Bernoulli **20**(2), 2014, 1006–1028
DOI: 10.3150/13-BEJ514

A consistent test of independence based on a sign covariance related to Kendall's tau

WICHER BERGSMA* and ANGELOS DASSIOS**

London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom.
*E-mail: *w.p.bergsma@lse.ac.uk; **a.dassios@lse.ac.uk*

The most popular ways to test for independence of two ordinal random variables are by means of Kendall's tau and Spearman's rho. However, such tests are not consistent, only having power for alternatives with "monotonic" association. In this paper, we introduce a natural extension of Kendall's tau, called τ^*, which is non-negative and zero if and only if independence holds, thus leading to a consistent independence test. Furthermore, normalization gives a rank correlation which can be used as a measure of dependence, taking values between zero and one. A comparison with alternative measures of dependence for ordinal random variables is given, and it is shown that, in a well-defined sense, τ^* is the simplest, similarly to Kendall's tau being the simplest of ordinal measures of monotone association. Simulation studies show our test compares well with the alternatives in terms of average p-values.

Keywords: concordance; copula; discordance; measure of association; ordinal data; permutation test; sign test

1. Introduction

A random variable X is called *ordinal* if its possible values have an ordering, but no distance is assigned to pairs of outcomes. Ordinal variables may be continuous, categorical, or mixed continuous/categorical. Ordinal data frequently arise in many fields, though especially often in social and biomedical science (Kendall and Gibbons [13], Agresti [1]). Ordinal data methods are also often applied to real-valued (interval level) data in order to achieve robustness.

The two most popular measures of association for ordinal random variables X and Y are Kendall's tau (τ) (Kendall [14]) and Spearman's rho (ρ_S) (Spearman [25]), which may be defined as

$$\tau = E\operatorname{sign}\big[(X_1 - X_2)(Y_1 - Y_2)\big], \qquad \rho_S = 3E\operatorname{sign}\big[(X_1 - X_2)(Y_1 - Y_3)\big],$$

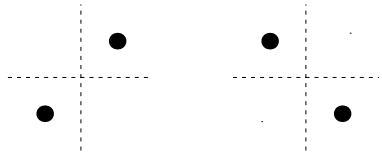
where the (X_i, Y_i) are independent replications of (X, Y) (Kruskal [16]). The factor 3 in the expression for ρ_S occurs to obtain a measure whose range is $[-1, 1]$. Both τ and ρ_S are proportional to sign versions of the ordinary covariance, which can be seen from the following expressions for the covariance:

$$\operatorname{cov}(X, Y) = \tfrac{1}{2}E(X_1 - X_2)(Y_1 - Y_2) = E(X_1 - X_2)(Y_1 - Y_3).$$

From the definitions, probabilistic interpretations of τ and ρ_S can be derived. Firstly,

$$\tau = \Pi_{C_2} - \Pi_{D_2}, \tag{1}$$

1350-7265



(a) Concordant pair (b) Discordant pair

Figure 1. Concordant and discordant pairs of points associated with Kendall's tau.

where Π_{C_2} is the probability that two observations are concordant and Π_{D_2} the probability that they are discordant (see Figure 1). Secondly,

$$\rho_S = \Pi_{C_3} - \Pi_{D_3},$$

where Π_{C_3} is the probability that three observations are concordant and Π_{D_3} the probability that they are discordant (see Figure 2). It can be seen that τ is simpler than ρ_S, in the sense that it can be defined using only two rather than three independent replications of (X, Y), or, more specifically, in terms of probabilities of concordance and discordance of two rather than three points. This was a reason for Kruskal to prefer τ to ρ_S (Kruskal [16], end of Section 14).

An alternative definition of ρ_S, which was originally given by Spearman, is as a Pearson correlation between uniform rank scores of the X and Y variables. For continuous random variables, both this and the aforementioned definition lead to the same quantity. However, with this definition, ρ_S is to some extent an *ad hoc* measure, since the choice of scores is arbitrary, and alternative scores (e.g., normal scores) might be used.

A test of independence based on i.i.d. data can be obtained by application of the permutation test to an estimator of τ or ρ_S, which is easy to implement and fast to carry out with modern computers. Such ordinal tests are also used as a robust alternative to tests based on the Pearson correlation.

A drawback for certain applications is that τ and ρ_S may be zero even if there is an association between X and Y, so tests based on them are inconsistent for the alternative of a general association. For this reason, alternative coefficients have been devised. The best known of these are

(a) Concordant triples (b) Discordant triples

Figure 2. Concordant and discordant triples of points associated with Spearman's rho.

those introduced by Hoeffding [11] and Blum, Kiefer and Rosenblatt [4]. With F_{12} the joint distribution function of (X, Y), and F_1 and F_2 the marginal distribution functions of X, respectively, Y, Hoeffding's coefficient is given as

$$H = \int \big[F_{12}(x, y) - F_1(x)F_2(y)\big]^2 \, \mathrm{d}F_{12}(x, y), \tag{2}$$

and the Blum–Kiefer–Rosenblatt (henceforth: BKR) coefficient as

$$D = \int \big[F_{12}(x, y) - F_1(x)F_2(y)\big]^2 \, \mathrm{d}F_1(x) \, \mathrm{d}F_2(y). \tag{3}$$

Both can be seen to be non-negative with equality to zero under independence. Furthermore, $D = 0$ can also be shown to imply independence. However, the Hoeffding coefficient has a severe drawback, namely that it may be zero even if there is an association, that is, it does not lead to a consistent independence test. An example is the case that $P(X = 0, Y = 1) = P(X = 1, Y = 0) = 1/2$ (Hoeffding [11], page 548).

A third option, especially suitable for categorical data, is the Pearson chi-square test; it is directly applicable to categorical data and can be used for continuous data after a suitable categorization. However, the chi-square test does not take the ordinal nature of the data into account, leading to potential power loss for "ordinal" alternatives; effectively the chi-square test treats the data as nominal rather than ordinal (see also Agresti [1]).

Although H and D have simple mathematical formulas, they seem to be rather arbitrary, and many variants are possible (see also Section 3.3). For this reason, we decided to develop a probabilistic interpretation of H (given in Section 3 of this paper). However, we then noticed that H and D were unnecessarily complex, and that a clearly simpler and natural alternative coefficient was possible. Our new coefficient is a direct modification of Kendall's τ, which we call τ^*. It is non-negative and zero if and only if independence holds. Like τ and ρ_S, we show that H and τ^* equal the difference of concordance and discordance probabilities of a number of independent replications of (X, Y). Analogously to the aforementioned way that τ is simpler than ρ_S, τ^* is simpler than H in that only four independent replications of (X, Y) are required, whereas H needs five. It appears to us that relative simplicity of interpretation of a coefficient is of utmost importance, and that this is also the main reason for the current popularity of Kendall's tau. In particular, when it was introduced in the pre-computer age in 1938, the sample value of Kendall's tau was much harder to compute than the sample value of Spearman's rho, which had been in use since 1904 (Kruskal [16]). In spite of this, judging by the number of Google Scholar hits, both currently appear to be about equally popular.[1]

As a remark on the two-sample case, if one of the variables is binary, a test that $\tau^* = 0$ is equivalent to the Cramér von Mises test, as shown in Section 3 in Dassios and Bergsma [3].

The organization of this paper is as follows. In Section 2, we first define τ^*, and then state our main theorem that $\tau^* \geq 0$ with equality if and only if independence holds. Furthermore, we provide a probabilistic interpretation in terms of concordance and discordance probabilities

[1]The Google Scholar search "`kendall's tau`" `OR` "`kendall tau`" gave us 16,400 hits and the search "`spearman's rho`" `OR` "`spearman rho`" 18,500.

of four points. Section 5 contains the proof of the main theorem. The proof turns out to be surprisingly involved for such a simple to formulate coefficient, and the ideas in the proof may be useful for other related research. A comparison with the Hoeffding, the BKR and some more recent coefficients is given in Section 3, and a new probabilistic interpretation for the former is given. In Section 4, we give a description of independence testing via the permutation test and a simulation study compares average p-values of our test and the aforementioned other two tests. Our test compares well with the other two in this respect.

2. Definition of τ^* and statement of its properties

We denote i.i.d. sample values by $(x_1, y_1), \ldots, (x_n, y_n)$, but will also use $\{(X_i, Y_i)\}$ to denote i.i.d. replications of (X, Y) in order to define population coefficients. The empirical value t of Kendall's tau is

$$t = \frac{1}{n^2} \sum_{i,j=1}^{n} \operatorname{sign}(x_i - x_j) \operatorname{sign}(y_i - y_j),$$

and its population version is

$$\tau = E \operatorname{sign}(X_1 - X_2) \operatorname{sign}(Y_1 - Y_2).$$

(Kruskal [16], Kendall and Gibbons [13]). With

$$\begin{aligned} s(z_1, z_2, z_3, z_4) &= \operatorname{sign}(z_1 - z_4)(z_3 - z_2) \\ &= \operatorname{sign}\bigl(|z_1 - z_2|^2 + |z_3 - z_4|^2 - |z_1 - z_3|^2 - |z_2 - z_4|^2\bigr), \end{aligned}$$

we obtain

$$t^2 = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} s(x_i, x_j, x_k, x_l) s(y_i, y_j, y_k, y_l)$$

and

$$\tau^2 = E s(X_1, X_2, X_3, X_4) s(Y_1, Y_2, Y_3, Y_4).$$

Replacing squared differences in s by absolute values of differences, we define

$$a(z_1, z_2, z_3, z_4) = \operatorname{sign}\bigl(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|\bigr). \tag{4}$$

This leads to a modified version of t^2,

$$t^* = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} a(x_i, x_j, x_k, x_l) a(y_i, y_j, y_k, y_l) \tag{5}$$

and the corresponding population coefficient

$$\tau^* = \tau^*(X, Y) = Ea(X_1, X_2, X_3, X_4)a(Y_1, Y_2, Y_3, Y_4).$$

The quantities t^* and τ^* are new, and the main result of the paper is the following:

Theorem 1. *Assume (X, Y) has a bivariate discrete or continuous distribution, or a mixture of the two, that is, assume there exists a probability mass function f and a density function $\tilde{f}$ such that*

$$P(X < x, Y < y) = \sum_{u_i < x, v_i < y} f(u_i, v_i) + \int_{u<x, v<y} \tilde{f}(u, v)\,\mathrm{d}u\,\mathrm{d}v.$$

It holds true that $\tau^*(X, Y) \geq 0$ with equality if and only if X and Y are independent.

The proof is given in Section 5. We conjecture that the condition of the theorem is not necessary, that is, that for arbitrary (X, Y) it holds that $\tau^*(X, Y) \geq 0$ with equality if and only if X and Y are independent.

If the sign functions are omitted from τ^*, we obtain the covariance introduced by Bergsma [2] and Székely, Rizzo and Bakirov [26]. They showed that for arbitrary real random variables X and Y, this covariance is non-negative with equality to zero if and only if X and Y are independent. (See Section 3.4 for further details.)

By the Cauchy–Schwarz inequality, the normalized value

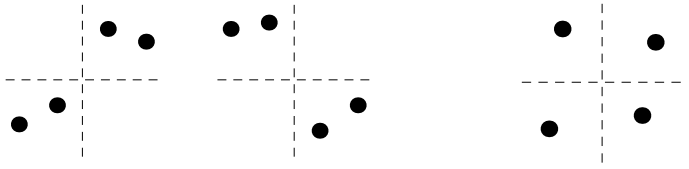
$$\tau_b^* = \frac{\tau^*(X, Y)}{\sqrt{\tau^*(X, X)\tau^*(Y, Y)}}$$

does not exceed one. (Note that this notation is in line with Kendall's τ_b, defined analogously.)

The definition of τ^* can easily be extended to X and Y in arbitrary metric spaces, but unfortunately Theorem 1 does not extend then, as it is possible that $\tau^* < 0$. This is shown by the following example. Consider a set of points $\{u_1, \ldots, u_8\} \subset \mathbb{R}^8$, where $u_i = (u_{i1}, \ldots, u_{i8})'$ such that $u_{ii} = 3$, $u_{ij} = -1$ if $i \neq j$ and $i, j \leq 4$ or $i, j \geq 5$, and $u_{ij} = 0$ otherwise. Suppose Y is uniformly distributed on $\{0, 1\}$, and given $Y = 0$, X is uniformly distributed on $u_1, \ldots, u_4$, and given $Y = 1$, X is uniformly distributed on $u_5, \ldots, u_8$. Then $\tau^* = -1/64$.

Note that $\tau^*(X, Y)$ is a function of the copula, which is the joint distribution of $F_1(X)$ and $F_2(Y)$, where F_1 and F_2 are the cumulative distribution functions of X and Y. More generally, for any strictly monotone (increasing or decreasing) functions g and h, $\tau^*(X, Y) = \tau^*(g(X), h(Y))$. Nelsen [18], Chapter 5, explores the way in which copulas can be used in the study of dependence between random variables, paying particular attention to Kendall's tau and Spearman's rho.

We now give a probabilistic interpretation of τ^*. Recall that Kendall's tau is the probability that a pair of points is concordant minus the probability that a pair of points is discordant. Our τ^* is proportional to the probability that two pairs are "jointly" concordant, plus the probability that two pairs are "jointly" discordant, minus the probability that, "jointly", one pair is discordant and the other concordant. Here, "jointly" refers to there being a common axis separating the two points of each of the two pairs.



(a) Concordant points (b) Discordant points

Figure 3. Configurations of concordant and discordant quadruples of points associated with τ^*. The dotted axes indicate strict separation of points in different quadrants; within a quadrant, no restrictions apply on the relative positions of points.

To use a slightly different terminology which will be convenient, we say that a set of four points is concordant if two pairs are either "jointly" concordant or "jointly" discordant, while four points are called discordant if, "jointly", one pair is concordant and the other is discordant. These configurations are given in Figure 3. In mathematical notation, a set of four points $\{(x_1, y_1), \ldots, (x_4, y_4)\}$ is concordant if there is a permutation (i, j, k, l) of $(1, 2, 3, 4)$ such that

$$(x_i, x_j < x_k, x_l)\&\big[(y_i, y_j < y_k, y_l)||(y_i, y_j > y_k, y_l)\big],$$

and discordant if there is a permutation (i, j, k, l) of $(1, 2, 3, 4)$ such that

$$\big[(x_i, x_j < x_k, x_l)||(x_i, x_j > x_k, x_l)\big]\&\big[(y_i, y_k < y_j, y_l)||(y_i, y_k > y_j, y_l)\big],$$

where $||$ and $\&$ are logical OR, respectively, AND, and $I(z_1, z_2 < z_3, z_4)$ is shorthand for $I(z_1 < z_3 \& z_1 < z_4 \& z_2 < z_3 \& z_2 < z_4)$. It is straightforward to verify that

$$\begin{aligned} a(z_1, z_2, z_3, z_4) = {} & I(z_1, z_3 < z_2, z_4) + I(z_1, z_3 > z_2, z_4) \\ & - I(z_1, z_2 < z_3, z_4) - I(z_3, z_4 < z_1, z_2), \end{aligned}$$

where I is the indicator function. Hence,

$$\begin{aligned} \tau^* = {} & 4P(X_1, X_2 < X_3, X_4 \& Y_1, Y_2 < Y_3, Y_4) \\ & + 4P(X_1, X_2 < X_3, X_4 \& Y_1, Y_2 > Y_3, Y_4) \\ & - 8P(X_1, X_2 < X_3, X_4 \& Y_1, Y_3 < Y_2, Y_4). \end{aligned} \tag{6}$$

Denoting the probability that four randomly chosen points are concordant as Π_{C_4} and the probability that they are discordant as Π_{D_4}, we obtain that the sum of the first two terms on the right-hand side of (6) equals $\Pi_{C_4}/6$, while the last term equals $\Pi_{D_4}/24$. Hence,

$$\tau^* = \frac{2\Pi_{C_4} - \Pi_{D_4}}{3}. \tag{7}$$

It can be seen that t^* and τ^* do not depend on the scale at which the variables are measured, but only on the ranks or grades of the observations. Four points are said to be *tied* if they are

neither concordant nor discordant. Clearly, for continuous distributions the probability of tied observations is zero. Hence, under independence, when all configurations are equally likely, $\Pi_{C_4} = 1/3$ and $\Pi_{D_4} = 2/3$, and if one variable is a strictly monotone function of the other, then $\Pi_{C_4} = 1$ and $\Pi_{D_4} = 0$.

3. Comparison to other tests

The two most popular (almost) consistent tests of independence for ordinal random variables are those based on Hoeffding's H and BKR's D, given in (2) and (3). We compare τ^* with these coefficients as well as with the recently introduced non-ordinal measures of Székely, Rizzo and Bakirov [26] and Gretton *et al.* [9]. We give a probabilistic interpretation for H and show that τ^* is simpler. Since $H = 0$ does not imply independence if the distributions are discrete, it should perhaps not be used, and we are left with two ordinal coefficients, τ^* and D, of which τ^* is the simplest. Further discussions of ordinal data and non-parametric methods for independence testing are given Agresti [1], Hollander and Wolfe [12] and Sheskin [24].

3.1. Probabilistic interpretation of Hoeffding's H

Hoeffding's [11] coefficient for measuring deviation from independence for a bivariate distribution function is given by (2) (see also Blum, Kiefer and Rosenblatt [4], Hollander and Wolfe [12] and Wilding and Mudholkar [27]). An alternative formulation given by Hoeffding is

$$H = \tfrac{1}{4} E\phi(X_1, X_2, X_3)\phi(X_1, X_4, X_5)\phi(Y_1, Y_2, Y_3)\phi(Y_1, Y_4, Y_5),$$

where $\phi(z_1, z_2, z_3) = I(z_1 \geq z_2) - I(z_1 \geq z_3)$. Hoeffding's H can be zero for some discrete dependent (X, Y). An example is the case that $P(X = 0, Y = 1) = P(X = 1, Y = 0) = 1/2$ (Hoeffding [11], page 548).

Interestingly, Hoeffding's H has an interpretation in terms of concordance and discordance probabilities closely related to the interpretation of τ^*. With

$$\begin{aligned}
F_{12}(x, y) &= P(X \leq x, Y \leq y),\\
F_{1\bar{2}}(x, y) &= P(X \leq x, Y > y) = F_1(x) - F_{12}(x, y),\\
F_{\bar{1}2}(x, y) &= P(X > x, Y \leq y) = F_2(y) - F_{12}(x, y),\\
F_{\bar{1}\bar{2}}(x, y) &= P(X > x, Y > y) = 1 - F_1(x) - F_2(y) + F_{12}(x, y),
\end{aligned}$$

we have the equality

$$F_{12} - F_1 F_2 = F_{12} F_{\bar{1}\bar{2}} - F_{1\bar{2}} F_{\bar{1}2}. \tag{8}$$

Let five points be H-concordant if four are configured as in Figure 3(a) and the fifth is on the point where the axes cross and, analogously, five points are H-discordant if four are configured

as in Figure 3(b) and the fifth is on the point where the axes cross. Denote the probabilities of H-concordance and discordance by Π_{C_5} and Π_{D_5}. Then, omitting the arguments x and y,

$$\int \bigl(F_{12}^2 F_{\overline{12}}^2 + F_{1\overline{2}}^2 F_{\overline{12}}^2\bigr)\,\mathrm{d}F_{12} = \frac{2!2!1!}{5!}\Pi_{C_5} = \frac{1}{30}\Pi_{C_5}$$

and

$$\int F_{12}F_{1\overline{2}}F_{\overline{1}2}F_{\overline{12}}\,\mathrm{d}F_{12} = \frac{1}{5!}\Pi_{D_5} = \frac{1}{120}\Pi_{D_5}.$$

Hence, using (8),

$$H = \int (F_{12}F_{\overline{12}} - F_{1\overline{2}}F_{\overline{1}2})^2\,\mathrm{d}F_{12} = \frac{2\Pi_{C_5} - \Pi_{D_5}}{60}.$$

We can see that Hoeffding's H has two drawbacks compared to τ^*. Firstly, it is more complex in that it is based on concordance and discordance of five points rather than four and, secondly, it can be zero under dependence for certain discrete distributions.

3.2. The Blum–Kiefer–Rosenblatt coefficient and Spearman's rho

The coefficient D is given by (3), and tests based on it were first studied by Blum, Kiefer and Rosenblatt [4]. It follows from results in Bergsma [2] that in the continuous case, with

$$\begin{aligned} h(z_1,z_2,z_3,z_4) &= |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|, \\ D &= Eh\bigl(F_1(X_1), F_1(X_2), F_1(X_3), F_1(X_4)\bigr) \\ &\quad \times h\bigl(F_2(Y_1), F_2(Y_2), F_2(Y_3), F_2(Y_4)\bigr). \end{aligned} \tag{9}$$

A similar formulation was given by Feuerverger [8], who used characteristic functions for its derivation. This connection of Feuerverger's work to that of Blum, Kiefer and Rosenblatt does not appear to have been noted before.

Replacing absolute values in h by squares, it is straightforward to show that a thus modified D reduces to

$$\begin{aligned} &4\bigl(E\bigl[F_1(X_1) - F_1(X_2)\bigr]\bigl[F_2(Y_1) - F_2(Y_2)\bigr]\bigr)^2 \\ &\quad = 16\bigl(E\bigl[F_1(X_1) - EF_1(X)\bigr]\bigl[F_2(Y_1) - EF_2(Y)\bigr]\bigr)^2 = \tfrac{1}{9}\tilde{\rho}_S^2, \end{aligned}$$

where

$$\tilde{\rho}_S = 12E\bigl[F_1(X_1) - EF_1(X)\bigr]\bigl[F_2(Y_1) - EF_2(Y)\bigr]$$

is a version of Spearman's correlation which coincides with ρ_S given in Section 1 for continuous distributions (see Section 5 in Kruskal [16], for more details).

Following Kruskal's [16] preference for Kendall's tau over Spearman's rho due to its relative simplicity, the same preference might be expressed for τ^* compared to D.

3.3. Comparison to other ordinal consistent tests of independence

We now describe further approaches to obtaining consistent independence tests for ordinal variables described in the literature. It may be noted that H and D are special cases of a general family of coefficients, which can be formulated as

$$Q_{g,h} = Q_{g,h}(X, Y) = \int g\big(\big|F_{12}(x, y) - F_1(x)F_2(y)\big|\big)\,\mathrm{d}\big[h(F_{12})(x, y)\big]. \tag{10}$$

For appropriately chosen g and h, $Q_{g,h} = 0$ if and only if X and Y are independent. Instances were studied by de Wet [5], Deheuvels [6], Schweizer and Wolff [21] and Feuerverger [8] (where the former two focussed on asymptotic distributions of empirical versions, while the latter two focussed on population coefficients). Alternatively, Rényi [19] proposed *maximal correlation*, defined as

$$\rho^+ = \sup_{g,h} \rho\big(g(X), h(Y)\big),$$

where the supremum is taken over square integrable functions. Though applicable to ordinal random variables, ρ^+ does not utilize the ordinal nature of the variables. Furthermore, it is hard to estimate, and has the drawback that it may equal one for distributions arbitrarily "close" to independence (Kimeldorf and Sampson [15]). An ordinal variant, proposed by Kimeldorf and Sampson [15], was to maximize the correlation over non-decreasing square integrable functions.

3.4. Comparison to non-ordinal consistent tests of independence

Recently Székely, Rizzo and Bakirov [26] introduced a consistent test of independence for Euclidean random variables. With ψ_{XY} the characteristic function of the distribution of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$, and ψ_X and ψ_Y the characteristic functions of the corresponding marginal distributions, they defined

$$\mathrm{dcov}^2(X, Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^p \times \mathbb{R}^q} \frac{|\psi_{XY}(s, t) - \psi_X(s)\psi_Y(t)|^2}{\|t\|^{1+p}\|s\|^{1+q}}\,\mathrm{d}s\,\mathrm{d}t, \tag{11}$$

where c_p and c_q are constants. It holds true that $\mathrm{dcov}^2(X, Y) \geq 0$ with equality if and only if X and Y are independent, which is easy to show from the definition. The expression (11) was originally introduced by Feuerverger [8], but only for real X and Y ($p = q = 1$).

It was shown that dcov can equivalently be defined as

$$\mathrm{dcov}^2(X, Y) = E\|X_1 - X_2\|\|Y_1 - Y_2\| + E\|X_1 - X_2\|E\|Y_1 - Y_2\| - 2E\|X_1 - X_2\|\|Y_1 - Y_3\|.$$

From this, it is straightforward to derive that

$$\mathrm{dcov}^2(X, Y) = \tfrac{1}{4}Eh(X_1, X_2, X_3, X_4)h(Y_1, Y_2, Y_3, Y_4),$$

where h is defined by (9). Hence, for the case that X and Y are real (i.e., $p = q = 1$), dcov is closely related to τ^*, τ^* being a sign version.

With Z_1 and Z_2 independent with distribution F, let

$$h_F(z_1, z_2) = -\tfrac{1}{2} E h(z_1, z_2, Z_1, Z_2),$$

where h is defined by (9). It can be verified that

$$\mathrm{dcov}^2(X, Y) = E h_{F_1}(X_1, X_2) h_{F_2}(Y_1, Y_2).$$

As shown by Bergsma [2] for the case that X and Y are real and Sejdinovic *et al.* [22] (see also Sejdinovic *et al.* [23]) for the case that X and Y are Euclidean, h_F is a positive definite kernel implying non-negativity of dcov^2, while further properties of h_F imply equality to zero if and only if X and Y are independent. In fact, as shown explicitly by Sejdinovic *et al.*, dcov^2 falls in a general class of association measures based on positive definite kernels described by Gretton *et al.* [9], which they called the Hilbert–Schmidt independence criterion (HSIC). This criterion is a generalization of Escoufier's vector covariance (Escoufier [7], Robert and Escoufier [20]). It appears that τ^* is not an HSIC.

Although dcov^2 and τ^* are similar in form, proofs of their basic properties are very different. In particular, in spite of its simple mathematical description, the proof for τ^* is much more complex. The reason for this is that it appears hard to formulate τ^* in terms of positive definite kernels, or as the expectation of a squared norm of a random quantity (see also Lyons [17]).

Finally, another recent consistent test of independence for Euclidean random variables is given by Heller, Gorfine and Heller [10], which is based on the summation of Pearson chi-square statistics for well-chosen collapsing of the bivariate distribution onto 2×2 contingency tables.

4. Testing independence

A suitable test for independence is a permutation test which rejects the independence hypothesis for large values of t^*, the empirical value of τ^*. As an exact permutation test is too time consuming for moderately large n, we use a Monte Carlo approximation, which is also called a resampling test, and which is carried out as follows. For $r = 1, 2, \ldots$, let $(i_{r1}, \ldots, i_{rn})$ be a random permutation of $(1, \ldots, n)$, and let t_r^* be t^* computed for the rth resample $(X_1, Y_{i_{r1}}), \ldots, (X_n, Y_{i_{rn}})$. Then the Monte Carlo permutation p-value based on R resamples is computed as

$$\text{Monte Carlo } p\text{-value} = \frac{1}{R} \sum_{r=1}^{R} I\big(t_r^* > t^*\big).$$

A further computational problem is the evaluation of t^* itself (and of the t_r^*), which requires computational time $\mathrm{O}(n^4)$, and may be practically infeasible for moderately large samples. However, t^* can be well-approximated by taking a sufficiently large random sample of subsets of four observations to approximate the sum in (5).

As is well known, the permutation test conditions on the empirical marginal distributions, which are sufficient statistics for the independence model. In categorical data analysis, it is usually referred to as an exact conditional test. Note that there does not seem to be a need for an asymptotic approximation to the sampling distribution of t^*.

Table 1. Artificial contingency table containing multinomial counts. Permutation tests based on Kendall's tau and the Pearson chi-square statistic do not yield a significant association ($p = 0.99$, resp., $p = 0.25$), but a permutation test based on t^* yields $p = 0.035$

	Y						
X	1	2	3	4	5	6	7
1	2	1	0	0	0	1	2
2	1	2	0	0	0	2	1
3	0	0	2	1	2	0	0
4	0	0	1	1	1	0	0
5	0	0	1	2	1	0	0

In this section, we compare various tests of independence using an artificial and a real data set and via a simulation study.

4.1. Examples

An artificial multinomial table of counts is given in Table 1, where X and Y are ordinal variables with 5 and 7 categories. Visually, we can detect an association pattern, but as it is non-monotonic a test based on Kendall's tau does not yield a significant p-value. The chi-square test also yields a non-significant $p = 0.252$, while a permutation test based on t^* yields $p = 0.032$, giving evidence of an association. We also did tests based on D, which yields $p = 0.047$, and the test based on Hoeffding's H yields $p = 0.028$. In this example, using a consistent test designed for ordinal data, evidence for an association can be found, which is not possible with a nominal data test like the chi-square test or with a test based on Kendall's tau. For all tests except Hoeffding's $R = 10^6$ resamples were used, and for Hoeffding's test $R = 4000$ resamples were used.

Table 2 shows data from a randomized study to compare two treatments for a gastric ulcer crater, and was previously analyzed in Agresti [1]. Using $R = 10^5$ resamples, the chi-square test yields $p = 0.118$, Kendall's tau yields $p = 0.019$, t^* yields $p = 0.028$, D yields $p = 0.026$, and using 10^4 resamples Hoeffding's H yields $p = 0.006$.

Table 2. Results of study comparing two treatments of gastric ulcer

	Change in size of Ulcer Crater (Y)			
Treatment group (X)	Larger	Healed ($< \frac{2}{3}$)	Healed ($\geq \frac{2}{3}$)	Healed
A	6	4	10	12
B	11	8	8	5

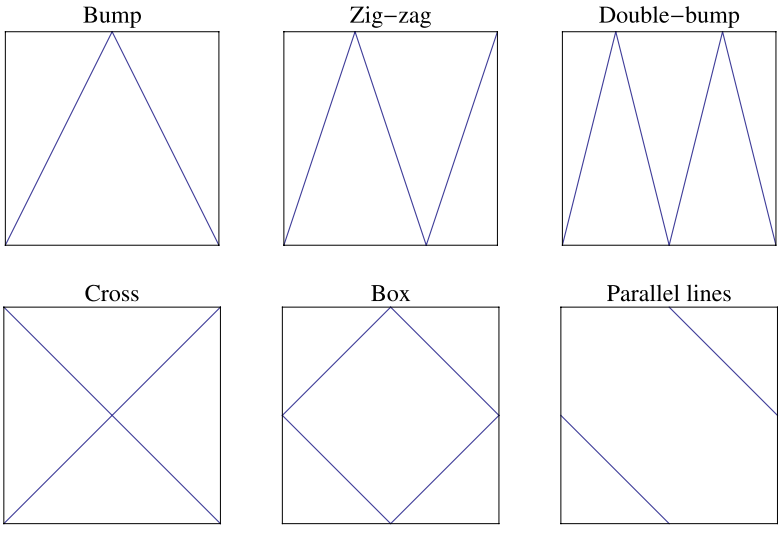


Figure 4. Simulations were done for data generated from the uniform distribution on the lines within each of the six boxes. For all except the Zig–zag and the Parallel lines, the ordinary correlation is zero.

4.2. Simulated average p-values for independence tests based on D, H, and τ^*

Any of the three tests can be expected to have most power of the three for certain alternatives, and least power of the three for others. Given the broadness of possible alternatives, it cannot be hoped to get a simple description of alternatives for which any single test is the most powerful. However, some insight may be gained by looking at average p-values for a set of carefully selected alternatives.

In Figure 4, six boxes with lines in them are represented, and we simulated from the uniform distribution on these lines. The first five maximize or minimize the correlation between some simple orthogonal functions for given uniform marginals. In particular, say the boxes represent the square $[0,1] \times [0,1]$, then the Bump, Zig–zag and Double bump distributions maximize, for given uniform marginals,

$$\rho\big[\cos(2\pi X), \cos(\pi Y)\big], \qquad \rho\big[\cos(3\pi X), \cos(\pi Y)\big] \quad \text{and} \quad \rho\big[\cos(4\pi X), \cos(\pi Y)\big],$$

respectively. The Cross and Box distributions respectively maximize and minimize, for given uniform marginals,

$$\rho\big[\cos(2\pi X), \cos(2\pi Y)\big].$$

As they represent in this sense extreme forms of association, these distributions should yield good insight in the comparative performance of the tests. Furthermore, the Parallel lines distribution was chosen because it is simple and demonstrates a weakness of Hoeffding's test, as it has comparatively very little power here (we did not manage to find a distribution where D or τ^*

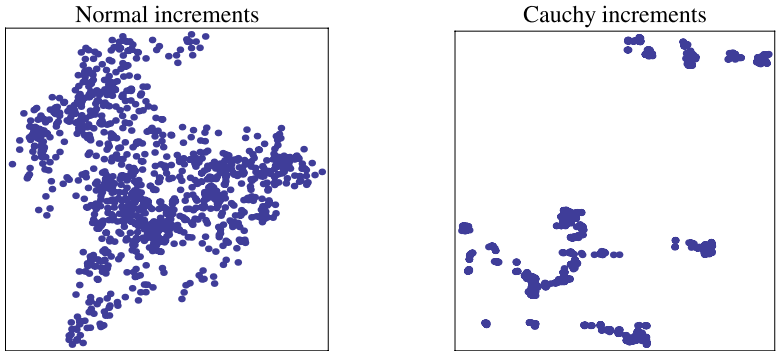


Figure 5. 1000 points of a random walk. In the first plot the (x, y) increments are independent normals, in the second they are independent Cauchy variables.

fare so comparatively poorly). Note that all six distributions have uniform marginals and so are copulas, and several were also discussed in Nelsen [18].

We also did a Bayesian simulation, based on random distributions with dependence. In particular, the data are $(X_1, Y_1), \ldots, (X_n, Y_n)$, where, for i.i.d. $(\varepsilon_{1i}, \varepsilon_{2i})$,

$$(X_1, Y_1) = (\varepsilon_{11}, \varepsilon_{21}),$$
$$(X_{i+1}, Y_{i+1}) = (X_i, Y_i) + (\varepsilon_{1i}, \varepsilon_{2i}) \qquad i = 1, \ldots, n-1.$$

Of course, the (X_i, Y_i) are not i.i.d., but conditioning on the empirical marginals the permutations of the Y-values give equally likely data sets under the null hypothesis of independence, so the permutation test is valid. Two distributions for the increments $(\varepsilon_{1i}, \varepsilon_{2i})$ were used: independent normals and independent Cauchy distributions. In Figure 5, points generated in this way are plotted. Note that for the Cauchy increments, the heavy tails of the marginal distributions are automatically taken care of by the use of ranks, so in that respect the three tests described here are particularly suitable.

Finally, we also simulated normally distributed data with correlation 0.5.

Average p-values are given in Table 3, where all averages are over at least 40,000 simulations (for D, we did 200,000 simulations). Hoeffding's test compares extremely badly with our test for the parallel lines distribution, and is worse than our test for the random walks, but outperforms our test for the Zig–zag, Double-bump, Cross and Box distributions. The reason for the poor performance of Hoeffding's test for the parallel lines distribution is that five points can only be concordant (see Section 3.1) if they all lie on a single line (a discordant set of five points has zero probability). Similarly, for the Zig–zag, Double-bump and Cross concordant sets of five points can be seen to be especially likely, so these choices of distributions favour the Hoeffding test. Note that Hoeffding's test is less suitable for general use because it is not necessarily zero under independence if there is a positive probability of tied observations.

The BKR test fares slightly worse than ours for the random walk with Cauchy increments, and significantly worse than ours for the Bump, Zig–zag, Cross and Box distributions, and does somewhat better than ours for the normal distribution. It appears that the BKR test has more

Table 3. Average p-values. See Figures 4 and 5 and the text for explanations

Distribution	Sample size n	Average p-value		
		D	H	τ^*
Random walk (normal increments)	50	0.061	0.080	0.061
Random walk (Cauchy increments)	30	0.039	0.065	0.031
Bump	12	0.087	0.061	0.045
Zig–zag	25	0.083	0.011	0.036
Double-bump	30	0.056	0.005	0.019
Cross	50	0.052	0.003	0.021
Box	50	0.070	0.008	0.019
Parallel lines	10	0.055	0.710	0.076
Normal distribution ($\rho = 0.5$)	30	0.055	0.052	0.073

power than ours for a monotone alternative (such as the normal distribution), at the cost of less power for some more complex alternatives.

5. Proof of Theorem 1

Here we give the proof of Theorem 1 for arbitrary real random variables X and Y. A shorter proof for continuous X and Y is given by Dassios and Bergsma [3]. Readers wishing to gain an understanding of the essence of the proof may wish to study the shorter proof first.

First, consider three real valued random variables U, V and W. They have continuous densities $\tilde{f}(x)$, $\tilde{g}(x)$ and $\tilde{k}(x)$ as well as probability masses $f(x_i)$, $g(x_i)$ and $k(x_i)$ at points $x_1, x_2, \ldots$. We also define

$$F(x) = P(U < x) = \sum_{x_i < x} f(x_i) + \int_{y<x} \tilde{f}(y)\,\mathrm{d}y,$$

$$G(x) = P(V < x) = \sum_{x_i < x} g(x_i) + \int_{y<x} \tilde{g}(y)\,\mathrm{d}y$$

and

$$K(x) = P(W < x) = \sum_{x_i < x} k(x_i) + \int_{y<x} \tilde{k}(y)\,\mathrm{d}y.$$

We will also use $H(x) = \frac{K(x)}{G(x)}$. Note that $H(x)$ also admits the representation

$$H(x) = \sum_{x_i < x} h(x_i) + \int_{y<x} \tilde{h}(y)\,\mathrm{d}y$$

but unlike the other three function that are non-decreasing $\tilde{h}(x)$ and $h(x_i)$ can take negative values.

We start by proving the following intermediate result.

Lemma 1. *Assume that $G(x) = 1$ implies $F(x) = K(x) = 1$ and that there is a constant c such that $F(x) \le cG(x)$ and $K(x) \le cG^2(x)$ for all x. Define*

$$\begin{aligned} S = {} & 2\sum \big(F(x_i) - G(x_i)\big)\big(F(x_i)g(x_i) - G(x_i)f(x_i)\big)\frac{K(x_i)}{G^2(x_i)} \\ & - \sum \big(F(x_i)g(x_i) - G(x_i)f(x_i)\big)^2 \frac{K(x_i)}{G^2(x_i)} \\ & + 2\int \big(F(x) - G(x)\big)\big(F(x)\tilde{g}(x) - G(x)\tilde{f}(x)\big)\frac{K(x)}{G^2(x)}\,\mathrm{d}x, \end{aligned}$$

where summation is over all x_i such that $K(x_i) > 0$ and at least one of $f(x_i)$ and $g(x_i)$ is positive, and integration is over all x such that $K(x) > 0$.

We then have $S \ge 0$ with equality iff $F(x) = G(x)$ for all x such that $K(x) > 0$.

Proof. The conditions stated in the lemma ensure that the sums and integral exist. We can rewrite

$$\begin{aligned} S = {} & 2\sum \big(F(x_i) - G(x_i)\big)\big(F(x_i)g(x_i) - G(x_i)f(x_i)\big)\frac{H(x_i)}{G(x_i)} \\ & - \sum \big(F(x_i)g(x_i) - G(x_i)f(x_i)\big)^2 \frac{H(x_i)}{G(x_i)} \\ & + 2\int \big(F(x) - G(x)\big)\big(F(x)\tilde{g}(x) - G(x)\tilde{f}(x)\big)\frac{H(x)}{G(x)}\,\mathrm{d}x. \end{aligned}$$

For simplicity, we denote $F(x), G(x), H(x), f(x_i), g(x_i), h(x_i), \tilde{f}(x), \tilde{g}(x)$ and $\tilde{h}(x)$ by $F, G, H, f, g, h, \tilde{f}, \tilde{g}$ and $\tilde{h}$. We have

$$\begin{aligned} S = {} & 2\sum (F - G)\big((F - G)g - G(f - g)\big)\frac{H}{G} \\ & + 2\int (F - G)\big((F - G)\tilde{g} - G(\tilde{f} - \tilde{g})\big)\frac{H}{G}\,\mathrm{d}x \\ & - \sum \big((F - G)g - G(f - g)\big)^2 \frac{H}{G} \\ = {} & 2\sum (F - G)^2 \frac{H}{G} g + 2\int (F - G)^2 \frac{H}{G}\tilde{g}\,\mathrm{d}x \qquad (12) \\ & - 2\sum H(F - G)(f - g) \end{aligned}$$

$$
\begin{aligned}
&- 2\int H(F-G)(\tilde{f}-\tilde{g})\,\mathrm{d}x \\
&- \sum\bigl((F-G)g - G(f-g)\bigr)^2\frac{H}{G}.
\end{aligned}
$$

The function $H(F-G)^2$ vanishes at $-\infty$ (because of the conditions of the lemma) and $+\infty$. Considering its integral and sum representation we have

$$
\begin{aligned}
&2\sum H(F-G)(f-g) + 2\int H(F-G)(\tilde{f}-\tilde{g})\,\mathrm{d}x \\
&\quad + \sum (F-G)^2 h + \int (F-G)^2\tilde{h}\,\mathrm{d}x \\
&\quad + 2\sum (F-G)(f-g)h + \sum (f-g)^2 h + \sum H(f-g)^2 = 0,
\end{aligned}
$$

and therefore

$$
\begin{aligned}
&-2\sum H(F-G)(f-g) - 2\int H(F-G)(\tilde{f}-\tilde{g})\,\mathrm{d}x \\
&\quad = \sum (F-G)^2 h + \int (F-G)^2\tilde{h}\,\mathrm{d}x \\
&\qquad + 2\sum (F-G)(f-g)h + \sum (f-g)^2 h + \sum H(f-g)^2.
\end{aligned}
\tag{13}
$$

Moreover,

$$
\begin{aligned}
&\frac{H}{G}\bigl((F-G)g - G(f-g)\bigr)^2 \\
&\quad = (F-G)^2 g^2\frac{H}{G} + GH(f-g)^2 - 2(F-G)(f-g)Hg.
\end{aligned}
\tag{14}
$$

Substituting (13) and (14) into (12), and denoting $M = F-G$, $m = f-g$ and $\tilde{m} = \tilde{f}-\tilde{g}$ we have

$$
\begin{aligned}
S &= \sum M^2\left(2g\frac{H}{G} + h - g^2\frac{H}{G}\right) + 2\sum Mm(h+gH) + \sum m^2(H+h-GH) \\
&\quad + \int M^2\left(2\tilde{g}\frac{H}{G} + \tilde{h}\right)\mathrm{d}x \\
&= \sum (M+m)^2\left(g\frac{H}{G+g} + h\right) + \sum M^2\left(2g\frac{H}{G} - g\frac{H}{G+g} - g^2\frac{H}{G}\right) \\
&\quad - 2\sum Mm\left(g\frac{H}{G+g} - gH\right) + \sum m^2\left(H - GH - g\frac{H}{G+g}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\int M^2\left(\tilde{g}\frac{H}{G}+\tilde{h}\right)\mathrm{d}x+\int M^2\tilde{g}\frac{H}{G}\,\mathrm{d}x\\
&=\sum(M+m)^2\left(g\frac{H}{G+g}+h\right)+\int M^2\left(\tilde{g}\frac{H}{G}+\tilde{h}\right)\mathrm{d}x+\int M^2\tilde{g}\frac{H}{G}\,\mathrm{d}x\\
&\quad+\sum M^2\left(g\frac{H}{G}+g^2\frac{H(1-G-g)}{G(G+g)}\right)-2\sum Mm\left(g\frac{H(1-G-g)}{G+g}\right)\\
&\quad+\sum m^2\frac{H}{G+g}\big((1-G)G-gG\big).
\end{aligned}
$$

Observe now that since $K=HG$

$$
g\frac{H}{G+g}+h=\frac{gH+hG+hg}{G+g}=\frac{k}{G+g}\geq 0
$$

and

$$
\tilde{g}\frac{H}{G}+\tilde{h}=\frac{\tilde{k}}{G}\geq 0.
$$

Moreover, the quadratic form

$$
\begin{aligned}
&M^2\left(g\frac{H}{G}+g^2\frac{H(1-G-g)}{G(G+g)}\right)-2Mm\left(g\frac{H(1-G-g)}{G+g}\right)+m^2\frac{H}{G+g}\big((1-G)G-gG\big)\\
&\quad=\frac{M^2gH}{G}+(Mg-mG)^2\frac{H(1-G-g)}{G(G+g)}.
\end{aligned}
$$

All terms in S are non-negative and are equal to zero iff $M(x)=0$ for all x such that $K(x)>0$, that is the two distributions F and G are identical for all x such that $K(x)>0$. □

Before we prove Theorem 1, we will prove another result as it will be used repeatedly.

Lemma 2. *Let A, B and C be events in the same probability space as the random variable X and define*

$$
\begin{aligned}
L\big(x^{(1)},x^{(2)}\big)&=\big(P\big(A|X=x^{(1)}\big)-P\big(A|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\quad\times\big(P\big(A|X=x^{(2)}\big)-P\big(A|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\quad\times P\big(B|X<x^{(1)}\wedge x^{(2)}\big)P\big(C|X<x^{(1)}\wedge x^{(2)}\big)\big(P\big(X<x^{(1)}\wedge x^{(2)}\big)\big)^2.
\end{aligned}
$$

We then have

$$
E\big(L(X_1,X_2)\big)\geq 0
$$

with equality iff $P(X<x)=P(X<x|A)$ for all x such that $P(X<x|B)P(X<x|C)>0$.

Proof. Let X have continuous density $\tilde{g}(x)$ and probability masses $g(x_i)$ at points $x_1, x_2, \ldots$ and let X have continuous density $\tilde{g}_A(x)$ and probability masses $g_A(x_i)$ at points $x_1, x_2, \ldots$ conditionally on A. Define also

$$G(x) = P(X < x) = \sum_{x_i < x} g(x_i) + \int_{y<x} \tilde{g}(y)\,\mathrm{d}y$$

and

$$G_A(x) = P\big(X < x|A\big) = \sum_{x_i < x} g_A(x_i) + \int_{y<x} \tilde{g}_A(y)\,\mathrm{d}y.$$

Conditioning on values of $X_1 \wedge X_2$ and using Bayes' theorem, we can see that

$$\begin{aligned}
&E\big(L(X_1, X_2)\big)\\
&\quad = \big(P(A)\big)^2 \sum P\big(B|X < x_i\big) P\big(C|X < x_i\big)\\
&\qquad \times \big\{2\big(\big(1 - G_A(x_i)\big)G(x_i) - \big(1 - G(x_i)\big)G_A(x_i)\big)\big(g_A(x_i)G(x_i) - g(x_i)G_A(x_i)\big)\\
&\qquad\quad - \big(g_A(x_i)G(x_i) - g(x_i)G_A(x_i)\big)^2\big\}\\
&\qquad + \big(P(A)\big)^2 \int P\big(B|X < x\big) P\big(C|X < x\big)\\
&\qquad\qquad \times 2\big(\big(1 - G_A(x)\big)G(x) - \big(1 - G(x)\big)G_A(x)\big)\big(\tilde{g}_A(x)G(x) - \tilde{g}(x)G_A(x)\big)\,\mathrm{d}x\\
&\quad = P(B)P(C)\big(P(A)\big)^2 \sum \frac{K(x_i)}{G^2(x_i)} \cdot \big\{2\big(G(x_i) - G_A(x_i)\big)\big(g_A(x_i)G(x_i) - g(x_i)G_A(x_i)\big)\\
&\qquad\qquad - \big(g_A(x_i)G(x_i) - g(x_i)G_A(x_i)\big)^2\big\}\\
&\qquad + P(B)P(C)\big(P(A)\big)^2 \int \frac{K(x)}{G^2(x)} 2\big(G(x) - G_A(x)\big)\big(\tilde{g}_A(x)G(x) - \tilde{g}(x)G_A(x)\big)\,\mathrm{d}x,
\end{aligned}$$

where

$$K(x) = P\big(X < x|B\big) P\big(X < x|C\big).$$

The result then follows from Lemma 1 ($F = G_A$). It is easy to see that the conditions in Lemma 1 are satisfied. For example, $P(X < x|B) \le \frac{P(X<x)}{P(B)}$. □

Proof of Theorem 1. We need to prove that

$$\begin{aligned}
&P(Y_1 \wedge Y_2 > Y_3 \vee Y_4, X_3 \vee X_4 < X_1 \wedge X_2)\\
&\quad + P(Y_1 \vee Y_2 < Y_3 \wedge Y_4, X_3 \vee X_4 < X_1 \wedge X_2)\\
&\quad - P(Y_1 \wedge Y_3 > Y_2 \vee Y_4, X_3 \vee X_4 < X_1 \wedge X_2)\\
&\quad - P(Y_1 \vee Y_3 < Y_2 \wedge Y_4, X_3 \vee X_4 < X_1 \wedge X_2) \ge 0
\end{aligned}$$

with equality in the independence case.

Let (X, Y) represent any of the pairs (X_i, Y_i). Define now $F_1(y) = P(Y < y|X = x^{(1)})$, $F_2(y) = P(Y < y|X = x^{(2)})$ and $G(y) = P(Y < y|X < x^{(1)} \wedge x^{(2)})$ with the representations

$$F_1(y) = \sum_{y_i < y} f_1(y_i) + \int_{z<y} \widetilde{f_1}(z)\,\mathrm{d}z,$$

$$F_2(y) = \sum_{y_i < y} f_2(y_i) + \int_{z<y} \widetilde{f_2}(z)\,\mathrm{d}z$$

and

$$G(y) = \sum_{y_i < y} g(y_i) + \int_{z<y} \tilde{g}(z)\,\mathrm{d}z.$$

Note that conditionally on the event

$$\Theta = \left\{X_1 = x^{(1)}, X_2 = x^{(2)}, X_3 < x^{(1)} \wedge x^{(2)}, X_4 < x^{(1)} \wedge x^{(2)}\right\},$$

the distribution of the minimum of Y_1 and Y_2 has density $(1-F_1)\widetilde{f_2} + (1-F_2)\widetilde{f_1}$ and probability masses $(1-F_1)f_2 + (1-F_2)f_1 - f_1 f_2$ at $y_1, y_2, \ldots$, the distribution of the minimum of Y_3 and Y_4 has density $2(1-G)\tilde{g}$ and probability masses $2(1-G)g - g^2$, the distribution of the minimum of Y_1 and Y_3 has density $(1-F_1)\tilde{g} + (1-G)\widetilde{f_1}$ and probability masses $(1-F_1)g + (1-G)f_1 - f_1 g$ and the distribution of the minimum of Y_2 and Y_4 has density $(1-F_2)\tilde{g} + (1-G)\widetilde{f_2}$ and probability masses $(1-F_2)g + (1-G)f_2 - f_2 g$. We therefore have (suppressing the arguments of the functions)

$$\begin{aligned}
&P\big(Y_1 \wedge Y_2 > Y_3 \vee Y_4|\Theta\big) + P\big(Y_1 \vee Y_2 < Y_3 \wedge Y_4|\Theta\big) \\
&\quad - P\big(Y_1 \wedge Y_3 > Y_2 \vee Y_4|\Theta\big) - P\big(Y_1 \vee Y_3 < Y_2 \wedge Y_4|\Theta\big) \\
&= \sum\big((1-F_1)f_2 + (1-F_2)f_1 - f_1 f_2\big)G^2 + \sum\big(2(1-G)g - g^2\big)F_1 F_2 \\
&\quad - \sum\big((1-F_1)g + (1-G)f_1 - f_1 g\big)F_2 G - \sum\big((1-F_2)g + (1-G)f_2 - f_2 g\big)F_1 G \\
&\quad + \int\big((1-F_1)\widetilde{f_2} + (1-F_2)\widetilde{f_1}\big)G^2\,\mathrm{d}y + \int 2(1-G)\tilde{g}F_1 F_2\,\mathrm{d}y \\
&\quad - \int\big((1-F_1)\tilde{g} + (1-G)\widetilde{f_1}\big)F_2 G\,\mathrm{d}y - \int\big((1-F_2)\tilde{g} + (1-G)\widetilde{f_2}\big)F_1 G\,\mathrm{d}y \\
&= \sum(F_1 - G)(F_2 g - G f_2) + \sum(F_2 - G)(F_1 g - G f_1) - \sum(F_1 g - G f_1)(F_2 g - G f_2) \\
&\quad + \int(F_1 - G)(F_2\tilde{g} - G\widetilde{f_2})\,\mathrm{d}y + \int(F_2 - G)(F_1\tilde{g} - G\tilde{f_1})\,\mathrm{d}y \\
&= 2\sum(F_1 - G)(F_2 - G)g - \sum(F_1 - G)(f_2 - g)G - \sum(F_2 - G)(f_1 - g)G
\end{aligned}$$

$$
\begin{aligned}
&- \sum (F_1 g - G f_1)(F_2 g - G f_2) + 2 \int (F_1 - G)(F_2 - G)\tilde{g}\,\mathrm{d}y \\
&- \int (F_1 - G)(\tilde{f}_2 - \tilde{g})G\,\mathrm{d}y - \int (F_2 - G)(\tilde{f}_1 - \tilde{g})G\,\mathrm{d}y.
\end{aligned}
$$

The function $G(F_1 - G)(F_2 - G)$ vanishes at $-\infty$ and $+\infty$. Considering its integral and sum representation, we have

$$
\begin{aligned}
&- \sum (F_1 - G)(f_2 - g)G - \sum (F_2 - G)(f_1 - g)G \\
&\quad - \int (F_1 - G)(\tilde{f}_2 - \tilde{g})G\,\mathrm{d}y - \int (F_2 - G)(\tilde{f}_1 - \tilde{g})G\,\mathrm{d}y \\
&= \sum (F_1 - G)(F_2 - G)g + \sum (F_1 - G)(f_2 - g)g + \sum (F_2 - G)(f_1 - g)g \\
&\quad + \sum (f_1 - g)(f_2 - g)G + \sum (f_2 - g)(f_1 - g)g \\
&\quad + \int (F_1 - G)(F_2 - G)\tilde{g}\,\mathrm{d}y \\
&= \sum (F_1 + f_1 - G - g)(F_2 + f_2 - G - g)g \\
&\quad + \sum (f_1 - g)(f_2 - g)G + \int (F_1 - G)(F_2 - G)\tilde{g}\,\mathrm{d}y.
\end{aligned} \tag{15}
$$

Moreover,

$$
\begin{aligned}
(F_1 g - G f_1)(F_2 g - G f_2) &= (F_1 - G)(F_2 - G)g^2 - (F_1 - G)(f_2 - g)Gg \\
&\quad - (F_2 - G)(f_1 - g)Gg + (f_1 - g)(f_2 - g)G^2 \\
&= (F_1 - G)(F_2 - G)g^2 + (f_1 - g)(f_2 - g)G^2 \\
&\quad - (F_1 + f_1 - G - g)(F_2 + f_2 - G - g)gG \\
&\quad + (F_1 - G)(F_2 - G)gG + (f_1 - g)(f_2 - g)gG \\
&= (F_1 - G)(F_2 - G)g(G + g) + (f_1 - g)(f_2 - g)G(G + g) \\
&\quad - (F_1 + f_1 - G - g)(F_2 + f_2 - G - g)gG.
\end{aligned} \tag{16}
$$

Using (16) and (15), we have

$$
\begin{aligned}
&P\big(Y_1 \wedge Y_2 > Y_3 \vee Y_4 | \Theta\big) + P\big(Y_1 \vee Y_2 < Y_3 \wedge Y_4 | \Theta\big) \\
&\quad - P\big(Y_1 \wedge Y_3 > Y_2 \vee Y_4 | \Theta\big) - P\big(Y_1 \vee Y_3 < Y_2 \wedge Y_4 | \Theta\big) \\
&= \sum (F_1 - G)(F_2 - G)g + \sum (F_1 - G)(F_2 - G)g(1 - G - g) \\
&\quad + \sum (F_1 + f_1 - G - g)(F_2 + f_2 - G - g)g
\end{aligned}
$$

$$
\begin{aligned}
&+\sum(F_1+f_1-G-g)(F_2+f_2-G-g)gG\\
&+\sum(f_1-g)(f_2-g)G(1-G-g)+3\int(F_1-G)(F_2-G)\tilde{g}\,\mathrm{d}y.
\end{aligned}
$$

We therefore conclude that conditionally on $\{X_1=x^{(1)},X_2=x^{(2)}\}$,

$$
\begin{aligned}
&P(Y_1\wedge Y_2>Y_3\vee Y_4,X_3\vee X_4<X_1\wedge X_2)\\
&\quad+P(Y_1\vee Y_2<Y_3\wedge Y_4,X_3\vee X_4<X_1\wedge X_2)\\
&\quad-P(Y_1\wedge Y_3>Y_2\vee Y_4,X_3\vee X_4<X_1\wedge X_2)\\
&\quad-P(Y_1\vee Y_3<Y_2\wedge Y_4,X_3\vee X_4<X_1\wedge X_2)\\
&=\sum\big(P\big(Y<y|X=x^{(1)}\big)-P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times\big(P\big(Y<y|X=x^{(2)}\big)-P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times P\big(Y=y|X<x^{(1)}\wedge x^{(2)}\big)\big(P\big(X<x^{(1)}\wedge x^{(2)}\big)\big)^2\\
&\quad+\sum\big(P\big(Y<y|X=x^{(1)}\big)-P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times\big(P\big(Y<y|X=x^{(2)}\big)-P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times P\big(Y=y|X<x^{(1)}\wedge x^{(2)}\big)P\big(Y>y|X<x^{(1)}\wedge x^{(2)}\big)\big(P\big(X<x^{(1)}\wedge x^{(2)}\big)\big)^2\\
&\quad+\sum\big(P\big(Y\le y|X=x^{(1)}\big)-P\big(Y\le y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times\big(P\big(Y\le y|X=x^{(2)}\big)-P\big(Y\le y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times P\big(Y=y|X<x^{(1)}\wedge x^{(2)}\big)\big(P\big(X<x^{(1)}\wedge x^{(2)}\big)\big)^2\\
&\quad+\sum\big(P\big(Y\le y|X=x^{(1)}\big)-P\big(Y\le y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times\big(P\big(Y\le y|X=x^{(2)}\big)-P\big(Y\le y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times P\big(Y=y|X<x^{(1)}\wedge x^{(2)}\big)P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)\big(P\big(X<x^{(1)}\wedge x^{(2)}\big)\big)^2\\
&\quad+\sum\big(P\big(Y=y|X=x^{(1)}\big)-P\big(Y=y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times\big(P\big(Y=y|X=x^{(2)}\big)-P\big(Y=y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)P\big(Y>y|X<x^{(1)}\wedge x^{(2)}\big)\big(P\big(X<x^{(1)}\wedge x^{(2)}\big)\big)^2\\
&\quad+3\int\big(P\big(Y<y|X=x^{(1)}\big)-P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times\big(P\big(Y<y|X=x^{(2)}\big)-P\big(Y<y|X<x^{(1)}\wedge x^{(2)}\big)\big)\\
&\qquad\times P\big(Y\in\mathrm{d}y|X<x^{(1)}\wedge x^{(2)}\big)\big(P\big(X<x^{(1)}\wedge x^{(2)}\big)\big)^2.
\end{aligned}
$$

All of the above terms lead to non-negative expressions because of Lemma 2 (for the first, third and sixth term we take $C = \Omega$, the set of all possible outcomes). We then see that the expression can be zero iff X and Y are independent. The condition stated in Theorem 1 is needed to avoid complications when integrating over $x^{(1)}$ in the application of Lemma 2 to terms such as $\sum P(Y = y|X = x^{(1)})$. □

Acknowledgements

We would like to thank the anonymous referee for useful comments. We would also like to thank the Associate Editor for many insightful and important suggestions that greatly improved this paper.

Supplementary Material

A shorter proof of the main theorem for the continuous case and some miscellaneous further results (DOI: 10.3150/13-BEJ514SUPP; .pdf). The supplement contains the following results: (i) a shorter proof of the main theorem, but only for the continuous case, (ii) the Cramér von Mises test as a special case, (iii) a shorter proof of main theorem for the case that one of the variables is binary, and (iv) a result for an extension to the case of variables in metric spaces.

References

[1] Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR2742515

[2] Bergsma, W.P. (2006). A new correlation coefficient, its orthogonal decomposition, and associated tests of independence. Available at arXiv:math/0604627v1 [math.ST].

[3] Bergsma, W. and Dassios, A. (2014). Supplement to "A consistent test of independence based on a sign covariance related to Kendall's tau." DOI:10.3150/13-BEJ514SUPP.

[4] 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. MR0125690

[5] de Wet, T. (1980). Cramér–von Mises tests for independence. *J. Multivariate Anal.* **10** 38–50. MR0569795

[6] Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Multivariate Anal.* **11** 102–113. MR0612295

[7] Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* **29** 751–760. MR0334416

[8] Feuerverger, A. (1993). A consistent test for bivariate dependence. *International Statistical Review/Revue Internationale de Statistique* **61** 419–433.

[9] Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory. Lecture Notes in Computer Science* **3734** 63–77. Berlin: Springer. MR2255909

[10] Heller, R., Heller, Y. and Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. Available at arXiv:1201.3522.

[11] Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Math. Statistics* **19** 546–557. MR0029139

[12] Hollander, M. and Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd ed. *Wiley Series in Probability and Statistics*: *Texts and References Section*. New York: Wiley. MR1666064
[13] Kendall, M. and Gibbons, J.D. (1990). *Rank Correlation Methods*, 5th ed. *A Charles Griffin Title*. London: Edward Arnold. MR1079065
[14] Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.
[15] Kimeldorf, G. and Sampson, A.R. (1978). Monotone dependence. *Ann. Statist.* **6** 895–903. MR0491562
[16] Kruskal, W.H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.* **53** 814–861. MR0100941
[17] Lyons, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41** 3284–3305.
[18] Nelsen, R.B. (2006). *An Introduction to Copulas*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2197664
[19] Rényi, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hungar.* **10** 441–451. MR0115203
[20] Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **25** 257–265. MR0440801
[21] Schweizer, B. and Wolff, E.F. (1981). On nonparametric measures of dependence for random variables. *Ann. Statist.* **9** 879–885. MR0619291
[22] Sejdinovic, D., Gretton, A., Sriperumbudur, B. and Fukumizu, K. (2012). Hypothesis testing using pairwise distances and associated kernels. In *Proc. International Conference on Machine Learning*. Edinburgh, UK: ICML.
[23] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2012). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Available at arXiv:1207.6076.
[24] Sheskin, D.J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th ed. Boca Raton, FL: Chapman & Hall/CRC. MR2296053
[25] Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology* **15** 72–101.
[26] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665
[27] Wilding, G.E. and Mudholkar, G.S. (2008). Empirical approximations for Hoeffding's test of bivariate independence using two Weibull extensions. *Stat. Methodol.* **5** 160–170. MR2424751

Received July 2012 and revised January 2013