1}{5!} \sum_{(j_1,\ldots,j_5) \in \mathcal{P}_5} a\{\rho(x_{j_1},x_{j_2}), \rho(x_{j_1}, x_{j_3}), \rho(x_{j_1}, x_{j_4}), \rho(x_{j_1}, x_{j_5})\}$$
$$a\{\zeta(y_{j_1}, y_{j_2}), \zeta(y_{j_1}, y_{j_3}), \zeta(y_{j_1}, y_{j_4}), \zeta(y_{j_1}, y_{j_5})\}.$$

*Remark* 4. There exists a more straightforward expression of Eq. (2). Let

$$A_i = \frac{1}{\binom{n-1}{4}} \sum_j \frac{1}{4!} \sum_{\pi \in \mathcal{P}_4} a\{\rho(x_i, x_{j_{\pi(1)}}), \ldots, \rho(x_i, x_{j_{\pi(4)}})\} a\{\zeta(y_i, y_{j_{\pi(1)}}), \ldots, \zeta(y_i, y_{j_{\pi(4)}})\},$$

where the outer sum is taken over the set of all ordered subsets $j$ of 4 different integers chosen from $\{1, 2, \ldots, n\}/\{i\}$. Here $A_i$ is a U-statistic of order 4 and gives rise to the $\tau^*$ between the interpoint distance-induced random variables, $\rho(x_i, X)$ and $\zeta(y_i, Y)$. Then $\sum_i A_i/n$ equals the empirical IPR-$\tau^*(X, Y)$.

Let

$$\phi_i\{(x_1, y_1), \ldots, (x_i, y_i)\} = E[\phi\{(x_1, y_1), \ldots, (x_i, y_i), (X^{i+1}, Y^{i+1}), \ldots, (X^5, Y^5)\}]$$

and $\sigma_i^2 = \text{var}(\phi_i)$ for $i = 1, \ldots, 5$.

We first present the general results based on the large-sample theory of the $U$-statistics (Section 5.5 of Serfling (2009)).

Lemma 1. *If $\sigma_5^2 < \infty$, we have*

$$n^{1/2}\{H_n(X, Y) - \textit{IPR-}\tau^*(X, Y)\} \to N(0, 5^2\sigma_1^2)$$

*in distribution.*

In the case that $\sigma_1^2 = 0$, the above Gaussian limit is degenerate, and we refer to a second lemma.

Lemma 2. *If $\sigma_5^2 < \infty$ and $0 = \sigma_1^2 < \sigma_2^2$, we have*

$$n\{H_n(X, Y) - \textit{IPR-}\tau^*(X, Y)\} \to \binom{5}{2} \sum_{m=1}^{\infty} \lambda_m (\chi_{1m}^2 - 1) \tag{3}$$

*in distribution, where $\chi_{1m}^2$ are independent $\chi_1^2$ variables and $\lambda_m$ are the solutions of the eigen equation*

$$\int_{\mathcal{X} \times \mathcal{Y}} \phi_2\{(x, y), (x', y')\}\psi(x', y')d\theta = \lambda\psi(x, y), \psi \in l_2(\theta). \tag{4}$$

The integral operator in equation (4) is bounded and compact, and we have the expansion $\phi_2\{(x, y), (x', y')\} = \sum_{m=1}^{\infty} \lambda_m \psi_m(x, y)\psi_m(x', y')$, with mean-square convergence.

Under the null hypothesis, interpoint distances arising from fixed $(x_1, y_1)$ are still independent and identically distributed random variables with no association. Therefore, $\phi_1\{(x_1, y_1)\}$ equals zero (Nandy et al., 2016) and so does $\sigma_1^2$. Then $H_n$ is a degenerate U-statistic, and Lemma 2 applies. Under the alternative hypothesis, $H_n(X, Y)$ converges to IPR-$\tau^*(X, Y)$ and IPR-$\tau^* > 0$ by Theorem 1, so $nH_n(X, Y) \to \infty$ as $n \to \infty$.

We propose to reject the null hypothesis when $nH_n(X, Y) > C_\alpha$, where $C_\alpha$ is the $\alpha$-level critical value from the null distribution. Combining Theorem 1 and Lemmas 1 - 2, we can obtain the following theorem.

THEOREM 2. *If $X$ and $Y$ are jointly distributed as specified in Theorem 1, $nH_n(X, Y)$ can serve as a test statistic for a test of independence which is consistent against the alternatives. Specifically,*
*(a) If $X$ and $Y$ are independent,*

$$nH_n(X, Y) \to \binom{5}{2} \sum_{m=1}^{\infty} \lambda_m (\chi_{1m}^2 - 1) \tag{5}$$

*in distribution, where $\chi_{1m}^2$ are independent $\chi_1^2$ variables and $\lambda_m$ are the solutions of the eigen equation*

$$\int_{\mathcal{X}\times\mathcal{Y}} \phi_2((x, y), (x', y'))\psi(x', y')d\theta = \lambda\psi(x, y), \psi \in L_2. \tag{6}$$

*(b) If $X$ and $Y$ are dependent, $nH_n(X, Y) \to \infty$ in probability.*

For $n$ sample points, efficient algorithms to compute the sign covariance has been developed in Heller & Heller (2016b) with $\mathcal{O}(n^2)$ operations and later in Even-Zohar & Leng (2021) with $\mathcal{O}(nlogn)$ operations. Our statistics is a summation of $n$ different sign covariances of interpoint distances. In practice, since the critical value of null distribution does not have an easy expression, a permutation procedure is needed to approximate the critical value. Several recent results showed that the permutation critical value converge to the critical value from the null distribution for a class of U-statistic based independence tests; see Theorem A.1 of Kim et al. (2020), Theorem 1 of Xu & Zhu (2020), and Proposition 18 of Berrett et al. (2020). Alternatively, we can approximate the null distribution by empirically solving the eigen equation as in equation (6). However, our numerical experiments showed that the permutation method with 1000 permutation samples is generally faster than the null distribution approximation.

Finally, as we show now, the ideas developed in this paper can be used to generalize other univariate dependence measures, such as Hoeffding's $D$, Blum-Kiefer-Rosenblatt's $R$, the squared Kendall's $\tau$, and the univariate distance covariance. These coefficients can be estimated by a $U$-statistic; see for example Drton et al. (2020). Let

$$U_m = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \ldots < i_m \le n} \phi\{(x_{i_1}, y_{i_1}), \ldots, (x_{i_m}, y_{i_m})\},$$

where the kernel $\phi\{(x_1, y_1), \ldots, (x_m, y_m)\}$ is defined as

$$\frac{1}{(m)!} \sum_{(j_1, \ldots, j_m) \in \mathcal{P}_m} h\{(x_{j_1}, y_{j_1}), \ldots, (x_{j_m}, y_{j_m})\}.$$

Then we can generalize $U_m$ to work for multivariate data or more general objects by introducing an extra independent pair,

$$U_{m+1} = \frac{1}{\binom{n}{m+1}} \sum_{1 \le i_1 < \ldots < i_{m+1} \le n} \tilde{\phi}\{(x_{i_1}, y_{i_1}), \ldots, (x_{i_{m+1}}, y_{i_{m+1}})\},$$

where the new kernel $\tilde{\phi}\{(x_1, y_1), \ldots, (x_{m+1}, y_{m+1})\}$ is defined as

$$\frac{1}{(m+1)!} \sum_{(j_1, \ldots, j_{m+1}) \in \mathcal{P}_{m+1}} h[\{\rho(x_{j_1}, x_{j_2}), \zeta(y_{j_1}, y_{j_2})\}, \ldots, \{\rho(x_{j_1}, x_{j_{m+1}}), \zeta(y_{j_1}, y_{j_{m+1}})\}].$$

When this generalization is applied to Hoeffding's $D$ or Blum-Kiefer-Rosenblatt's $R$, the empirical performance is very similar to that of IPR-$\tau^*$. In this paper, we focus on generalizing the sign covariance test because it has nice theoretical properties and also the construction of $\tau^*$ is simpler than the construction of Hoeffding's $D$ and Blum-Kiefer-Rosenblatt's $R$.

In addition, for $X \in \mathbb{R}$ and $Y \in \mathbb{R}$, the distance covariance (Székely et al., 2007) has the equivalent formula

$$dCov = 1/4E\{b(X^1, X^2, X^3, X^4)b(Y^1, Y^2, Y^3, Y^4)\},$$

with $b(z_1, z_2, z_3, z_4) = |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|$. Here the $b$ function differs from the $a$ function used in $\tau^*$ only by a sign operator (Bergsma & Dassios, 2014). Székely et al. (2007) showed that the distance covariance test works for multivariate data if the absolute distance $|z_1 - z_2|$ is replaced by a Euclidean distance on $\mathbb{R}^p$. If we estimate the univariate distance covariance by $U_m$ as analogous to that for the sign covariance $\tau^*$, the form of $U_{m+1}$ naturally provides another way to generalize the univariate distance covariance to the multivariate case.

## 4. Simulations

We study the empirical performance of our proposed interpoint-ranking sign covariance test through three simulations: Simulation I, Multivariate data; Simulation II, Manifold-valued functional data, and Simulation III, Manifold-valued data, which is in the supplementary material. We compare the Type-I error and statistical power with several existing tests of independence: the distance covariance test denoted by "dCov" using the R package *energy* (Székely et al., 2007), the test based on the summation of Pearson chi-square statistic denoted by "HHG" using the R package *HHG* (Heller et al., 2012), the Ball covariance test with a constant weight denoted by"BCov1" and a probability weight denoted by "BCov2" using the R package *Ball* (Pan et al., 2019) and the Hilbert-Schmidt independence criterion with Gaussian kernel denoted by "HSIC" using the R package *HSIC* (Gretton et al., 2008). Our proposed test is implemented in R. The code is available on the author's website. The distance covariance test is consistent for variables in metric spaces of strong negative type and the HSIC method requires the kernels to be positive definite and characteristic. These conditions are in general hard to check for non-Hilbertian data, and we know some of the non-Hilbertian data are not of strong negative type. Also the original "HHG" paper (Heller et al., 2012) only proved consistency of the test for $\mathbb{R}^p$ and $\mathbb{R}^q$. Nevertheless, we still applied these methods to all of the examples since these methods are all based on pairwise distances and can be empirically applied to complex objects as long as an appropriate distance or metric can be defined. In the implementation, we use the same metric for all methods, and p-values are all based on 1000 permutations. In all of the following settings, we report the results with sample sizes 20, 50, 100 and 200. The significance level is 0.05, and powers are based on 1000 simulations.

Table 1. *Empirical Type-I error rates at nominal significance level 0.05 using n=100 in Simulation I. Results are based on 1,000 simulations.*

| Test | Normal | t(1) | t(2) |
|---|---|---|---|
| IPR-$\tau^*$ | 0.035 | 0.048 | 0.045 |
| dCov | 0.054 | 0.038 | 0.060 |
| HHG | 0.051 | 0.046 | 0.054 |
| BCov1 | 0.047 | 0.045 | 0.056 |
| BCov2 | 0.054 | 0.046 | 0.056 |
| HSIC | 0.053 | 0.040 | 0.057 |

In Simulation I, we consider multivariate variables $X = (X_1, X_2, X_3, X_4, X_5)$ and $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ with the regular Euclidean distance. Examples 1-2 assess Type-I error rates and Examples 3-11 compare power performances. Similar settings have been used in Székely et al. (2007), Heller et al. (2012) and Pan et al. (2019).

*Example* 1. (Type I) $X$, $Y$ are generated from a multivariate normal distribution with mean $\mathbf{0}$ and $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(Y_i, Y_j) = 0.1$ for $i \neq j$, $i, j = 1, 2, 3, 4, 5$. There is no correlation between $X$ and $Y$ components.

*Example* 2. (Type I) $X$, $Y$ are generated from a multivariate $t(v)$ distribution for $v = 1, 2$.

*Example* 3. (Linear) $X$,$Y$ are generated from a jointly normal distribution with mean $\mathbf{0}$ and $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(Y_i, Y_j) = 0.1$ for $i \neq j$, $\mathrm{cov}(X_i, Y_i) = 0.3$ for $i, j = 1, 2, 3, 4, 5$.

In Examples 4-8, $X$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ and $\mathrm{cov}(X_i, X_j) = 0.1$ for $i \neq j$, $i, j = 1, 2, 3, 4, 5$.

*Example* 4. (Quadratic) $Y_i = 0.5X_i^2 + \epsilon$ with $\epsilon \sim N(0, 1)$ for $i = 1, 2, 3, 4, 5$.

*Example* 5. (Y=$X\epsilon$) $Y_i = X_i\epsilon$ with $\epsilon \sim N(0, 1)$ for $i = 1, 2, 3, 4, 5$.

*Example* 6. (Y=Inv(X)) $Y_i = 1/|X_i|$ for $i = 1, 2, 3, 4, 5$.

*Example* 7. (Concave) $Y_i = \pm 1/|X_i|$ with random signs of equal probability, for $i = 1, 2, 3, 4, 5$.

*Example* 8. (X-shape) $Y_i = \pm X_i$ with random signs of equal probability, for $i = 1, 2, 3, 4, 5$.

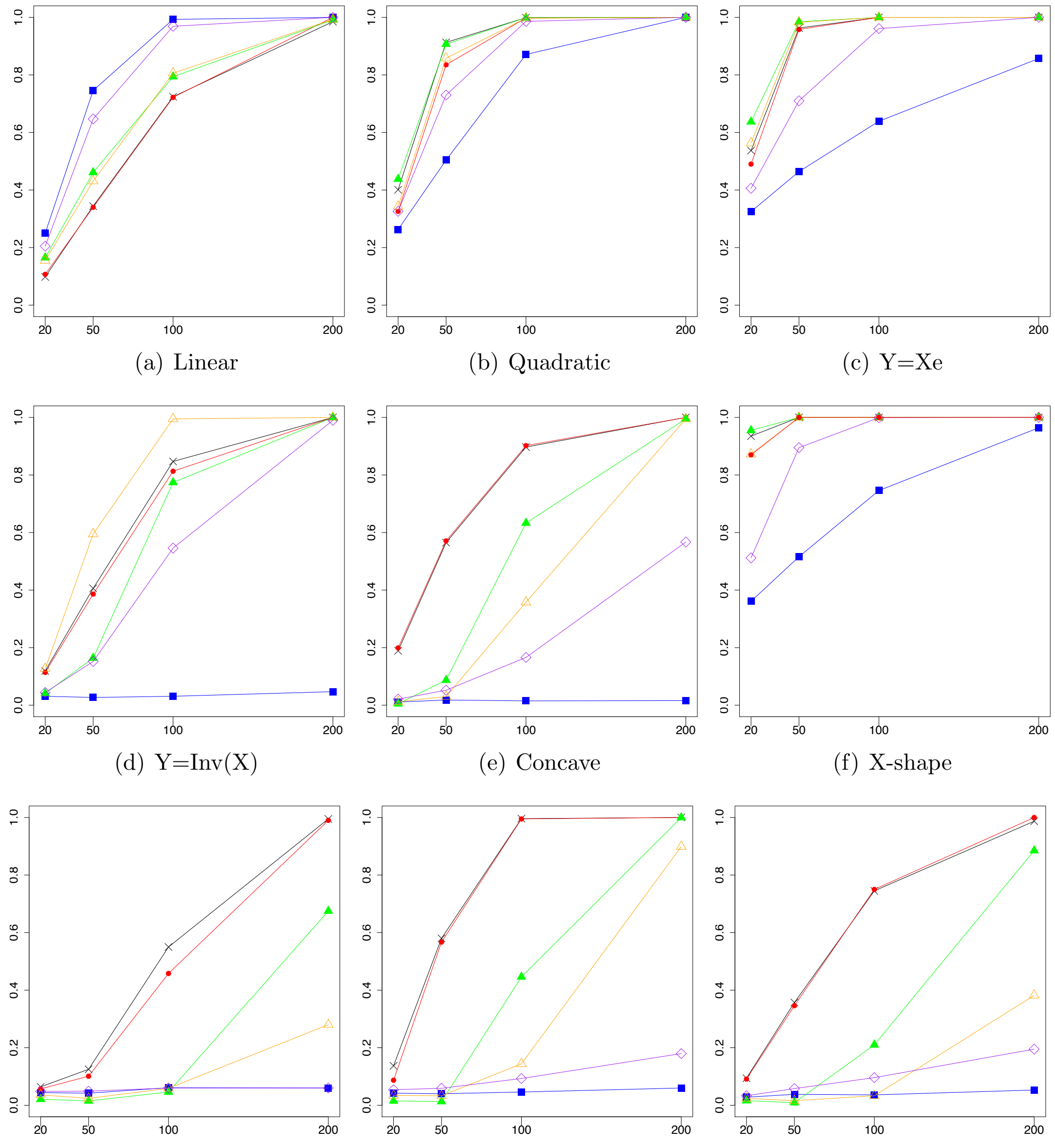
In Examples 9-11, $Z = (Z_1, Z_2, Z_3, Z_4, Z_5)$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ and $\mathrm{cov}(Z_i, Z_j) = 0.1$ for $i \neq j$, $i, j = 1, 2, 3, 4, 5$.

*Example* 9. (Circle) $X_i = 2logit^{-1}(Z_i) - 1$ and $Y_i = \pm(1 - X_i^2)^{1/2}$ with random signs of equal probability, for $i = 1, 2, 3, 4, 5$.

*Example* 10. (Diamond) $X_i = 2logit^{-1}(Z_i) - 1$ and $Y_i = \pm(1 - |X_i|)$ with random signs of equal probability, for $i = 1, 2, 3, 4, 5$.

*Example* 11. (Two-pieces) $X_i = 2logit^{-1}(Z_i) - 1$ and $Y_i = (0.9I_{\{|X_i|<0.5\}} + 0.1)\epsilon$ with $\epsilon \sim N(0, 0.1)$ for $i = 1, 2, 3, 4, 5$.

Type-I error rates are well-controlled for all methods for all sample sizes. See Table 1 for the result with $n$=100. Figure 1 summarizes the empirical powers. The proposed method "IPR-$\tau^*$" shows a good performance, and the powers reach one as sample size increases to 200. In



(a) Linear

(b) Quadratic

(c) Y=Xe

(d) Y=Inv(X)

(e) Concave

(f) X-shape

(g) Circle

(h) Diamond

(i) Two-pieces

Fig. 1. Simulation I: Empirical power of the tests for IPR-$\tau^*$ (●, red), dCov (■, blue), HHG (×, black), BCov1 (▲, green), BCov2 (△, orange) and HISC (◇, purple). Power values are computed for each of the sample sizes 20, 50, 100, 200 with 1,000 simulations.

general, we see that "IPR-$\tau^*$" and "HHG" always belong to a group with the highest power except for the linear case. While the distance covariance test has the best power for the linear case, it generally has the lowest power for other non-linear cases. The power of "dCov" is poor for "$Y = Inv(X)$", "Concave", "Circle", "Diamond" and "Two-pieces" even with the sample size 200. The performances of "BCov1", "BCov2" and "HSIC" are somewhat in between. The power of "HSIC" with Gaussian kernel is seen to be poor for "Circle", "Diamond" and "Two-pieces".

In Simulation II, we consider manifold-valued functional trajectories $X(t) = (\theta(t), \phi(t))_s$ with $t \in [0, 1]$. For each time point $t$, $X(t)$ is a point on the unit sphere $S^2$, so $\theta(t) \in [-\pi, \pi]$ and $\phi(t) \in [-\pi/2, \pi/2]$. We generate data as follows.

$$\theta(t) =\{\eta_0 t + \sum_{j=1}^{20} \eta_j \text{sin}(j\pi t)\}(\text{mod } 2\pi) - \pi,$$

$$\phi(t) =[\{\xi_0 t + \sum_{j=1}^{20} \xi_j \text{sin}(j\pi t)\}/2 \vee \pi/2] \wedge -\pi/2$$

with coefficients $\eta$ and $\xi$ drawn independently from a normal distribution with mean zero. Standard deviations are 1 for $\eta_0, \xi_0$ and $j^{-6/5}$ for $\eta_j, \xi_j$ $(j = 1, \ldots, 20)$. When defining $\phi(t)$, we made upper and lower bounds to ensure $\phi(t) \in [-\pi/2, \pi/2]$, even though $\{\xi_0 t + \sum_{j=1}^{20} \xi_j \text{sin}(j\pi t)\}/2$ exceeding $\pi/2$ or below $-\pi/2$ is unlikely in this setting. Examples of sample trajectories $X(t)$ are shown in Figure 2.

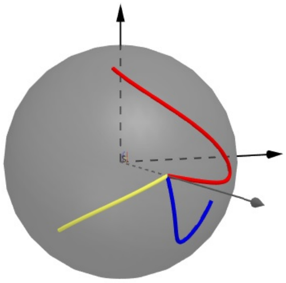


Fig. 2. A demonstration of sample curves $X_i(t)$ generated in Simulation II.

The distance between two trajectories is measured by $d_1(X_i(t), X_{i'}(t)) = sup_{t\in[0,1]} d(X_i(t), X_{i'}(t))$ where $d$ refers a great-circle distance between two points on a unit sphere. In our simulation, we generate $n = 20, 50, 100, 200$ sample curves with 101 observed points on each trajectory where the observed points are equally spaced between $[0, 1]$.

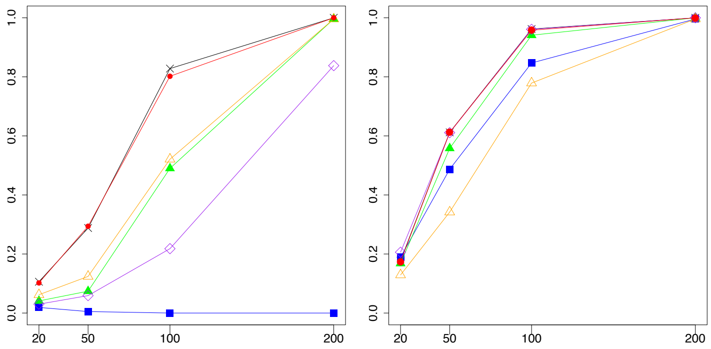
*Example* 12. (Type I) $X(t) = (\theta(t), \phi(t))_s, Y \sim N(0, 1)$.

*Example* 13. (FT1) $X(t) = (\theta(t), \phi(t))_s, Y = (1/\int_0^1 |\theta(t)|dt, 1/\int_0^1 |\phi(t)|dt)$.

*Example* 14. (FT2) $X(t) = (\theta(t), \phi(t))_s, Y = (\int_0^1 |\theta(t)|dt + \epsilon_1, \int_0^1 |\phi(t)|dt + \epsilon_2)$ with $\epsilon_1, \epsilon_2 \sim N(0, 0.5)$.

In Simulation II, the Type-I errors are well-controlled for all the methods. Figure 3 show that the powers of all methods increase as the sample size increases, except for the "dCov" method in Example 13. The proposed test "IPR-$\tau^*$" and "HHG" maintain good power in both cases.

Overall, numerical experiments have confirmed that the proposed test is consistent against general alternatives. The empirical powers for "IPR-$\tau^*$" and "HHG" are similar and within the highest power group, although the test statistics are derived from very different perspectives. The ball covariance methods are powerful in most cases, and we find that different weights lead to different performances with no obvious winner. The ball covariance is proved to be asymptotically equivalent to the "HHG" test if a chi-square type weight is used (Pan et al., 2019), and we only focus on finite sample performance here. The HSIC method with Gaussian kernel is



(a) FT1 (b) FT2

Fig. 3. Simulation II: Empirical power of the tests for IPR-$\tau^*$ (●,red), dCov (■, blue), HHG (×, black), BCov1 (▲, green), BCov2 (△, orange) and HISC (◇, purple). Power values are computed for each of the sample sizes 20, 50, 100, 200 with 1,000 simulations.

usually in the middle group, but has poor power for some cases such as Examples 9 and 10. The performance of "dCov" is somehow divided. The method tends to perform better than others in normal and linear cases, but clearly less competitive in terms of power in all other cases.

## 5. Data Examples

### 5.1. *DLBCL Data*

We first apply the proposed test of independence to study the relationship between gene expression and survival outcomes. We use the data provided by Rosenwald et al. (2002), in which the survival time of patients with diffuse large B-cell lymphoma after chemotherapy is recorded as well as the related gene expression profiles. As the survival time is believed to be influenced by the molecular features of tumor, previous papers, including Bair & Tibshirani (2004), Bair et al. (2006), Bøvelstad et al. (2007), Chen et al. (2011) and Chen et al. (2017), have tried to build predictive models for survival time of a patient, using gene-expression patterns as predictors.

In the study, 240 patients were examined for 7399 gene expression profiles with the use of DNA microarrays. Following the same approach as the above-cited authors, we pre-screen the genes and use only 240 most relevant ones. The subset selection is performed by fitting a univariate Cox regression model of each gene expression value on survival one-by-one and ranking the obtained Cox scores from largest to smallest (Chen et al., 2011). We apply the proposed method "IPR-$\tau^*$" as well as "dCov", "HHG", "BCov1", "BCov2", "HSIC" to test independence between 240 gene expressions and the survival time. Euclidean distance is used to measure the distance between two gene expressions for all the methods.

All of the methods detect the dependency with 0.05 significance level. More efforts can be put into building predictive models after the nonparametric tests show statistically significant results. As a comparison, we test with the 240 genes that have the lowest Cox scores and repeat the same sets of tests. No method concludes the dependency.

### 5.2. *Farm Data*

We investigate the dependency between the annual crop yields and the temperature using the dataset described in Wong et al. (2019). In this dataset, the annual yield of two major crops, corn and soybean, is recorded in bushels per acre from 105 counties of Kansas from 1999 and 2011, provided by the National Agricultural Statistics Agency at https://quickstats.nass.usda.gov/. The weather data is from the National Climatic Data Center at https://www.ncdc.noaa.gov/data-access and contains the daily minimum, maximum temperature aggregated at the county level.

Following the source paper, we let $Y$ be the annual corn or soybean yield for a specific year and county, respectively, $X_1(t)$ and $X_2(t)$ be the daily maximum and minimum temperatures for the same year and county, and $X(t) = (X_1(t), X_2(t))$. We aim at testing independence between the annual crop yields and temperature trajectories. For this purpose, we introduce the daily heat unit accumulation defined as, $HU(X, t) = \{(X_1(t) + X_2(t))/2 - T_{base}\}_+$, based on which phenological development of plants occurs according to the EPIC plant growth model (Williams et al., 1989), i.e., no growth occurs at or below $T_{base}$. In the formula, $T_{base}$ is a crop-specific base temperature. Following the reference paper, we use $T_{base} = 8°C$ for corn and $10°C$ for soybean. $c_+$ denotes the positive part of $c$. The dissimilarity between temperatures is measured by

$$d_X(X(t), X'(t)) = \left[\int_0^{365} \{(\frac{X_1(t) + X_2(t)}{2} - T_{base})_+ - (\frac{X_1'(t) + X_2'(t)}{2} - T_{base})_+\}^2 dt\right]^{1/2},$$

and Euclidean distance is used for $Y$.

All six methods have the same statistical conclusion at the 0.05 significance level. Significant dependence between temperature and annual crop yields are found for both soybean data and the corn data.

Bergsma & Dassios (2014) derived that $\tau^* = (2\Pi_{C_4} - \Pi_{D_4})/3$, where $\Pi_{C_4}$ and $\Pi_{D_4}$ denote the probabilities that four randomly chosen pairs are concordant and discordant, respectively. Under independence, $\Pi_{C_4} = 1/3$ and $\Pi_{D_4} = 2/3$. If one variable is a strictly monotone function of the other, then $\Pi_{C_4} = 1$ and $\Pi_{D_4} = 0$. In addition Drton et al. (2020) proved that for bivariate normal data, the sign covariance is a monotone function of the correlation $|\rho|$. The interpoint-ranking sign covariance is an average of the sign covariances between $X$- and $Y$-interpoint distances arising from all anchor points $(X^0, Y^0)$ in $\mathcal{X} \times \mathcal{Y}$. In this data example, we find that the largest sign covariance values arise from anchor points in year 2004, which indicates that the $X$ and $Y$-interpoint distances computed for a data point in year 2004 tend to be highly correlated. Further interpretation and visualization of the dependence could be a topic for future research.

## SUPPLEMENTARY MATERIAL

Supplementary material includes the proof of Theorem 1 and the result of simulation III.

## REFERENCES

ADRIAENSSENS, N., COENEN, S., VERSPORTEN, A., MULLER, A., MINALU, G., FAES, C., VANKERCKHOVEN, V., AERTS, M., HENS, N., MOLENBERGHS, G. et al. (2011). European surveillance of antimicrobial consumption (esac): outpatient quinolone use in europe (1997–2009). *Journal of antimicrobial chemotherapy* **66**, vi47–vi56.

BAIR, E., HASTIE, T., PAUL, D. & TIBSHIRANI, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

BAIR, E. & TIBSHIRANI, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**, e108.

BERGSMA, W. & DASSIOS, A. (2014). A consistent test of independence based on a sign covariance related to kendall's tau. *Bernoulli* **20**, 1006–1028.

BERRETT, T. B., KONTOYIANNIS, I. & SAMWORTH, R. J. (2020). Optimal rates for independence testing via $u$-statistic permutation tests. *arXiv preprint arXiv:2001.05513* .

BERRETT, T. B. & SAMWORTH, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika* **106**, 547–566.

BISWAS, M., SARKAR, S. & GHOSH, A. K. (2016). On some exact distribution-free tests of independence between two random vectors of arbitrary dimensions. *Journal of Statistical Planning and Inference* **175**, 78–86.

Blum, J. R., Kiefer, J. & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics* , 485–498.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A. & Lingjærde, O. C. (2007). Predicting survival from microarray data—a comparative study. *Bioinformatics* **23**, 2080–2087.

Chen, K., Chen, K., Müller, H.-G. & Wang, J.-L. (2011). Stringing high-dimensional data for functional analysis. *Journal of the American Statistical Association* **106**, 275–284.

Chen, K., Zhang, X., Petersen, A. & Müller, H.-G. (2017). Quantifying infinite-dimensional data: Functional data analysis in action. *Statistics in Biosciences* **9**, 582–604.

Dai, X., Müller, H.-G. et al. (2018). Principal component analysis for functional data on riemannian manifolds and spheres. *The Annals of Statistics* **46**, 3334–3361.

Deb, N. & Sen, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *arXiv preprint arXiv:1909.08733* .

del Barrio, E., Cuesta-Albertos, J. A., Hallin, M. & Matrán, C. (2018). Center-outward distribution functions, quantiles, ranks, and signs in Rd. *arXiv preprint arXiv:1806.01238* .

Dhar, S. S., Dassios, A., Bergsma, W. et al. (2016). A study of the power and robustness of a new test for independence against contiguous alternatives. *Electronic Journal of Statistics* **10**, 330–351.

Drton, M., Han, F. & Shi, H. (2020). High dimensional independence testing with maxima of rank correlations. *The Annals of Statistics* **to appear**.

Even-Zohar, C. & Leng, C. (2021). Counting small permutation patterns. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM.

Free, S., O'Higgins, P., Maudgil, D., Dryden, I., Lemieux, L., Fish, D. & Shorvon, S. (2001). Landmark-based morphometrics of the normal adult brain using mri. *Neuroimage* **13**, 801–813.

Friedman, J. H., Rafsky, L. C. et al. (1983). Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics* **11**, 377–391.

Gretton, A., Bousquet, O., Smola, A. & Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*. Springer.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B. & Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in neural information processing systems*.

Guo, L. & Modarres, R. (2020). Nonparametric tests of independence based on interpoint distances. *Journal of Nonparametric Statistics* **32**, 225–245.

Heller, R. & Heller, Y. (2016a). Multivariate tests of association based on univariate tests. *Advances in Neural Information Processing Systems* **29**, 208–216.

Heller, R., Heller, Y. & Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100**, 503–510.

Heller, Y. & Heller, R. (2016b). Computing the bergsma dassios sign-covariance. *arXiv preprint arXiv:1605.08732* .

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics* , 546–557.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30**, 81–93.

Kim, I., Balakrishnan, S. & Wasserman, L. (2020). Robust multivariate nonparametric tests via projection-averaging. *Annals of Statistics* , 1–34.

Kleiber, M. & Pervin, W. (1969). A generalized banach-mazur theorem. *Bulletin of The Australian Mathematical Society* **1**, 169–173.

Kuratowski, C. (1935). Quelques problèmes concernant les espaces métriques non-séparables. *Fundamenta Mathematicae* **25**, 534–545.

Leung, D., Drton, M. et al. (2018). Testing independence in high dimensions with sums of rank correlations. *The Annals of Statistics* **46**, 280–307.

Lyons, R. et al. (2013). Distance covariance in metric spaces. *The Annals of Probability* **41**, 3284–3305.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research* **27**, 209–220.

Masucci, A. P., Smith, D., Crooks, A. & Batty, M. (2009). Random planar graphs and the london street network. *The European Physical Journal B* **71**, 259–271.

Nandy, P., Weihs, L., Drton, M. et al. (2016). Large-sample theory for the bergsma-dassios sign covariance. *Electronic Journal of Statistics* **10**, 2287–2311.

Nash, J. (1956). The imbedding problem for riemannian manifolds. *Annals of Mathematics* , 20–63.

Pan, W., Wang, X., Zhang, H., Zhu, H. & Zhu, J. (2019). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association* , 1–24.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58**, 240–242.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine* **346**, 1937–1947.

SARKAR, S. & GHOSH, A. K. (2018). Some multivariate tests of independence based on ranks of nearest neighbors. *Technometrics* **60**, 101–111.

SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A., FUKUMIZU, K. et al. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics* **41**, 2263–2291.

SERFLING, R. J. (2009). *Approximation theorems of mathematical statistics*, vol. 162. John Wiley & Sons.

SHEN, C., PRIEBE, C. E. & VOGELSTEIN, J. T. (2019). From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association* , 1–22.

SHI, H., DRTON, M. & HAN, F. (2020). Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association* , 1–16.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *The American journal of psychology* **15**, 72–101.

SZÉKELY, G. J. & RIZZO, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis* **117**, 193–213.

SZÉKELY, G. J., RIZZO, M. L., BAKIROV, N. K. et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

WEIHS, L., DRTON, M. & MEINSHAUSEN, N. (2018). Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika* **105**, 547–562.

WILLIAMS, J., JONES, C., KINIRY, J. & SPANEL, D. A. (1989). The epic crop growth model. *Transactions of the ASAE* **32**, 497–0511.

WONG, R. K., LI, Y. & ZHU, Z. (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association* **114**, 406–418.

XU, K. & ZHU, L. (2020). Power analysis of projection-pursuit independence tests. *Statistica Sinica* **to appear**.

ZHENG, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)* **6**, 1–41.

ZHU, C., ZHANG, X., YAO, S., SHAO, X. et al. (2020). Distance-based and rkhs-based dependence metrics in high dimension. *Annals of Statistics* **48**, 3366–3394.

ZHU, L., XU, K., LI, R. & ZHONG, W. (2017). Projection correlation between two random vectors. *Biometrika* **104**, 829–843.

[*Received on* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]