Sankhyā : The Indian Journal of Statistics
2021, Volume 83-B, Supplement 1, pp. S12-S25
© 2021, Indian Statistical Institute

A Basic Treatment of the Distance Covariance

Dominic Edelmann and Tobias Terzer
German Cancer Research Center, Heidelberg, Germany
Donald Richards
Pennsylvania State University, University Park, USA

Abstract

The distance covariance of Székely et al. (*Ann. Statist.*, **35**, 2769–2794 207, 2009), a powerful measure of dependence between sets of multivariate random variables, has the crucial feature that it equals zero if and only if the sets are mutually independent. Hence the distance covariance can be applied to multivariate data to detect arbitrary types of non-linear associations between sets of variables. We provide in this article a basic, albeit rigorous, introductory treatment of the distance covariance. Our investigations yield an approach that can be used as the foundation for presentation of this important and timely topic even in advanced undergraduate- or junior graduate-level courses on mathematical statistics.

AMS (2000) subject classification. Primary 62G10, 62H20; Secondary 60E10, 62G20.
Keywords and phrases. Asymptotic distribution, Distance correlation, Multivariate tests of independence, Orthogonal transformations, U-statistics

1 Introduction

The distance covariance, a measure of dependence between multivariate random variables X and Y, was introduced by Székely et al. (2007) and has since received extensive attention in the statistical literature. A crucial feature of the distance covariance is that it equals zero if and only if X and Y are mutually independent. Hence the distance covariance is sensitive to arbitrary dependencies; this is in contrast to the classical covariance, which is generally capable of detecting only linear dependencies. This property is illustrated in Fig. 1, which illustrates that tests based on the distance covariance are able to detect numerous types of non-linear associations even when tests based on the classical covariance may fail to detect many such statistical relationships.

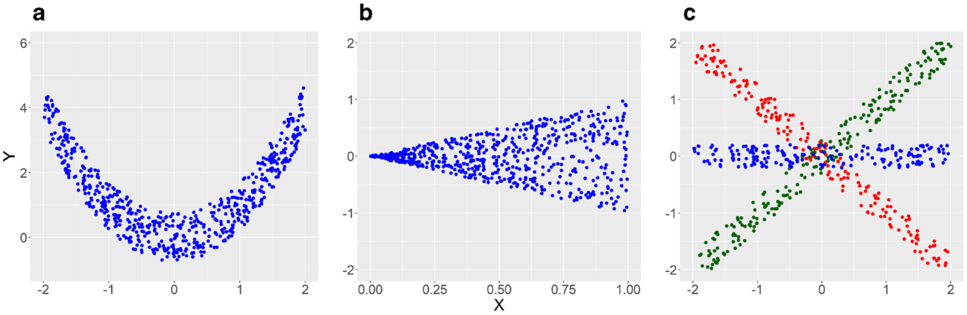

Figure 1: The sub-figures **A-C** represent scatter-plots of bivariate samples $(\boldsymbol{X}, \boldsymbol{Y})$ with $n = 600$ data points to which independence tests, based on the distance covariance and classical covariance, were applied. In each case a distance covariance permutation test using 100,000 permutations yields p-values of 10^{-5}, demonstrating that the distance covariance is able to detect these dependencies. The p-values of permutation tests based on the classical covariance with 100,000 permutations are 0.663, 0.129, and 0.889 for **A**, **B**, and **C**, respectively

While the dependencies illustrated in Fig. 1 clearly represent purely illustrative examples, the sensitivity of the distance covariance to arbitrary dependencies can be very useful for applications. This is demonstrated in Fig. 2, where we show three dependencies between expression values genes in the breast cancer data set by Van De Vijver et al. (2002); all these dependencies can be detected by the distance covariance but not by the classical covariance.

For comparisons of the distance covariance and classical covariance in applications to data, see the examples given by Székely and Rizzo (2009, Section 5.2) and Dueck et al. (2014, Section 5); for extensive numerical experiments and fast algorithms for computing the distance covariance, see Huo and Székely (2016, Section 5). We also refer to Sejdinovic et al. (2013), Dueck et al. (2014), Székely and Rizzo (2009), Székely and Rizzo (2014), & Huo and Székely (2016), and Edelmann et al. (2020), representing only a few of the many authors who have given further theoretical results on the distance covariance and distance correlation coefficients; and to Zhou (2012) & Fiedler (2016), and Edelmann et al. (2019) as among the applications to time series analyses. Many applications to data analysis of the distance correlation coefficient and the distance covariance are now available, including: Kong et al. (2012) on data in sociology, Martínez-Gómez et al. (2014) and

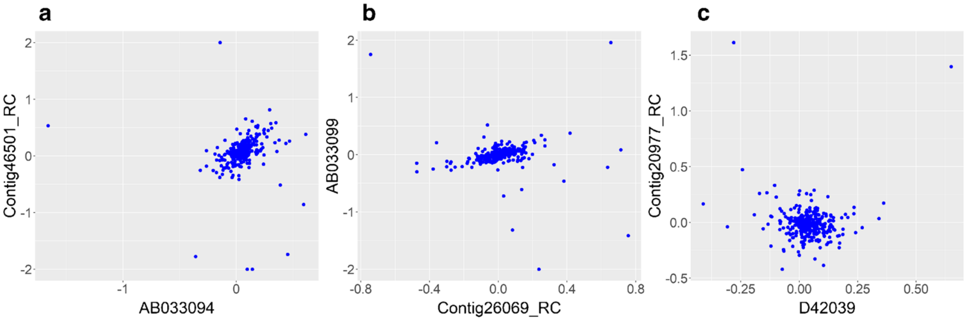

Figure 2: Sub-figures **A-C** represent three scatter-plots of the expression values of genes in a breast cancer data set provided by Van De Vijver et al. (2002) ($n = 295$ samples) on which permutation tests, based on the distance covariance and classical covariance, were applied. The p-values of the distance covariance permutation tests using 100,000 permutations are **A**: 10^{-5} ; **B**: 10^{-5}; **C**: 3.00×10^{-4} . The p-values of permutation tests based on the classical covariance with 100,000 permutations are **A**: 0.079 ; **B**: 0.503; **C**: 0.930

Richards et al. (2014) on astrophysical databases and galaxy clusters, Dueck et al. (2014) on time series analyses of wind vectors at electricity-generating facilities, Richards (2017) on the relationship between the strength of gun control laws and firearm-related homicides, Zhang et al. (2018) for remote sensing applications, and Ohana-Levi et al. (2020) on grapevine transpiration.

The original papers of Székely et al. (2007, 2009) are now widely recognized as seminal and important contributions to measuring dependence between sets of random variables; however, the exposition therein includes some ingenious arguments that may make the material challenging to readers not having an advanced background in mathematical statistics. With the benefit of hindsight, we are able to provide in this article a simpler, albeit mathematically rigorous, introduction to the distance covariance that can be taught even in an undergraduate-level course covering the basic theory of U-statistics. Other than standard U-statistics theory and some well-known properties of characteristic functions, the requirements for our treatment are a knowledge of multidimensional integration and trigonometric inequalities, as covered in a course on undergraduate-level advanced calculus. Consequently, we hope that this treatment will prove to be beneficial to non-mathematical statisticians.

Our presentation introduces the distance covariance as an important alternative to the classical covariance. Moreover, the distance covariance constitutes a particularly interesting example of a U-statistic since it includes both the "non-degenerate" and "first-order degenerate" cases of the asymptotic distribution theory of U-statistics, these corresponding to the situations in which X and Y are dependent, leading to the non-degenerate case, or X and Y are independent, leading to the first-order degenerate case of the asymptotic theory.

Throughout the exposition, $\|\cdot\|$ denotes the Euclidean norm and $\langle\cdot,\cdot\rangle$ the corresponding inner product. Also, we denote by $|\cdot|$ the modulus in $\mathbb{C}$ or the absolute value in $I\!\!R$, and the imaginary unit is $i = \sqrt{-1}$.

2 The Fundamental Integral of Distance Covariance Theory

Following Székely et al. (2007), we first establish a closed-form expression for an integral that plays a central work in this article, leading to the equivalence of two crucial expressions for the distance covariance. The first expression displays the distance covariance as an integrated distance between the joint characteristic function of (X, Y) and the product of the marginal characteristic functions of X and Y; we will deduce from this expression that the distance covariance equals zero if and only if X and Y are independent. The second expression allows us to derive consistent distance covariance estimators that are expressible as polynomials in the distances between random samples.

Since the ability to characterize independence and the existence of easily computable estimators are arguably the most important properties of the distance covariance, we will refer to this integral as the *fundamental integral of distance covariance.*

Lemma 2.1. *For $x \in \mathbb{R}^p$,*

$$\int_{\mathbb{R}^p} \frac{1 - \cos\langle t, x\rangle}{\|t\|^{p+1}} \mathrm{d}t = \frac{\pi^{(p+1)/2}}{\Gamma\big((p+1)/2\big)} \|x\|. \tag{2.1}$$

Proof. Since (2.1) is valid for $x = 0$, we need only treat the case in which $x \neq 0$.

Denote by I_p the integral in Eq. 2.1. For $p = 1$, replacing t by t/x yields

$$I_1 = \int_{-\infty}^{\infty} \frac{1 - \cos tx}{t^2} \mathrm{d}t = |x| \int_{-\infty}^{\infty} \frac{1 - \cos t}{t^2} \mathrm{d}t. \tag{2.2}$$

Denoting the latter integral in Eq. 2.2 by c_1, it follows by integration-by-parts that

$$c_1 = 2\int_0^\infty \frac{1-\cos t}{t^2}\mathrm{d}t = 2\int_0^\infty \frac{\sin t}{t}\mathrm{d}t = \pi, \tag{2.3}$$

the last equality being classical in calculus (Spivak 1994, Chapter 19, Problem 43).

For general p, note that I_p is invariant under orthogonal transformations H of x:

$$\begin{aligned}\int_{\mathbb{R}^p} \frac{1-\cos\langle t, Hx\rangle}{\|t\|^{p+1}}\mathrm{d}t &= \int_{\mathbb{R}^p} \frac{1-\cos\langle Ht, Hx\rangle}{\|H\,t\|^{p+1}}\mathrm{d}t \\ &= \int_{\mathbb{R}^p} \frac{1-\cos\langle t, x\rangle}{\|t\|^{p+1}}\mathrm{d}t,\end{aligned}$$

where the first equality follows from the transformation $t \mapsto Ht$, which leaves the Lebesgue measure dt unchanged; and the second equality holds because the norm and the inner product are orthogonally invariant. Therefore, in evaluating I_p we may replace x by $\|x\|(1, 0, \ldots, 0)$; letting $t = (t_1, \ldots, t_p)$, we obtain

$$I_p = \int_{\mathbb{R}^p} \frac{1-\cos(t_1\|x\|)}{\|t\|^{p+1}}\mathrm{d}t = \|x\| \int_{\mathbb{R}^p} \frac{1-\cos t_1}{\|t\|^{p+1}}\mathrm{d}t, \tag{2.4}$$

the last equality obtained by replacing t_j by $t_j/\|x\|$, $j = 1, \ldots, p$.

Denoting by c_p the latter integral in Eq. 2.4, we substitute in that integral $t_j = v_j$, $j = 1, \ldots, p-1$, and $t_p = p^{-1/2}(v_1^2 + \cdots + v_{p-1}^2)^{1/2} v_p$. As the Jacobian of this transformation is $p^{-1/2}(v_1^2 + \cdots + v_{p-1}^2)^{1/2}$, we obtain

$$\begin{aligned}c_p &= p^{-1/2} \int_{\mathbb{R}^{p-1}} \frac{1-\cos v_1}{(v_1^2+\cdots+v_{p-1}^2)^{p/2}}\mathrm{d}v_1 \cdots \mathrm{d}v_{p-1} \cdot \int_{-\infty}^\infty \frac{\mathrm{d}v_p}{(1+p^{-1}v_p^2)^{(p+1)/2}} \\ &= p^{-1/2} c_{p-1} \int_{-\infty}^\infty \frac{\mathrm{d}v_p}{(1+p^{-1}v_p^2)^{(p+1)/2}}. \end{aligned} \tag{2.5}$$

As the remaining integral in Eq. 2.5 is the familiar normalizing constant of the Student's t-distribution on p degrees-of-freedom, we obtain

$$c_p = \frac{\pi^{1/2}\Gamma(p/2)}{\Gamma\big((p+1)/2\big)}\, c_{p-1}.$$

Starting with $c_1 = \pi$, we solve this recursive equation for c_p, obtaining (2.1).

3 Two Representations for the Distance Covariance

We now introduce the representations of the distance covariance mentioned above. Following Székely et al. (2007), we define the distance covariance through its characteristic function representation. For jointly distributed random vectors $X \in I\!R^p$ and $Y \in I\!R^q$, let $\phi_{X,Y}(s,t) = \mathbb{E}e^{i\langle s,X\rangle + i\langle t,Y\rangle}$ be the joint characteristic function of (X,Y) and $\phi_X(s) = \phi_{X,Y}(s,0)$ and $\phi_Y(t) = \phi_{X,Y}(0,t)$ be the corresponding marginal characteristic functions.

Definition 3.1. *The distance covariance $\mathcal{V}(X,Y)$ between X and Y is defined as the nonnegative square-root of*

$$\mathcal{V}^2(X,Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} \frac{|\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2}{\|s\|^{p+1}\,\|t\|^{p+1}} \,\mathrm{d}s\,\mathrm{d}t, \tag{3.1}$$

where c_p is the normalizing constant in Eq. 2.1.

As the integrand in Eq. 3.1 is nonnegative, it follows that $\mathcal{V}^2(X,Y) \geq 0$. Further, we will show in Corollary 3.4 that $\mathcal{V}^2(X,Y) < \infty$ whenever X and Y have finite first moments.

An advantage of the representation (3.1) is that it directly implies one of the most important properties of the distance covariance, viz., the characterization of independence.

Theorem 3.2. *For all X and Y, $\mathcal{V}^2(X,Y) = 0$ if and only if X and Y are independent.*

Proof. If X and Y are independent then $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$ for all s and t; hence $\mathcal{V}^2(X,Y) = 0$.

Conversely, if X and Y are not independent then the functions $\phi_{X,Y}(s,t)$ and $\phi_X(s)\phi_Y(t)$ are not identical (Van der Vaart 2000, Lemma 2.15). Since characteristic functions are continuous then there exists an open set $\mathcal{A} \subseteq I\!R^p \times I\!R^q$ such that $|\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2 > 0$ for all $(s,t) \in \mathcal{A}$. Hence, by Eq. 3.1, $\mathcal{V}^2(X,Y) > 0$.

For the purpose of deriving estimators for $\mathcal{V}^2(X,Y)$, we now apply Lemma 2.1 to obtain a second representation of the distance covariance.

Theorem 3.3. *Suppose that $(X_1,Y_1),\ldots,(X_4,Y_4)$ are independent, identically distributed (i.i.d.) copies of (X,Y). Then*

$$\mathcal{V}^2(X,Y) = \mathbb{E}\Big[\|X_1 - X_2\| \cdot \|Y_1 - Y_2\| - 2\|X_1 - X_2\| \cdot \|Y_1 - Y_3\| + \|X_1 - X_2\| \cdot \|Y_3 - Y_4\|\Big]. \tag{3.2}$$

Proof. First, we observe that the numerator in the integrand in Eq. 3.1 equals

$$\begin{aligned}&|\phi_{X,Y}(s,t)-\phi_X(s)\phi_Y(t)|^2\\ &=(\phi_{X,Y}(s,t)-\phi_X(s)\phi_Y(t))\,\overline{(\phi_{X,Y}(s,t)-\phi_X(s)\phi_Y(t))}\\ &=\mathbb{E}\big[e^{i\langle s,X_1-X_2\rangle+i\langle t,Y_1-Y_2\rangle}-2\,e^{i\langle s,X_1-X_2\rangle+i\langle t,Y_1-Y_3\rangle}+\,e^{i\langle s,X_1-X_2\rangle+i\langle t,Y_3-Y_4\rangle}\big].\end{aligned}$$

Since the latter expression is real, any term of the form e^{iz}, $z\in\mathbb{R}$, can be replaced by $\cos z$. Hence, by Eq. 3.1,

$$c_pc_q\mathcal{V}^2(X,Y)=\int_{\mathbb{R}^p}\int_{\mathbb{R}^q}\frac{A_{12}(s,t)-2\,A_{13}(s,t)+A_{34}(s,t)}{\|s\|^{p+1}\|t\|^{q+1}}\mathrm{d}s\mathrm{d}t \tag{3.3}$$

where, for each (j,k),

$$A_{jk}(s,t)=\mathbb{E}\cos\big(\langle s,X_1-X_2\rangle+\langle t,Y_j-Y_k\rangle\big). \tag{3.4}$$

Replacing t by $-t$ in Eq. 3.3, we also obtain

$$c_pc_q\mathcal{V}^2(X,Y)=\int_{\mathbb{R}^p}\int_{\mathbb{R}^q}\frac{A_{12}(s,-t)-2\,A_{13}(s,-t)+A_{34}(s,-t)}{\|s\|^{p+1}\|t\|^{q+1}}\mathrm{d}s\mathrm{d}t, \tag{3.5}$$

and by adding (3.3) and (3.5), we find that

$$c_pc_q\mathcal{V}^2(X,Y)=\int_{\mathbb{R}^p}\int_{\mathbb{R}^q}\frac{B_{12}(s,t)-2\,B_{13}(s,t)+B_{34}(s,t)}{\|s\|^{p+1}\|t\|^{q+1}}\mathrm{d}s\mathrm{d}t$$

where for each (j,k),

$$B_{jk}(s,t)=\frac{1}{2}\big(A_{jk}(s,t)+A_{jk}(s,-t)\big). \tag{3.6}$$

On applying to Eqs. 3.4 and 3.6 the trigonometric identity,

$$\cos(x+y)+\cos(x-y)=2\,\cos x\cos y,$$

we deduce that

$$B_{jk}(s,t)=\mathbb{E}\big[\cos\langle s,X_1-X_2\rangle\,\cos\langle t,Y_j-Y_k\rangle\big]. \tag{3.7}$$

For $j, k, \in \{1,2,3,4\}$, we apply to Eq. 3.7 the elementary identity,

$$\begin{aligned}\cos\langle s, X_1 - X_2\rangle \cos\langle t, Y_j - Y_k\rangle &= \big(1-\cos\langle s, X_1 - X_2\rangle\big)\big(1-\cos\langle t, Y_j - Y_k\rangle\big)\\ &\quad -1+\cos\langle s, X_1 - X_2\rangle + \cos\langle t, Y_j - Y_k\rangle; \quad (3.8)\end{aligned}$$

then we obtain

$$\begin{aligned}&c_p c_q \mathcal{V}^2(X,Y)\\ &\quad = \int_{\mathbb{R}^p}\int_{\mathbb{R}^q} \Big(\mathbb{E}\big[\big(1-\cos\langle s, X_1 - X_2\rangle\big)\big(1-\cos\langle t, Y_1 - Y_2\rangle\big)\big]\\ &\qquad -2\,\mathbb{E}\big[\big(1-\cos\langle s, X_1 - X_2\rangle\big)\big(1-\cos\langle t, Y_1 - Y_3\rangle\big)\big]\\ &\qquad +\mathbb{E}\big[\big(1-\cos\langle s, X_1 - X_2\rangle\big)\big(1-\cos\langle t, Y_3 - Y_4\rangle\big)\big]\Big)\frac{\mathrm{d}s\mathrm{d}t}{\|s\|^{p+1}\|t\|^{q+1}},\end{aligned}$$

which is obtained by decomposing all summands on the right-hand side using Eq. 3.8 and observing that all terms which are not of the form $\mathbb{E}[\cos\langle s, X_i - X_j\rangle \cos\langle t, Y_l - Y_k\rangle]$ cancel each other. By applying the Fubini-Tonelli Theorem and the linearity of expectation and integration, we obtain

$$\begin{aligned}&c_p c_q \mathcal{V}^2(X,Y)\\ &\quad = \mathbb{E}\int_{\mathbb{R}^p}\int_{\mathbb{R}^q} \Big[\big(1-\cos\langle s, X_1 - X_2\rangle\big)\big(1-\cos\langle t, Y_1 - Y_2\rangle\big)\\ &\qquad -2\big(1-\cos\langle s, X_1 - X_2\rangle\big)\big(1-\cos\langle t, Y_1 - Y_3\rangle\big)\\ &\qquad +\big(1-\cos\langle s, X_1 - X_2\rangle\big)\big(1-\cos\langle t, Y_3 - Y_4\rangle\big)\Big]\frac{\mathrm{d}s\mathrm{d}t}{\|s\|^{p+1}\|t\|^{q+1}}.\end{aligned}$$

The proof is completed by applying Lemma 2.1 to calculate these three integrals.

Before establishing estimators for $\mathcal{V}^2(X,Y)$, we remark briefly on the assumptions necessary for the existence of the distance covariance.

Corollary 3.4. *Suppose that* $\mathbb{E}\|X\| < \infty$ *and* $\mathbb{E}\|Y\| < \infty$. *Then* $\mathcal{V}^2(X,Y) < \infty$.

Proof. From the representation (3.2), we directly obtain the alternative representation

$$\begin{aligned}\mathcal{V}^2(X,Y) = \mathbb{E}\Big[&\|X_1 - X_2\|\,\|Y_1 - Y_2\| - \|X_1 - X_2\|\,\|Y_1 - Y_3\|\\ &- \|X_1 - X_2\|\,\|Y_2 - Y_3\| + \|X_1 - X_2\|\,\|Y_3 - Y_4\|\Big]. \quad (3.9)\end{aligned}$$

Applying the triangle inequality yields

$$\|X_1 - X_2\| \, \|Y_1 - Y_2\| - \|X_1 - X_2\| \, \|Y_1 - Y_3\| - \|X_1 - X_2\| \, \|Y_2 - Y_3\| \leq 0,$$

and hence

$$\begin{aligned} 0 \leq \mathcal{V}^2(X,Y) \qquad & \leq \mathbb{E}\|X_1 - X_2\| \, \|Y_3 - Y_4\| \\ & = \mathbb{E}\|X_1 - X_2\|\mathbb{E}\|Y_3 - Y_4\| \leq 4\, E\|X\|\mathbb{E}\|Y\|, \end{aligned}$$

where the last inequality follows again by the triangle inequality.

4 Asymptotic Theory for Estimating the Distance Covariance

Using the representation of the distance covariance given in Eq. 3.2, it is straightforward to derive a U-statistic estimator for $\mathcal{V}^2(X)$. Specifically, we define the symmetric kernel function

$$\begin{aligned} & h\big((X_1, Y_1), \ldots, (X_4, Y_4)\big) \\ = & \frac{1}{24} \sum \big(\|X_i - X_j\| \, \|Y_i - Y_j\| - 2\, \|X_i - X_j\| \, \|Y_i - Y_k\| + \|X_i - X_j\| \, \|Y_k - Y_l\|\big), \end{aligned} \tag{4.1}$$

where the sum is over all $i, j, k, l \in \{1, 2, 3, 4\}$ such that i, j, k, and l are distinct.

It follows from the representation (3.2) that each of the 24 summands in Eq. 4.1 has expectation $\mathcal{V}^2(X, Y)$. Therefore,

$$\mathbb{E}h\big((X_1, Y_1), \ldots, (X_4, Y_4)\big) = \mathcal{V}^2(X, Y).$$

Letting $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from (X, Y), we find that an unbiased estimator of $\mathcal{V}^2(X, Y)$ is

$$\widehat{\Omega} = \binom{n}{4}^{-1} \sum_{1 \leq i < j < k < l \leq n} h\big((X_i, Y_i), (X_j, Y_j), (X_k, Y_k), (X_l, Y_l)\big). \tag{4.2}$$

We can now derive the consistency and asymptotic distribution of this estimator using standard U-statistic theory (Lee, 2019). For this purpose, let us define

$$h_1(x, y) = \mathbb{E}\big[h\big((x, y), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4)\big)\big].$$

and

$$h_2((x_1, y_1), (x_2, y_2)) = \mathbb{E}\big[h\big((x_1, y_1), (x_2, y_2), (X_3, Y_3), (X_4, Y_4)\big)\big].$$

The preceding formulas and a classical result on U-statistics (Hoeffding 1948, Theorem 7.1) leads immediately to a proof of the following result.

Theorem 4.1. *Suppose that* $0 < Var(h_1(X,Y)) < \infty$. *Then* $\sqrt{n}\big(\widehat{\Omega} - \mathcal{V}^2(X,Y)\big) \xrightarrow{P} Z$ *as* $n \to \infty$, *where* $Z \sim \mathcal{N}\big(0, 16\,Var(h_1(X,Y)\big)$.

Except for pathological examples, Theorem 4.1 provides the asymptotic distribution of $\mathcal{V}^2(X,Y)$ if X and Y are dependent. For the crucial case of independent X and Y, however, the asymptotic distribution of $\sqrt{n}(\widehat{\Omega} - \mathcal{V}(X,Y)^2)$ is degenerate; in this case, the asymptotic distribution can be derived using results on first-order degenerate U-statistics (Lee 2019, Section 3.2.2).

Lemma 4.2. *Let* X *and* Y *be independent, and* (X_1, Y_1) *and* (X_2, Y_2) *be i.i.d. copies of* (X,Y). *Then* $h_1(x,y) \equiv 0$ *and* $Var\big(h_2((X_1,Y_1),(X_2,Y_2))\big) = \mathcal{V}^2(X,X)\,\mathcal{V}^2(Y,Y)/36$.

The proof follows by elementary, but lengthy, transformations and may be left as an exercise to students. A complete proof is provided by Huang and Huo (2017), Appendices B.6 and B.7.

Finally, the following result follows directly from Lemma 4.2 and classical results on the distributions of first-order degenerate U-statistics (Lee 2019, Section 3.2.2).

Theorem 4.3. *Let* X *and* Y *be independent, with* $\mathbb{E}(\|X\|) < \infty$ *and* $\mathbb{E}(\|Y\|) < \infty$. *Then,*

$$n\left(\widehat{\Omega} - \mathcal{V}^2(X,Y)\right) \xrightarrow{\mathcal{D}} 6 \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1), \tag{4.3}$$

as $n \to \infty$, *where* $Z_1, Z_2, \ldots$ *are i.i.d. standard normal random variables and* $\lambda_1, \lambda_2, \ldots$ *are the eigenvalues of the integral equation*

$$\mathbb{E}\big[h_2\big((x_1,y_1),(X_2,Y_2)\big)\, f(X_2,Y_2)\big] = \lambda f(x_1,y_1).$$

5 Concluding Remarks

In this article, we have derived under minimal technical requirements the most important statistical properties of the distance covariance. From this starting point, there are several additional interesting topics that can be explored, e.g., as instructional assignments:

(i) The estimator Eq. 4.2 is $O(n^4)$ and is computationally inefficient. A straightforward combinatorial computation shows that an $O(n^2)$ estimator of $\mathcal{V}$ is given by

$$\begin{aligned}\widetilde{\Omega} = \frac{1}{n\,(n-3)}\Bigg[& \sum_{i,j=1}^{n} \|X_i - X_j\| \|Y_i - Y_j\| \\ & + \frac{1}{(n-1)\,(n-2)} \sum_{i,j=1}^{n} \|X_i - X_j\| \cdot \sum_{i,j=1}^{n} \|Y_i - Y_j\| \\ & - \frac{2}{(n-2)} \sum_{i,j,k=1}^{n} \|X_i - X_j\| \|Y_i - Y_k\| \Bigg]; \end{aligned} \tag{5.1}$$

see Huo and Székely (2016).

(ii) We remark that although no assumption was provided in Theorem 4.1 to ensure that $\mathrm{Var}(h_1(X,Y)) < \infty$, it can be shown that this condition holds whenever X and Y have finite second moments; see Edelmann et al. (2020).

(iii) Important contributions of Székely et al. (2007) and (Székely and Rizzo, 2009) are based on the *distance correlation coefficient*, which is defined as the nonnegative square-root of

$$\mathcal{V}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}}.$$

Numerous properties of $\mathcal{V}^2(X,Y)$ (see, e.g., Székely et al. 2007, Theorem 3) may be derived using the methods that we have presented here.

We also remark on the fundamental integral, Eq. 2.1, that underpins the entire distance covariance and distance correlation theory. As noted by Dueck et al. (2015), the fundamental integral and variants of it have appeared in functional analysis (Gelfand and Shilov 1964, pp. 192–195), in Fourier analysis (Stein 1970, pp. 140 and 263), and in the theory of fractional Brownian motion on generalized random fields (Chilès and Delfiner P. 2012, p. 266; Reed et al. 1995).

The fundamental integral also extends further. For $m \in \mathbb{N}$ and $v \in \mathbb{R}$, define

$$\cos_m(v) := \sum_{j=0}^{m-1} (-1)^j \frac{v^{2j}}{(2j)!}, \tag{5.2}$$

the truncated Maclaurin expansion of the cosine function. Dueck et al. (2015) proved that for $\alpha \in \mathbb{C}$,

$$\int_{\mathbb{R}^d} \frac{\cos_m(\langle t, x\rangle) - \cos(\langle t, x\rangle)}{\|t\|^{d+\alpha}} \, \mathrm{d}t = \frac{2\pi^{p/2}\,\Gamma(1-\alpha/2)}{\alpha\, 2^\alpha\, \Gamma\big((p+\alpha)/2\big)} \, \|x\|^\alpha, \tag{5.3}$$

with absolute convergence if and only if $2(m-1) < \Re(\alpha) < 2m$. For $m = 1$ and $\alpha = 1$, Eq. 5.3 reduces to Eq. 2.1. Further, for $m = 1$ and $0 < \alpha < 2$, the integral (5.3) provides the Lévy-Khintchine representation of the negative definite function $\|x\|^\alpha$, thereby linking the fundamental integral to the probability theory of the stable distributions.

In conclusion, the statistical analysis of data through distance covariance and distance correlation theory, by means of the fundamental integral, is seen to be linked closely to many areas of the mathematical sciences.

Acknowledgments. D. Edelmann gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft (Grant number 417754611). The authors are grateful to the editor of the special issue and to a reviewer for comments on the manuscript.

References

CHILÈS, J. P. and DELFINER P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, 2nd edn. Wiley, New York.

DUECK, J., EDELMANN, D., GNEITING, T. and RICHARDS, D. (2014). The affinely invariant distance correlation. *Bernoulli*, **20**, 2305–2330.

DUECK, J., EDELMANN, D. and RICHARDS, D. (2015). A generalization of an integral arising in the theory of distance correlation. *Statist. Probab. Lett.*, **97**, 116–119.

EDELMANN, D., FOKIANOS, K. and PITSILLOU, M. (2019). An updated literature review of distance correlation and its applications to time series. *Internat. Statist. Rev.*, **87**, 237–262.

EDELMANN, D., RICHARDS, D. and VOGEL, D. (2020). The distance standard deviation. *Ann. Statist.*, **48**, 3395–3416.

FIEDLER, J. (2016). *Distances, Gegenbauer Expansions, Curls, and Dimples: On Dependence Measures for Random Fields.* Ph.D. dissertation, Heidelberg University.

GELFAND, I. M. and SHILOV, G. E. (1964). *Generalized Functions*, Volume 1. Academic Press, New York.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, **19**, 293–325.

HUANG, C. and HUO, X. (2017). A statistically and numerically efficient independence test based on random projections and distance covariance. Preprint, arXiv:1701.06054.

HUO, X. and SZÉKELY, G. J. (2016). Fast computing for distance covariance. *Technometrics*, **58**, 435–447.

KONG, J., KLEIN, B. E., KLEIN, R., LEE, K. E. and WAHBA, G. (2012). Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 20352–20357.

LEE, A. J. (2019). *U-Statistics: Theory and Practice.* CRC Press, Boca Raton.

MARTÍNEZ-GÓMEZ, E., RICHARDS, M. T. and RICHARDS, D. ST. P. (2014). Distance correlation methods for discovering associations in large astrophysical databases. *Astrophys. J.* **781**(v), 11.

OHANA-LEVI, N., MUNITZ, S., BEN-GAL, A., SCHWARTZ, A., PEETERS, A. and NETZER, Y. (2020). Multiseasonal grapevine water consumption - Drivers and forecasting. *Agric. For. Meteorol.*, **280**, 107796 (12 pp.)

REED, I. S., LEE, P. C. and TRUONG, T. K. (1995). Spectral representation of fractional Brownian motion in n dimensions and its properties. *IEEE Trans. Inform. Theory*, **41**, 1439–1451.

RICHARDS, D. ST. P. (2017). Distance correlation: A new tool for detecting association and measuring correlation between data sets. *Notices Amer. Math. Soc.*, **64**, 16–18.

RICHARDS, M. T., RICHARDS, D. ST. P. and MARTÍNEZ-GÓMEZ, E. (2014). Interpreting the distance correlation results for the COMBO-17 survey. *Astrophys. J.*, **784**, L34 (5 pp.)

SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, **41**, 2263–2291.

SPIVAK, M. (1994). *Calculus*, 3rd edn. Publish or Perish, Houston.

STEIN, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions.* Princeton University Press, Princeton.

SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Statist.*, **3**, 1236–1265.

SZÉKELY, G. J. and RIZZO, M. L (2014). Partial distance correlation with methods for dissimilarities. *Ann. Statist.*, **42**, 2382–2412.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing independence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press, New York.

VAN DE VIJVER, M. J., HE, Y. D., VAN'T VEER, L. J. and ET AL. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

ZHANG, X., KANO, M. and LI, Y. (2018). Quality-relevant independent component regression model for virtual sensing application. *Computers & Chem. Eng.*, **115**, 141–149.

ZHOU, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *J. Time Series Analysis*, **33**, 438–457.

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

Dominic EdelmannTobias Terzer
Division of Biostatistics, German
Cancer Research Center, Im
Neuenheimer Feld 280, 69120
Heidelberg, Germany
E-mail: dominic.edelmann@dkfz-heidelberg.de
t.terzer@dkfz-heidelberg.de

Donald Richards
Department of Statistics,
Pennsylvania State University,
University Park, PA 16802, USA
E-mail: richards@stat.psu.edu

Paper received: 16 October 2020; accepted 2 January 2021.