Summary and discussion of: "Brownian Distance Covariance"

Statistics Journal Club, 36-825

Benjamin Cowley and Giuseppe Vinci

October 27, 2014

1 Summary

1.1 Introduction

In the center of statistical analysis, one of the main tasks of a statistician is usually to find relationships between random quantities, i.e. to find *dependence*. For instance, discovering a *linear* relationship between two random variables X, Y defined by $Y = \beta_0 + \beta_1 X + \epsilon$, ϵ random error, could help to "predict Y by observing X". The Pearson's product moment correlation coefficient (PC) tells us how good the relationship between two random variables can be approximated by a linear function (see Figure 1 (A)). Similarly the Spearman's rank correlation coefficient (SC) is used to detect the presence of a *monotonic* relationship between two random variables (see Figure 1 (B)). Thus, extreme values of PC or SC ($\approx -1, 1$) highlight the presence of strong linear or at least monotonic relationships.

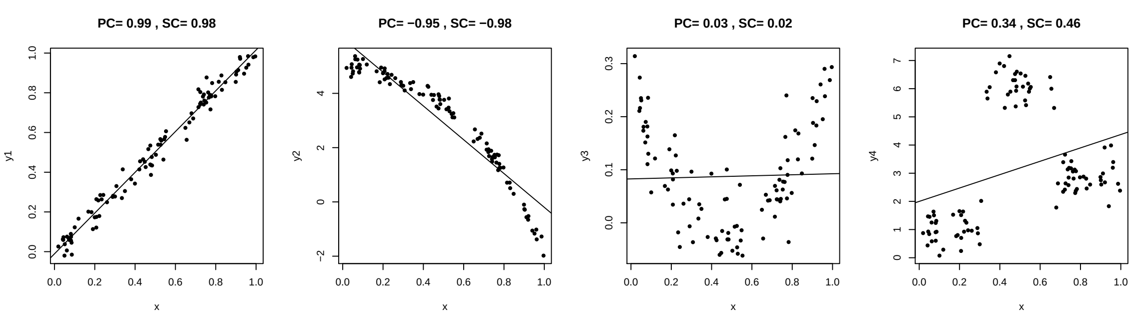

Figure 1: Examples of dependence between two random variables. A: linear relationship. B: monotonic relationship. C: U-shaped relationship. D: clustering relationship. The straight lines are the simple linear regression estimates.

Yet, consider Figure 1 (C)-(D). We can clearly *see* that X, Y are *not* independent, even if there is not any linear or monotonic relationship. In particular in Figure 1 (C) both the PC and SC are approximately zero. This is in fact an example where $PC = 0 \not\Rightarrow X \perp Y$ (independence), while it is always true that $X \perp Y \Rightarrow PC = 0$. Therefore, it seems clear that it is not correct in general to *test independence* of two random variables X, Y by testing $PC = 0$ or $SC = 0$. Furthermore, in the case when X, Y are vectors of possibly different dimensions, one would need to test for independence on all possible pairwise correlations, which can be infeasible. A solution could be to study the *joint distribution* of X, Y, or some transformations of it. In particular, by definition $X \perp Y$ if and only if $F_{XY} = \mathcal{F}_X F_Y$, where F_{XY} denotes the joint cumulative distribution function of vector (X, Y), and F_X, F_Y are the marginal cdf's of X, Y, respectively. Alternatively, we also have that $X \perp Y$ if and only if $\mathcal{F}_{XY} = \mathcal{F}_X \mathcal{F}_Y$, where $\mathcal{F}_{XY}$ denotes the joint characteristic function of vector (X, Y) and $\mathcal{F}_X, \mathcal{F}_Y$ are the marginal cf's of X, Y, respectively. Therefore, it seems to be intuitive that some *distances* $d(F_{XY}, F_X F_Y)$ or $d(\mathcal{F}_{XY}, \mathcal{F}_X \mathcal{F}_Y)$ could be much more appropriate to detect *any kind of relationship* between two random quantities X, Y. In Renyi (1959) we find the definition of Maximal Correlation

$$\mathrm{mCor} = \max_{f,g} \mathrm{Cor}(f(X), g(Y)), \quad \mathrm{s.t.Var}(f(X)), \mathrm{Var}(g(Y)) > 0 \tag{1}$$

where Cor denotes the Pearson correlation, and the maximum is taken over all the possible functions f, g. We have that mCor $= 0$ if and only if $X \perp Y$. An interesting result is given when we choose $f, g \in \mathcal{A} = \{h_t : \mathbb{R} \to \mathbb{C} : h_t(w) = e^{itw}, t \in \mathbb{R}\}$. We have in fact that $\mathrm{Cov}(e^{isX}, e^{itY}) = \mathbb{E}[e^{i(sX+tY)}] - \mathbb{E}[e^{isX}]\mathbb{E}[e^{itY}] = \mathcal{F}_{XY}(s,t) - \mathcal{F}_X(s)\mathcal{F}_Y(t)$, such that $\max_{f,g \in \mathcal{A}} \mathrm{Corr}(f(X), g(Y)) = 0$ if and only if $X \perp Y$, suggesting that focusing on characteristic functions is a reasonable choice.

In the past 60 years, we can find some attempts of addressing the idea of testing independence of two random variable X, Y based on estimates of the distribution functions or characteristic functions. In Hoeffding (1948), and Blum, Kiefer & Rosenblatt (1961), we can find tests of independence based on the following distance of the empirical cdf's:

$$R_n = \int\int [F_n(x,y) - F_n(x)F_n(y)]^2 dF_n(x,y) \tag{2}$$

and in Rosenblatt (1975) a test of independence was based on the distance between kernel density:

$$B_n = \int\int [\hat{p}_h(x,y) - \hat{p}_{h_X}(x)\hat{p}_{h_Y}(y)]^2 a(x,y) dx dy \tag{3}$$

where $a(,)$ is a weight function. We have that R_n is H_0-distribution free, while B_n is not. Moreover, tests based on B_n resulted to be less powerful than that based on R_n.

An important contribution was given in Feuerverger (1993). The author proposed a test based on the following distance between empirical characteristic functions:

$$T_n = \int\int [\mathcal{F}_n(s,t) - \mathcal{F}_n(s)\mathcal{F}_n(t)]^2 w(s,t) ds dt \tag{4}$$

and $nT_n \overset{D}{\to} \sum \lambda_j Z_j^2$, where Z_j's are iid $N(0,1)$, but estimating λ_i's resulted to be not so easy, such that approximations via simulations were preferable.

However, all these tests were limited to the bi-variate case, i.e. $X \in \mathbb{R}, Y \in \mathbb{R}$. An important contribution was recently given by Székely, Rizzo and Bakirov (2007). They proposed a quantity called *distance covariance*, which measures the degree of any kind of relationship between two vectors $\mathbf{X} \in \mathbb{R}^p, \mathbf{Y} \in \mathbb{R}^q$ of arbitrary dimensions p, q. Moreover in Székely-Rizzo (2009), the authors showed that an equivalence existed between a particular reasonable definition of distance covariance and the Brownian covariance of two random vectors of arbitrary dimensions.

In the rest of this paper we present the content of Székely and Rizzo (2009) and some simulations. We start by defining some notation. Then we define the distance covariance and distance correlation, their relationship with the Brownian covariance, some properties of distance covariance and hypothesis testing. Finally we present the results of some simulations and discussions.

1.2 Notation

- $X \perp Y$ means "X, Y are independent"
- $\|x\|$ is the L_2 norm of vector $\mathbf{x} \in \mathbb{R}^d$
- $\langle \mathbf{x}, \mathbf{y} \rangle$ is the inner product of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$
- $F_{\mathbf{W}}(\mathbf{w})$ is the cumulative distribution function (cdf) of random vector $\mathbf{W} \in \mathbb{R}^d$
- $\mathcal{F}_{\mathbf{W}}(\mathbf{t}) = \mathbb{E}\left[e^{i\langle \mathbf{t}, \mathbf{W} \rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle \mathbf{t}, \mathbf{W} \rangle} dF_{\mathbf{W}}(\mathbf{w})$ is the characteristic function (cf) of random vector $\mathbf{W} \in \mathbb{R}^d$, where $\mathbf{t}, \mathbf{W} \in \mathbb{R}^d$, and $i = \sqrt{-1}$
- $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ and $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$
- $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$, i.e. the Pearson's correlation coefficient

1.3 Distance Covariance

In this section, we define and discuss the distance covariance dependency measure.

We would like to see if there are any dependencies between X and Y. We have that $X \perp Y$ if and only if $f_{XY} = f_X f_Y$, where f_{xy}, f_X, f_X are the the probability density functions of $(X, Y), X, Y$, respectively. This motivates trying to find the distance between f_{XY} and $f_X f_X$. We might do this with a norm: $\|f_{XY} - f_X f_Y\|$. This norm is a dependency measure for X and Y. X and Y are independent if and only if $\|f_{XY} - f_X f_Y\| = 0$.

Unfortunately, the probability density functions (pdfs) do not always have nice properties for the distance covariance. A pdf does not always exist, and it may not be uniformly continuous on the entire support. This motivates the use of characteristic functions. A characteristic function of a column vector of p random variables $\mathbf{X}$ has the following form:

$$\mathcal{F}_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{i\langle \mathbf{t}, \mathbf{X} \rangle}\right] = \int_{\mathbb{R}^p} e^{i\langle \mathbf{t}, \mathbf{X} \rangle} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

where $i = \sqrt{-1}$. This is the inverse Fourier transform of pdf $f_{\mathbf{X}}$, and is guaranteed to exist (the integral is a bounded continuous function over a space whose measure is finite). Also, $\mathcal{F}_{\mathbf{X}}(\mathbf{t})$ is uniformly continuous on the entire support of $\mathbf{X}$. Finally, the most important property is that the characteristic function shares the independence property of the pdf, namely $\mathcal{F}_{\mathbf{XY}} = \mathcal{F}_{\mathbf{X}}\mathcal{F}_{\mathbf{Y}}$ if and only if $\mathbf{X} \perp \mathbf{Y}$. Thus, the characteristic function is a suitable replacement for the pdf when considering the distance between two pdfs.

Let $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$, where p does not necessarily need to be equal to q. Our new desired norm is the following:
$\|\mathcal{F}_{\mathbf{XY}} - \mathcal{F}_{\mathbf{X}}\mathcal{F}_{\mathbf{Y}}\|$, where $\|\mathcal{F}_{\mathbf{XY}} - \mathcal{F}_{\mathbf{X}}\mathcal{F}_{\mathbf{Y}}\| = 0 \iff \mathbf{X} \perp \mathbf{Y}$.

With this definition, the distance covariance, $\mathcal{V}(\mathbf{X}, \mathbf{Y})$ is defined as follows:
$\mathcal{V}^2(\mathbf{X}, \mathbf{Y}; w) = \|\mathcal{F}_{\mathbf{XY}}(t,s) - \mathcal{F}_{\mathbf{X}}(t)\mathcal{F}_{\mathbf{Y}}(s)\|_w^2$, where w denotes a specific weight function.
$\mathcal{V}^2(\mathbf{X}, \mathbf{Y}; w) = \int\limits_{\mathbb{R}^{p+q}} |\mathcal{F}_{\mathbf{XY}}(t,s) - \mathcal{F}_{\mathbf{X}}(t)\mathcal{F}_{\mathbf{Y}}(s)|^2 w(t,s)dtds$
We evoke a weight function $w(t,s)$ to enforce desired properties that results in the distance covariance measure. Different choices of $w(t,s)$ will lead to different types of covariances: Not every $w(t,s)$ leads to a dependence measure. Why would we evoke a weight function at all? One reason is that we will be taking the difference of empirical characteristic functions, which may have large noise in the higher frequencies that could negatively affect our estimate of the norm. A weight function w that gives large weights to small norms of t and s favor comparing the characteristic functions at lower frequencies (t and s correspond to frequencies in the Fourier domain).

We can also define a standardized version of the distance covariance, termed the distance correlation:

$$\mathcal{R}_w = \frac{\mathcal{V}(\mathbf{X}, \mathbf{Y}; w)}{\sqrt{\mathcal{V}(\mathbf{X}; w)\mathcal{V}(\mathbf{Y}; w)}}.$$

$\mathcal{R}_w$ is an unsigned correlation dependent on $w(t,s)$. To choose $w(t,s)$, we identify two desired properties of $\mathcal{R}_w$:
1. $\mathcal{R}_w \geq 0$; $\mathcal{R}_w = 0 \iff \mathbf{X} \perp \mathbf{Y}$
2. $\mathcal{R}_w$ is scale invariant (i.e., scaling $\mathbf{X}$ and $\mathbf{Y}$ by $\epsilon \in \mathbb{R}$ would not change $\mathcal{R}_w$).

A first attempt at choosing $w(t,s)$ was to consider integrable functions. However, this resulted in the problem that $\mathcal{R}_w$ could be arbitrarily close to zero, even if $\mathbf{X}$ and $\mathbf{Y}$ are dependent. Since we would like to test if $\mathbf{X}$ and $\mathbf{Y}$ are independent by seeing how close the distance correlation is to zero (i.e., the null distribution would have mass around zero), having a test statistic that can be arbitrarily close to the null distribution even under dependence is not desirable.

To address this issue, [6] used a non-integrable function:

$$w(t,s) = (C_p C_q \|t\|_p^{1+p} \|s\|_q^{1+q})^{-1} \tag{5}$$

where $\|t\|_m = \sqrt{(\sum_{j=1}^m t_j^2)}$ (i.e., the Euclidean norm for $\mathbb{R}^m$), $t \in \mathbb{R}^p$ and $s \in \mathbb{R}^q$, and $C_d = \pi^{\frac{1+d}{2}} / \Gamma\left(\frac{1+d}{2}\right)$. As mentioned above, this weight function will be small for high-frequency components of the characteristic functions (t and s represent frequencies).

1.3.1 Definition of distance covariance

This particular weight function defines the distance covariance, dCov:

$$\mathcal{V}^2(\mathbf{X},\mathbf{Y}) = \frac{1}{C_p C_q} \int\limits_{\mathbb{R}^{p+q}} \frac{|\mathcal{F}_{\mathbf{XY}}(t,s) - \mathcal{F}_{\mathbf{X}}(t)\mathcal{F}_{\mathbf{Y}}(s)|^2}{|t|^{1+p}|s|^{1+q}} dtds$$

Note that $\mathcal{V}(\mathbf{X},\mathbf{Y}) \geq 0$ and $\mathcal{V}(\mathbf{X},\mathbf{Y}) = 0 \iff \mathbf{X} \perp \mathbf{Y}$.

1.3.2 Sample distance covariance

One natural place to estimate the distance covariance is to first estimate the empirical characteristic functions (e.g., $\mathcal{F}_X^n$) and compute the sample distance covariance:

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = \|\mathcal{F}^n_{\mathbf{XY}}(t,s) - \mathcal{F}^n_{\mathbf{X}}(t)\mathcal{F}^n_{\mathbf{Y}}(s)\|^2$$

Evaluating the integral of this norm with the chosen $w(t,s)$ yields:

$$\begin{aligned}\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} |\mathbf{X}_k - \mathbf{X}_l|_p |\mathbf{Y}_k - \mathbf{Y}_l|_q + \frac{1}{n^2}\sum_{k,l=1}^{n} |\mathbf{X}_k - \mathbf{X}_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |\mathbf{Y}_k - \mathbf{Y}_l|_q \\ - 2\frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} |\mathbf{X}_k - \mathbf{X}_l|_p |\mathbf{Y}_k - \mathbf{Y}_m|_q \end{aligned} \quad (6)$$

This motivates how the distance covariance is a linear combination of expected values of norms (as seen through the equivalency of the distance covariance to the Brownian covariance, discussed in later sections):

$$\begin{aligned}\mathcal{V}^2(\mathbf{X},\mathbf{Y}) = E[|\mathbf{X}-\mathbf{X}'||\mathbf{Y}-\mathbf{Y}'|] + E[|\mathbf{X}-\mathbf{X}'|]E[|\mathbf{Y}-\mathbf{Y}'|] \\ -E[|\mathbf{X}-\mathbf{X}'||\mathbf{Y}-\mathbf{Y}''|] - E[|\mathbf{X}-\mathbf{X}''||\mathbf{Y}-\mathbf{Y}'|]\end{aligned}$$

Here, $\mathbf{X},\mathbf{X}',\mathbf{X}''$ are i.i.d (and similarly defined $\mathbf{Y},\mathbf{Y}',\mathbf{Y}''$).

Looking over the form of the sample distance covariance, we can define a simple procedure involving matrices and Euclidean distances between sample data points. Thus, the sample distance covariance is interpretable and computationally efficient, making it a valuable tool in practical settings.

1.3.3 Procedure to compute the sample distance covariance

We first describe the procedure to compute the sample distance covariance, and we then provide simple examples to intuit the measure.

Procedure to compute $\mathcal{V}_n(\mathbf{X},\mathbf{Y})$:

1. Let $(\mathcal{X}, \mathcal{Y}) = \{\mathbf{X}_i, \mathbf{Y}_i : i = 1, \ldots, n\}$ represent the n paired i.i.d random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$.

2. Compute the distance matrices a and b:
   $(a_{kl}) = (\|\mathbf{X}_k - \mathbf{X}_l\|_p)$ and $(b_{kl}) = (\|\mathbf{Y}_k - \mathbf{Y}_l\|_q)$.

3. Compute the means of the rows ($\bar{a}_{k\cdot}$), columns ($\bar{a}_{\cdot l}$), and all elements for the distance matrices ($\bar{a}_{\cdot\cdot}$). These terms correspond to terms in Equation 6.
   $\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}$, $\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}$, $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$. Compute $\bar{b}_{k\cdot}$, $\bar{b}_{\cdot l}$, and $\bar{b}_{\cdot\cdot}$ in a similar manner.

4. Define the re-centered matrices, A and B:
   $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$
   $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$

5. Finally, the sample distance covariance is computed as follows:
   $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}$

To gain intuition for the distance covariance, we ran three simple simulations, including a bi-variate non-linear case, a case of independence between two random vectors, and a case of dependence between two random vectors (see Section 1.7.1). We include heat maps of the various matrices in the procedure, and provide intuition for each step.

1.3.4 Summary of distance covariance

In this section, we provided a mathematical definition to the distance covariance. The sample distance covariance can be computed in an efficient manner, which allows the use of the random permutation test to test for significance (discussed in a later section). The distance covariance also has nice properties, including detecting non-linear interactions between variables and an interpretable procedure to compute the sample distance covariance. The distance covariance also extends the bi-variate case to test for dependencies between random vectors of different dimensions. Thus, the measure tests for dependencies between one variable in $\mathbf{X}$ and multiple variables in $\mathbf{Y}$. Note that dependencies can exist between variables in $\mathbf{X}$ but that $\mathbf{X}$ and $\mathbf{Y}$ can still be independent.

An application where the distance covariance could be used is in neuroscience. Currently, experimenters implant a multi-electrode array into the brain and simultaneously record the activity of a population of neurons in the cerebral cortex. Experiments in the near future will implant *two* multi-electrode arrays in *two* different areas of the brain, and ask if neurons in those two areas are communicating (i.e., do the areas interact?). One way to test for interactions between the two areas is to treat the activity from one array as a vector $\mathbf{X}$ with dimensionality equal to the number of neurons recorded from that array, and likewise define $\mathbf{Y}$ for the other array. Then, we can use the distance covariance measure to test for dependencies between the two populations (i.e., test if $\mathbf{X}$ and $\mathbf{Y}$ are independent).

1.4 Brownian Distance Covariance

We are going to give some definitions useful to introduce the notion of Brownian covariance and finally state the main result of Székely and Rizzo (2009).

Definition 1. *Let* $\mathbf{X} \in \mathbb{R}^p$ *be a random vector and let* $\{U(s) : s \in \mathbb{R}^p\}$ *be a random field. The* U*-centered version of* $\mathbf{X}$ *with respect to* U *is*

$$X_U = U(\mathbf{X}) - \mathbb{E}[U(\mathbf{X})|U] \tag{7}$$

where $X_U \in \mathbb{R}$.

Notice that X_U is random because both $\mathbf{X}$ and U are random. The stochastic transformation $\mathbf{X} \mapsto X_U$ converts random vector $\mathbf{X} \in \mathbb{R}^p$ into a random variable $X_U \in \mathbb{R}$. This is a key step that enable us to study the dependence of two random vectors $\mathbf{X}, \mathbf{Y}$ of arbitrary dimensions by compressing them onto the same univariate space.

Definition 2. *Let* $\mathbf{X} \in \mathbb{R}^p, \mathbf{Y} \in \mathbb{R}^q$ *be two random vectors, and let* $\{U(s) : s \in \mathbb{R}^p\}$, $\{V(t) : t \in \mathbb{R}^q\}$ *be two independent random fields. The covariance of* $\mathbf{X}, \mathbf{Y}$ *with respect to* U, V *is the positive number whose square is:*

$$\mathrm{Cov}^2_{U,V}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[X_U X'_U Y_V Y'_V] \tag{8}$$

where $(\mathbf{X}, \mathbf{Y})$ *and* $(\mathbf{X}', \mathbf{Y}')$ *are independent copies.*

We can see that the resulting dependence measure is determined by the choice of U, V. We can see also that this definition generalize Pearson product moment covariance:

Remark 1. *If* $p = q = 1$ *and* U, V *are identity functions* $\mathrm{id}(t) = t$, *then*

$$\mathrm{Cov}^2_{U,V}(X, Y) = (\mathrm{Cov}(X, Y))^2 \tag{9}$$

since $X_{\mathrm{id}} = X - \mathbb{E}[X]$, $Y_{\mathrm{id}} = Y - \mathbb{E}[Y]$.

In order to detect any kind of relationships between $\mathbf{X}$ and $\mathbf{Y}$ (see Remark 2), and also to get some tractable formulas, a natural choice of (U, V) is (W, W'), which are two independent Wiener fields.

Definition 3. *Wiener field. A Wiener random field on* $\mathbb{R}^d$ *is a random field* $\{W(s) : s \in \mathbb{R}^d\}$ *with the following properties:*

1. *trajectories are continuous a.s.*
2. $\mathrm{Cov}(W(s), W(t)) = \|s\| + \|t\| + \|s - t\|$
3. *increments are independent*

Definition 4. *Brownian covariance. Let* $\{W(s) : s \in \mathbb{R}^p\}$, $\{W'(t) : t \in \mathbb{R}^q\}$ *be two independent Wiener fields. The Brownian covariance of* $\mathbf{X}, \mathbf{Y}$ *is the positive number whose square is:*

$$\mathcal{W}^2(\mathbf{X}, \mathbf{Y}) = \mathrm{Cov}^2_{W,W'}(\mathbf{X}, \mathbf{Y}) \tag{10}$$

Theorem 1. *Let $w(s,t)$ be the weight function defined in expression (5). Then*

$$\mathcal{V}(\mathbf{X},\mathbf{Y};w) \equiv \mathcal{W}(\mathbf{X},\mathbf{Y}) \tag{11}$$

Proof. See proofs of Theorem 7 and Theorem 8 in [6]. □

Corollary 1. *We have that $\mathcal{W}(\mathbf{X},\mathbf{Y})$ has all the properties of $\mathcal{V}(\mathbf{X},\mathbf{Y};w)$.*

Remark 2. *The Brownian covariance of two random vectors $\mathbf{X} \in \mathbb{R}^p, \mathbf{Y} \in \mathbb{R}^q$ is obtained by averaging over all the possible realizations of two independent Wiener fields and with respect to the joint distribution of $(\mathbf{X},\mathbf{Y})$. Since a Wiener field has the property of having continuous trajectories a.s., we can see the Brownian covariance as an approximate average over all the possible continuous functions $f : \mathbb{R}^p \to \mathbb{R}$, $g : \mathbb{R}^q \to \mathbb{R}$ evaluated at random points $(\mathbf{X},\mathbf{Y})$. This intuition, combined with Theorem 1, can help to understand why tests based on the distance covariance $\mathcal{V}(\mathbf{X},\mathbf{Y})$ can detect any kind of relationship between $\mathbf{X}$ and $\mathbf{Y}$. We can see also that averaging with respect to Wiener processes is much easier than maximizing with respect to all the possible functions, as it is instead required to compute the maximal correlation (Renyi 1951).*

1.5 Properties of Distance Covariance

In this section, we briefly discuss some properties of the distance covariance. Proof of these properties can be found in [6].

1. As seen in the distance covariance section (Section 1.3), there is an equivalent definition for computing the sample distance covariance with empirical characteristic functions and the Euclidean norm for distances. This equates the desired norm $(\|\mathcal{F}_{\mathbf{XY}} - \mathcal{F}_{\mathbf{X}}\mathcal{F}_{\mathbf{Y}}\|_w)$ with a computationally-efficient calculation using distance matrices.

2. As desired from any sample measure, the sample distance covariance converges almost surely to the distance covariance (i.e., $\mathcal{V}_n \longrightarrow \mathcal{V}$), and the sample distance correlation converges almost surely to the distance correlation (i.e., $\mathcal{R}_n^2 \longrightarrow \mathcal{R}^2$).

3. The following properties hold for $\mathcal{V}(\mathbf{X},\mathbf{Y}), \mathcal{V}(\mathbf{X})$, and $\mathcal{R}(\mathbf{X},\mathbf{Y})$:

   (a)
   $$0 \leq \mathcal{R}(\mathbf{X},\mathbf{Y}) \leq 1 \text{ and } \mathcal{R} = 0 \text{ if and only if } \mathbf{X} \text{ and } \mathbf{Y} \text{ are independent.}$$

   (b) Let random vectors $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (Y_1, Y_2)$. If $\mathbf{X}$ is independent of $\mathbf{Y}$, then
   $$\mathcal{V}(X_1 + Y_1, X_2 + Y_2) \leq \mathcal{V}(X_1, X_2) + \mathcal{V}(Y_1, Y_2)$$
   Equality holds if X_1, X_2, Y_1, Y_2 are mutually independent.

   (c)
   $$\mathcal{V}(\mathbf{X}) = 0 \Longrightarrow \mathbf{X} = E[\mathbf{X}] \text{ almost surely}$$

(d)
$$\mathcal{V}(\mathbf{a}_1 + b_1 C_1 \mathbf{X}, \mathbf{a}_2 + b_2 C_2 \mathbf{Y}) = \sqrt{(|b_1 b_2|)}\mathcal{V}(\mathbf{X}, \mathbf{Y})$$
for constant vectors $a_1 \in \mathbb{R}^p, a_2 \in \mathbb{R}^q$, scalars $b_1, b_2 \in \mathbb{R}$, and orthonormal matrices $C_1 \in \mathbb{R}^p, C_2 \in \mathbb{R}^q$.

(e) If $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^p$ are independent, then $\mathcal{V}(\mathbf{X}+\mathbf{Y}) \leq \mathcal{V}(\mathbf{X}) + \mathcal{V}(\mathbf{Y})$. Equality holds if and only if one of the random vectors $\mathbf{X}$ or $\mathbf{Y}$ is constant.

4. The following properties hold for $\mathcal{R}_n$ and $\mathcal{V}_n$:

(a)
$$\mathcal{V}_n(\mathbf{X}, \mathbf{Y}) \geq 0$$

(b) $\mathcal{V}_n(\mathbf{X}) = 0$ if and only if every observed sample is identical.

(c)
$$0 \leq \mathcal{R}_n(\mathbf{X}, \mathbf{Y}) \leq 1$$

(d) $\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = 1$ implies that the linear subspaces spanned by $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^p$ respectively are almost surely equal. If we assume subspace equality, then
$$\mathbf{Y} = \mathbf{a} = b\mathbf{X}C$$
for some constant vector $a \in \mathbb{R}^p$, nonzero real number b, and orthogonal matrix C.

5. The following properties involve the test statistic $n\mathcal{V}_n^2$:

(a) If $\mathbf{X} \perp \mathbf{Y}$, then $\frac{n\mathcal{V}_n^2}{T_2} \xrightarrow[n\to\infty]{D} Q$, where $Q = \sum_{j=1}^{\infty} \lambda_j Z_j^2$. $Z_j \sim \mathcal{N}(0,1)$, $\{\lambda_j\}$ are nonnegative constants that depend on the distribution of $(\mathbf{X}, \mathbf{Y})$, and $[Q] = 1$.

(b) If $\mathbf{X} \perp \mathbf{Y}$, then $n\mathcal{V}_n^2 \xrightarrow[n\to\infty]{D} Q_1$, where Q_1 is a nonnegative quadratic form of centered Gaussian random variables and $E[Q_1] = E[|\mathbf{X} - \mathbf{X}'|]E[|\mathbf{Y} - \mathbf{Y}'|]$.

(c) If $\mathbf{X}$ and $\mathbf{Y}$ are dependent, then $\frac{n\mathcal{V}_n^2}{T_2} \xrightarrow[n\to\infty]{P} \infty$ and $n\mathcal{V}_n^2 \xrightarrow[n\to\infty]{P} \infty$.

6. The following properties address the case when (X, Y) has a bi-variate normal distribution, which is an assumption by the Pearson's correlation ρ to test for independence.

(a)
$$\mathcal{R}(\mathbf{X}, \mathbf{Y}) \leq |\rho|$$

(b)
$$\mathcal{R}^2(\mathbf{X}, \mathbf{Y}) = f(\rho) \text{ where } f \text{ has a closed form.}$$

(c)
$$\inf_{\rho \neq 0} \frac{\mathcal{R}(\mathbf{X}, \mathbf{Y})}{|\rho|} = \lim_{\to 0} \frac{\mathcal{R}(\mathbf{X}, \mathbf{Y})}{|\rho|} \approx 0.89$$

1.6 Hypothesis Testing

We would like to test for independence between $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$. Our null hypothesis is that $\mathbf{X}$ and $\mathbf{Y}$ are independent (e.g., $H_0 : \mathcal{F}_{XY} = \mathcal{F}_X \mathcal{F}_Y$). While it turns out that a test statistic can be found with a limiting distribution, this has weak properties. However, because the distance covariance is computationally-efficient to compute, we recommend using the random permutation test to test for independence. This is achieved by, for each run, randomly permuting the samples in the set of data points for $\mathbf{Y}$, $\mathcal{Y}$. This breaks the dependencies between $\mathbf{X}$ and $\mathbf{Y}$. We can then compute the shuffled sample distance covariance $\mathcal{V}_n^*$ for this permuted sample (i.e., under the null hypothesis), and repeat this many times. Then, we estimate the distribution of permuted $\mathcal{V}_n^*$'s, and find the fraction of how many are greater than the actual sample covariance distribution $\mathcal{V}_n$. This yields a p-value to test for independence.

For completeness, we include notes about the limiting distribution for the distance covariance.

First, it can be shown that, if $\mathbf{X}$ and $\mathbf{Y}$ are independent, then

$$\frac{n\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\bar{a}_{..}\bar{b}_{..}} = \frac{n\mathcal{V}_n^2}{T_2} \xrightarrow[n\to\infty]{D} Q$$

where

$$Q \stackrel{D}{=} \sum_{j=1}^{\infty} \lambda_j Z_j^2$$

as defined in the previous section. Since Q has a defined distribution, we can test if $P(Q \geq \chi^2_{q-\alpha}(1)) \leq \alpha)$, and the test statistic $\frac{n\mathcal{V}_n^2}{T_2}$ has a limiting distribution. Thus, we can reject independence if

$$\frac{n\mathcal{V}_n^2}{T_2} \geq \chi^2_{1-\alpha}(1)$$

Note that $\chi^2_{1-\alpha}(1)$ cannot be computed analytically, but "good" approximations exist. However, this test has been shown to be conservative.

If $\mathbf{X}$ and $\mathbf{Y}$ are dependent,

$$\frac{n\mathcal{V}_n^2}{T_2} \xrightarrow[n\to\infty]{P} \infty.$$

Thus, the distance covariance test of independence is statistically consistent against all types of dependence.

1.7 Simulations

1.7.1 Simulation 1: Examples of sample distance covariance

We provide intuition behind the distance covariance with three examples (a non-linear bivariate example, an example where the random vectors are independent, and an example where the data are clustered).

Example 1: Bi-variate non-linear dependency

In this example, $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. The variables were sampled from the following distributions:

$$X \sim \text{Uniform}([-10, 10])$$
$$Y \mid X = X^2 + \epsilon \qquad \text{where} \quad \epsilon \sim \mathcal{N}(0, 25)$$

Thus, X and Y have a strong dependency (Fig. 2). However, notice that this parabolic dependency is not captured by the Pearson's correlation (Fig. 1C). Thus, we investigate if the distance covariance can do a better job at identifying the non-linear dependency between X and Y.

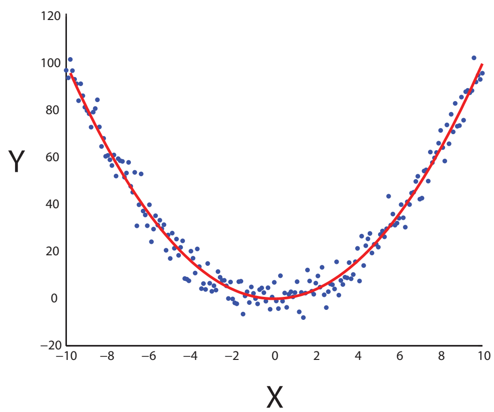

Figure 2: Non-linear relationship between two random variables. Each blue point represents one sample observation, and the red line represents $Y = X^2$.

First, we compute the distance matrices (**a** and **b**) for both sets of samples $\mathcal{X}$ and $\mathcal{Y}$ respectively, where we take all the pairwise distances between data points (Fig. 3). Note that if two data points of $\mathcal{X}$ are close together (e.g., $X_i = -10$ and $X_j = -9$), their distance is small (Fig. 3 left panel, **a**). However, for Y, the distance between two data points (e.g., Y_i and Y_j) is small if the distance between corresponding data points X_i and X_j is small *or* if $X_i \approx -X_j$ (Fig. 3 right panel, **b**).

When the distance matrices are re-centered (Fig. 4, **A** and **B** correspond to the re-centered distance matrices of X and Y, respectively), we see the distances that are the most "outlying." In particular, notice that the elements in the corners of both matrices have the largest magnitudes. In this example, the corners represent the greatest outlying data points (e.g., when $X \in \{[-10, -9], [9, 10]\}$) and correspond to the columns and rows that have the greatest difference in distances.

Finally, we compute the element multiplication of both **A** and **B** (Fig. 5). Three properties of the structure emerge from this matrix. First, most of the elements in the center of the matrix are small because in the range $X \in [-5, 5]$, X and Y appear to have little

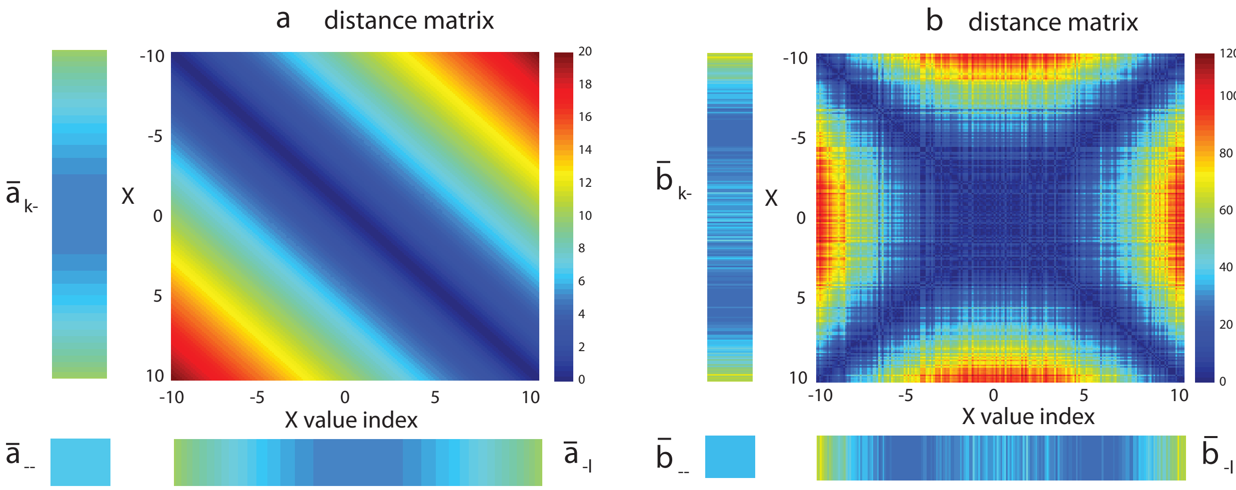

Figure 3: Distance matrices for the bi-variate non-linear dependency example. The distance matrix **a** represents the distances between observed data points in $\mathcal{X}$ (left panel), and likewise for **b** and $\mathcal{Y}$ (right panel). Each element is the Euclidean distance between two observed points. The matrices are indexed by the sorted values of X. For **b**, Y will have similar values when $X = \{-10, 10\}$ (corners of **b**), and thus have small distances. The surrounding bars correspond to the means of the rows ($\mathbf{a}_{k-}$ and $\mathbf{b}_{k-}$), means of the columns ($\mathbf{a}_{-\ell}$ and $\mathbf{b}_{-\ell}$), and the means of all elements in the matrix ($\mathbf{a}_{--}$ and $\mathbf{b}_{--}$). Note the scales are different for **a** and **b**.

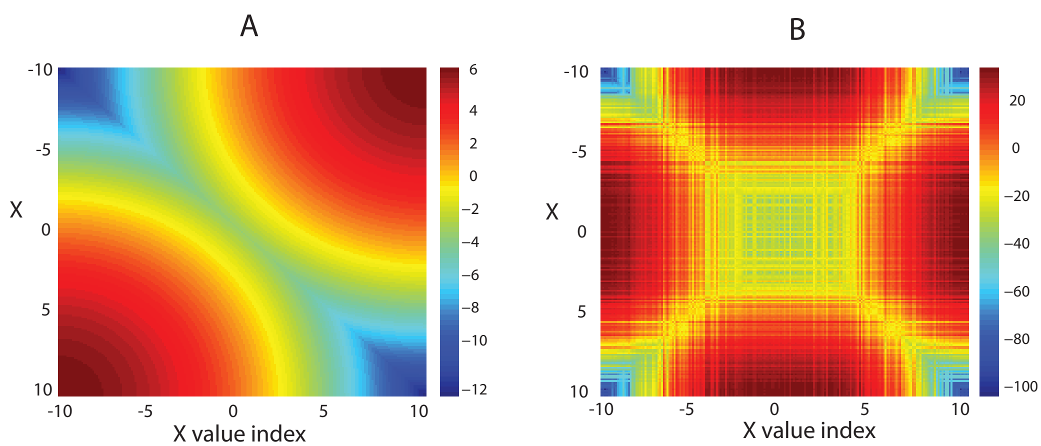

Figure 4: Re-centered distance matrices, **A** and **B**, that were computed by a linear combination of the distance matrices **a** and **b** and the means, respectively. The matrices are indexed in the same manner as in Figure 3.

dependency. Second, the upper left and bottom right corners (i.e., on the main diagonal) have large values because as the magnitude of X grows, one can find more dependency in Y (i.e., the ends of the parabola have strong non-zero correlation). Third, the upper right and bottom left corners (i.e., on the minor diagonal) have negative values because the distance covariance measure expects when the distance of two data points in one data set (X_i and X_j) is small, then the distance of the corresponding Y_i and Y_j should also be small, but this is not the case for this example. However, the magnitudes of the minor diagonal are smaller than the magnitudes of the main diagonal. Thus, the correlation at the ends of the parabola are stronger than the fact that in those regions the X's are very far apart whereas the Y's are similar. The resulting distance covariance is $\mathcal{V}_n^2(X,Y) = 22.22$, significantly larger than zero ($p < 0.01$, random permutation test).

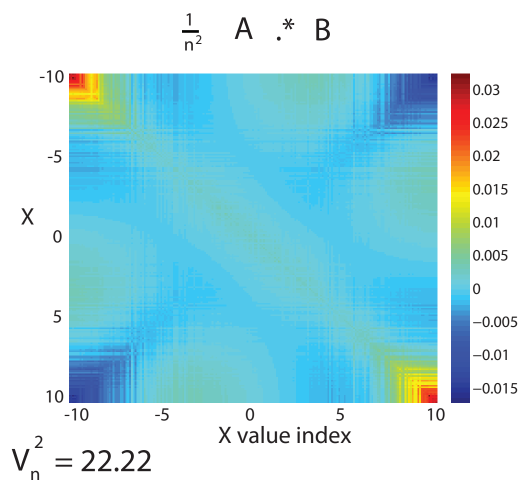

Figure 5: The normalized element-wise multiplication of **A** and **B** from Figure 4. The (k,ℓ)th element corresponds to $\frac{1}{n^2}\mathbf{A}_{k,\ell}\mathbf{B}_{k,\ell}$. The matrix is indexed in the same manner as in Figure 3. The square of the sample distance covariance was $\mathcal{V}_n^2(X,Y) \approx 22.22$, significantly different from zero ($p < 0.01$).

Example 2: Independent Gaussian random vectors

In the next example, we ask how well the sample distance covariance can assess truly independent data. The variables were sampled from the following distributions:

$$\mathbf{X} \sim \mathcal{N}(0, I)$$

$$\mathbf{Y} \sim \mathcal{N}(0, I)$$

where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^2$. Thus, $\mathbf{X}$ and $\mathbf{Y}$ have *no* dependency (Fig. 6). We test if $\mathbf{X}$ is independent of $\mathbf{Y}$.

As expected, the distance matrices (**a** and **b**) reveal small distances between data points in $\mathcal{X}$ and Y, respectively (Fig. 7). There is no apparent structure in **a** or **b**.

Re-centering the distance matrices (**A** and **B**) still reveals no structure (Fig. 8).

Taking the element-wise multiplication of the two matrices **A** and **B** reveals a matrix of elements with small magnitudes (Fig. 9). The resulting distance covariance $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = 0.076$ is not significantly far from zero ($p = 0.82$, random permutation test).

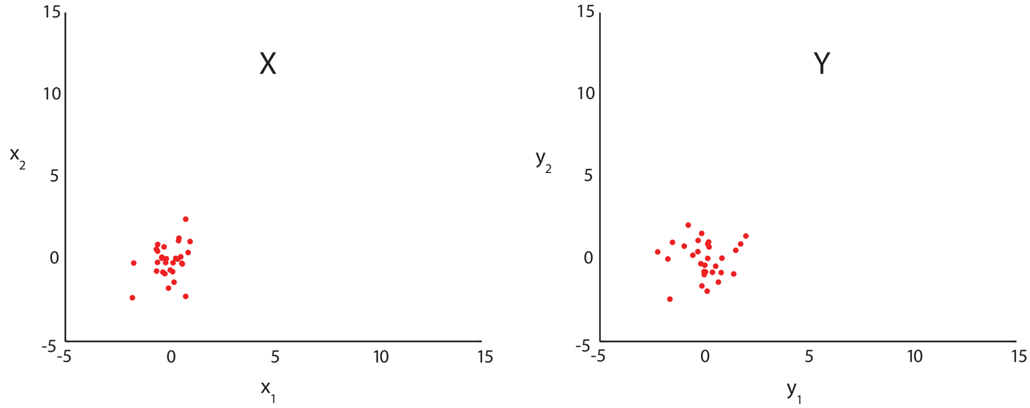

Figure 6: Independent Gaussian random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^2$ have no dependencies between them, and were drawn from a standard multivariate normal.

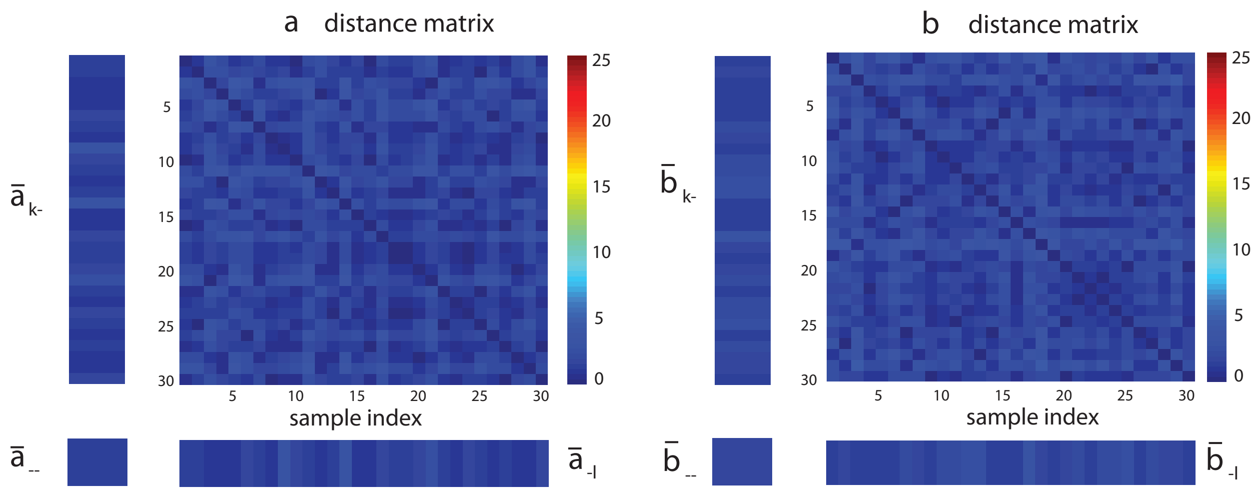

Figure 7: Distance matrices for the independent random vectors example. The distance matrix $\mathbf{a}$ represents the distances between observed data points in $\mathcal{X}$ (left panel), and likewise for $\mathbf{b}$ and $\mathcal{Y}$ (right panel). Each element is the Euclidean distance between two observed points. The matrices are indexed in no particular order. The surrounding bars correspond to the means of the rows ($\mathbf{a}_{k-}$ and $\mathbf{b}_{k-}$), means of the columns ($\mathbf{a}_{-\ell}$ and $\mathbf{b}_{-\ell}$), and the means of all elements in the matrix ($\mathbf{a}_{--}$ and $\mathbf{b}_{--}$).

Example 3: Clustered random vectors

Our final example involves clustered data (Fig. 10). Note that the clusters for $\mathbf{X} \in \mathbb{R}^2$ (Fig. 10, left panel) correspond (in color) to the clusters for $\mathbf{Y} \in \mathbb{R}^2$ (Fig. 10, right panel). Thus, there are dependencies between X and Y, but these dependencies cannot necessarily be captured by Pearson's correlation. We investigate if the distance covariance measure can pick up on the dependencies between $\mathbf{X}$ and $\mathbf{Y}$.

The distance matrices (Fig. 11, $\mathbf{a}$ and $\mathbf{b}$) reveal structure between the clusters. In

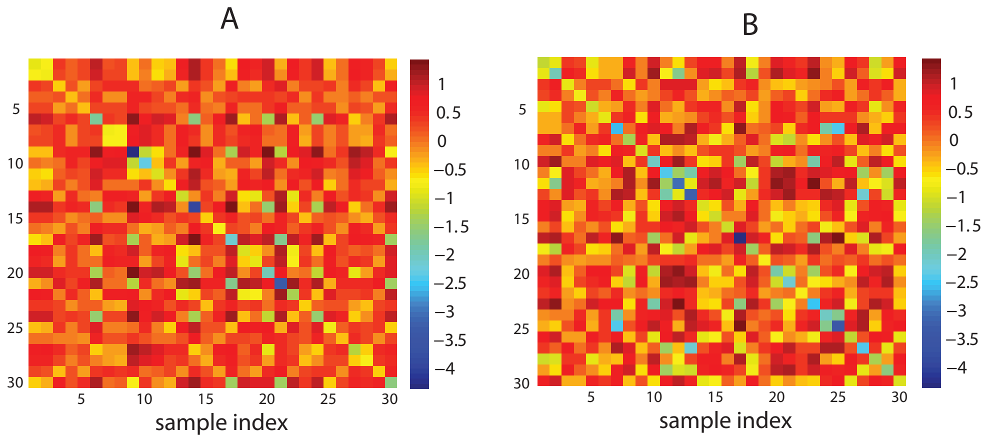

Figure 8: Re-centered distance matrices, **A** and **B**, that were computed by a linear combination of the distance matrices **a** and **b** and the means, respectively. The matrices are indexed in the same manner as in Figure 7.

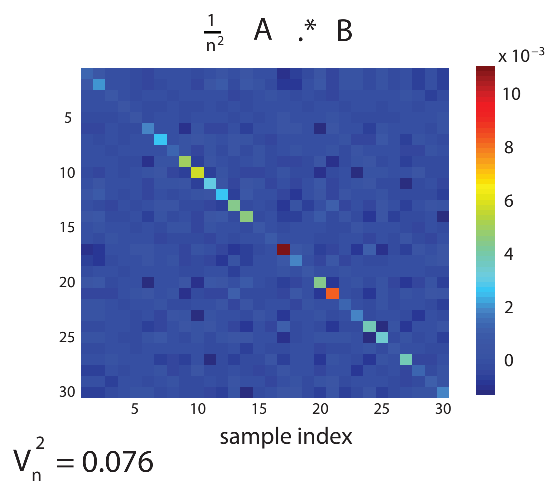

Figure 9: The normalized element-wise multiplication of **A** and **B** from Figure 8. The (k,ℓ)th element corresponds to $\frac{1}{n^2}\mathbf{A}_{k,\ell}\mathbf{B}_{k,\ell}$. The matrix is indexed in the same manner as in Figure 7. The square of the sample distance covariance was $\mathcal{V}_n^2(X,Y) \approx 0.076$, not significantly different from zero ($p = 0.82$).

particular, one can see in the top rows of the distance covariance matrix for **X** (**a**, Fig. 11, left panel) that the data points in the red cluster are far from the data points in the green cluster (the middle elements $\mathbf{a}_{1:10,11:20}$ are larger than those in $\mathbf{a}_{1:10,1:10}$) and even further from the data points in the blue cluster (the rightmost elements $\mathbf{a}_{1:10,21:30}$ are larger than those in $\mathbf{a}_{1:10,1:20}$). The same structure is maintained for the distances of **Y** (**b**, Fig. 11, right panel).

The re-centered matrices (**A** and **B**) maintain similar structure as the distance matrices (Fig. 12), but put more emphasis on the outlying data points (data points in the red and

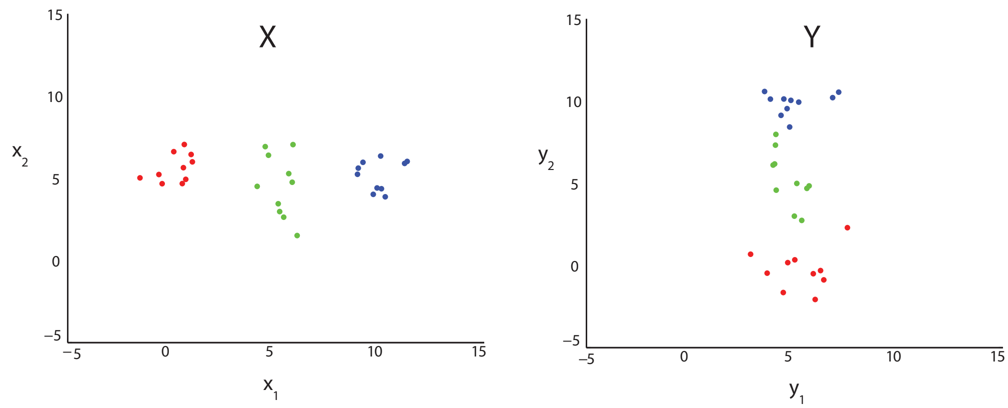

Figure 10: Example with clustered data with dependencies between $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^2$. The data of $\mathbf{X}$ were sampled from standard normals and then shifted to one of three means: {(0,5), (5,5), (10,5)}. The data of $\mathbf{Y}$ was generated in the same manner except with different means: {(5,0), (5,5), (5,10)}. The data of $\mathbf{X}$ and $\mathbf{Y}$ were ordered such that the data points in the red cluster for $\mathbf{X}$ corresponded to the data points in the red cluster for $\mathbf{Y}$. Thus, dependencies existed between clusters of the same color.

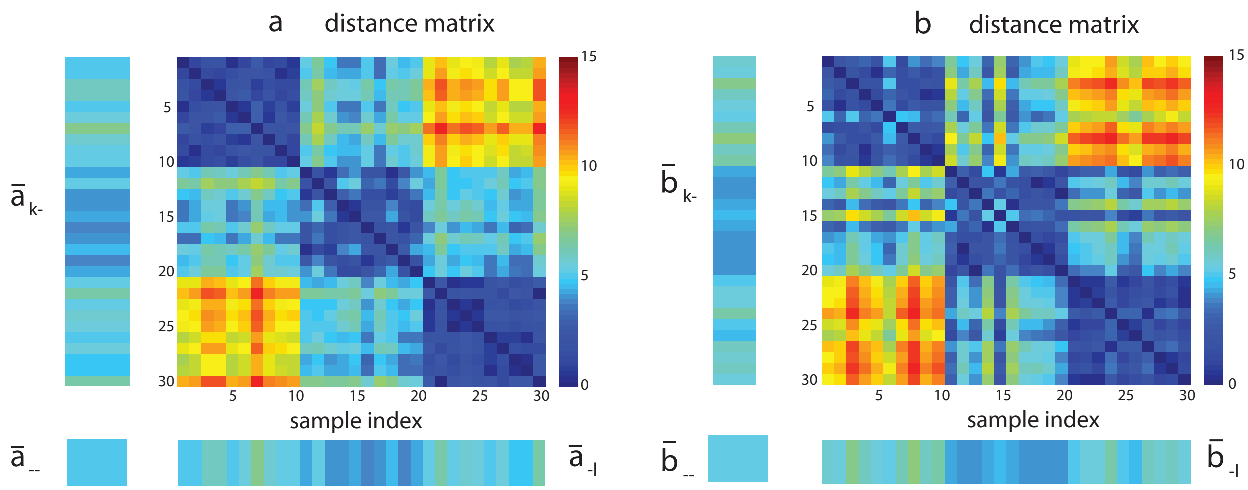

Figure 11: Distance matrices for the independent random vectors example. The distance matrix $\mathbf{a}$ represents the distances between observed data points in $\mathcal{X}$ (left panel), and likewise for $\mathbf{b}$ and $\mathcal{Y}$ (right panel). Each element is the Euclidean distance between two observed points. The matrices are indexed such that rows and columns 1 to 10 in $\mathbf{a}$ and $\mathbf{b}$ correspond to observed data points in the red clusters, rows and columns 11 to 20 correspond to the green clusters, and rows and columns 21 to 30 correspond to the blue clusters. The surrounding bars correspond to the means of the rows ($\mathbf{a}_{k-}$ and $\mathbf{b}_{k-}$), means of the columns ($\mathbf{a}_{-\ell}$ and $\mathbf{b}_{-\ell}$), and the means of all elements in the matrix ($\mathbf{a}_{--}$ and $\mathbf{b}_{--}$).

blue clusters, top left and bottom right corners of the matrices).

The resulting element-wise multiplication between $\mathbf{A}$ and $\mathbf{B}$ reveals strong dependencies

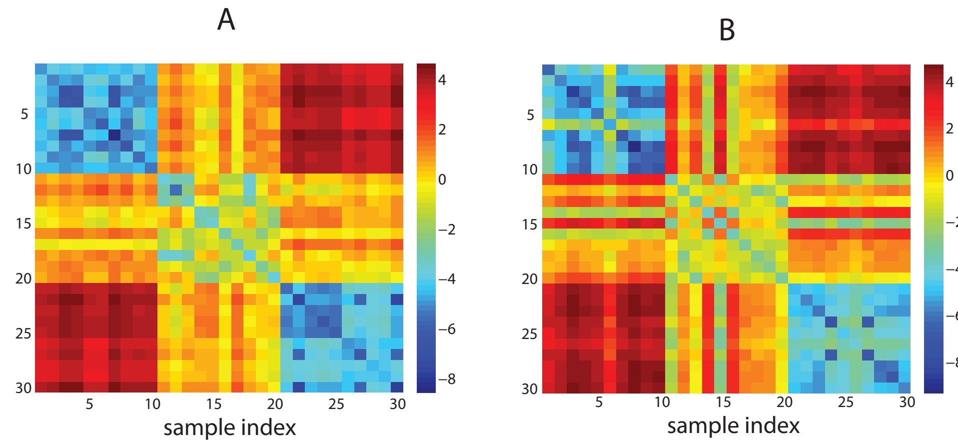

Figure 12: Re-centered distance matrices, **A** and **B**, that were computed by a linear combination of the distance matrices **a** and **b** and the means, respectively. The matrices are indexed in the same manner as in Figure 11.

within clusters (Fig. 13, top left and bottom right corners) and strong dependencies between outlying clusters (Fig. 13, top right and bottom left corners). This reveals the important assumptions of the distance covariance: Two data points that have small distance in X should have a small distance for the corresponding data points in $\mathcal{Y}$. Likewise, if $\mathbf{X}_i$ is an outlier in $\mathcal{X}$, then the corresponding $\mathbf{Y}_i$ should be an outlier in $\mathcal{Y}$. The sample distance covariance for this example was $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = 8.44$, significantly larger than zero ($p < 0.01$, random permutation test).

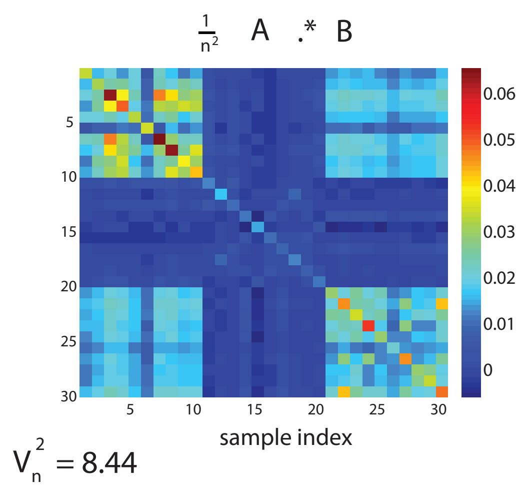

Figure 13: The normalized element-wise multiplication of **A** and **B** from Figure 12. The (k, ℓ)th element corresponds to $\frac{1}{n^2}\mathbf{A}_{k,\ell}\mathbf{B}_{k,\ell}$. The matrix is indexed in the same manner as in Figure 11. The square of the sample distance covariance was $\mathcal{V}_n^2(X, Y) \approx 8.44$, significantly different from zero ($p < 0.01$).

1.7.2 Simulation 2: Lack of monotonicity

Let $f_\lambda(x) = \sin(\lambda \frac{\pi}{2} x)$ and let $X \sim \text{Uniform}\,(0,1)$. We have that for $\lambda \in (0,1]$, function f_λ is monotone increasing in x, while for $\lambda > 1$ it is not monotone any more (see Figure 14). We can expect that, when trying to assess the strength of the relationship between random variables $(X, Y^{(\lambda)})$, where $Y^{(\lambda)} = f_\lambda(X) + \epsilon, \epsilon \sim N(0, \sigma_\epsilon^2)$ independent noise, a correlation test based on Pearson product moment, Spearman's rank coefficient or Kendall's rank coefficient, will not be so powerful for $\lambda \approx 2 + 4k$, with $k \in \mathbb{N}$. In particular, for the case of Pearson correlation, for those values of λ, the best linear approximation of the relationship between $Y^{(\lambda)}$ and $X \sim \text{Uniform}(0,1)$ will be approximately an horizontal regression line $y = \beta_0 + \beta_1 x$, where $\beta_1 = \rho_{XY} \frac{sd(Y)}{sd(X)} \approx 0$, where $\rho_{XY} \approx 0$ is the Pearson's correlation coefficient. Moreover, Spearman's and Kendall's rank coefficients will also be attenuated because of lack of monotonicity. We instead expect that a *distance covariance* test should be able to detect the presence of *any possible kind of dependence* (Szekely 2009), i.e. for any λ, a test $H_0 : X, Y$ independent based on the distance covariance should reject H_0 more easily.

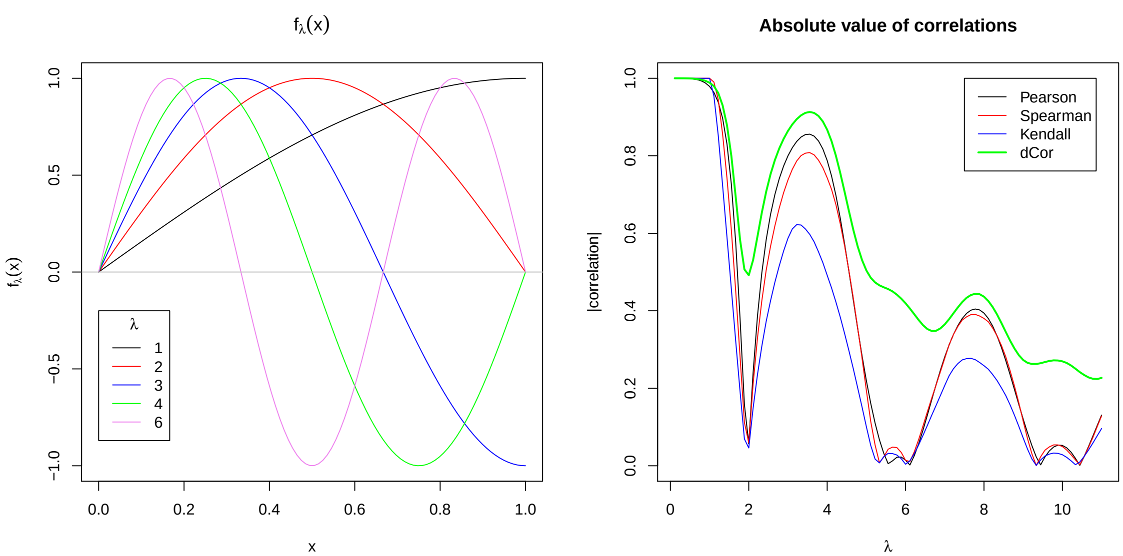

Figure 14: Left: plot of $f_\lambda(x)$ against x, for $\lambda = 1, 2, 3, 4, 6$. We can see that for $\lambda = 1$, $f_\lambda(x)$ is strictly increasing, while for $\lambda > 1$ the function is not monotone any more. Right: absolute value of Pearson, Spearman, Kendall and distance correlations of random variables $X, Y^{(\lambda)}$.

Correlation measures as functions of parameter λ

In this section we want to provide a preliminary intuition of how the different correlation measures mentioned above are affected by the lack of monotonicity. In Figure 14 we plot Pearson, Spearman, Kendall and distance correlations of random variables $X, Y^{(\lambda)}$, where $X \sim \text{Uniform}(0,1)$ and $Y^{(\lambda)} = f_\lambda(X)$ (so no additive noise). As introduced above, we can see that all the coefficients show their highest (absolute) values when f_λ is monotone ($\lambda \in (0,1)$), while for all the other cases ($\lambda > 1$) all the coefficients are attenuated, but the distance correlation does not vanish for values $\lambda \approx 2 + 4k, k \in \mathbb{N}$, which is the case of all the other correlation coefficients.

Power functions

We are going to estimate (by using R) the power and probability of TYPE I error of tests based on Pearson product moment, Spearman's rank coefficient, Kendall's rank coefficient, and distance covariance. In particular we plot the power of all the tests mentioned above, as function of sample size and parameter λ. We generate data $X_i \overset{iid}{\sim} \text{Uniform}(0,1)$ and $Y_i^{(\lambda)} = f_\lambda(X_i) + \epsilon_i$, with $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon)$. In particular we choose $\sigma_\epsilon = 0.5$ and for a pair (n, λ) we estimate the probability of rejecting hypothesis "$H_0 : X, Y$ independent" as follows:

1. Generate $\{(X_i, Y_i)\}_{i=1}^n$
2. Test H_0 based on data generated
3. Repeat 1-2 many times (e.g. 500) and compute the proportion of rejections.

In Figure 15 we can see that when testing independence when the random variables X, Y are linked by a monotonic relationship ($\lambda = 1$) then all the different tests are able to detect dependence, showing an increasing power with respect to the sample size n. Yet, when the relationship is not monotone, but concave ($\lambda = 2$) then, as expected, tests based on Pearson product moment, Spearman's rank coefficient, and Kendall's rank coefficient show a very small power even constant with respect to the sample size, while the distance covariance test shows an increasing power in the sample size. Moreover, we also compare the performances of distance covariance tests based on permutation and asymptotic approximation (χ_1^2). We can see that estimating the distribution of the distance covariance test statistic via permutation leads to a well controlled probability of TYPE I error, while the asymptotic approximation does not (at least for the sample sizes considered). This is why for the other simulations we will only use the permutation approximation.

In Figure 16 we plot the power of the tests as function of $\lambda \in (0, 12)$ and fixed sample size $n = 50$. We can see that the power of the distance covariance test, at least for $\lambda > 1$, is always larger than any other test that, as expected show very low power for $\lambda \approx 2 + 4k, k$ integer.

1.7.3 Simulation 3: heteroskedasticity

Let $Y = Zg(X)$, where $g(X)$ is some function of $X \sim \text{Uniform}(0,1)$, and $Z \sim N(0,1)$. We have that $\mathbb{E}[Y|X = x] = 0, \forall x$, but $\text{Var}(Y|X = x) = g(x)^2$. Therefore, for any nonconstant function $g(x)$, the distribution of $Y|X$ is affected by X, i.e. X, Y are not independent. In Figure 17 we consider the case $g(x) = \sin(3\pi x)$, and we compare the power of tests (by using R) based on Pearson, Spearman, Kendall, and Distance covariance. We can see that as the sample size increases, the power of the distance covariance test increases, while all the other tests do not show any improvement. This example suggests that, when fitting a

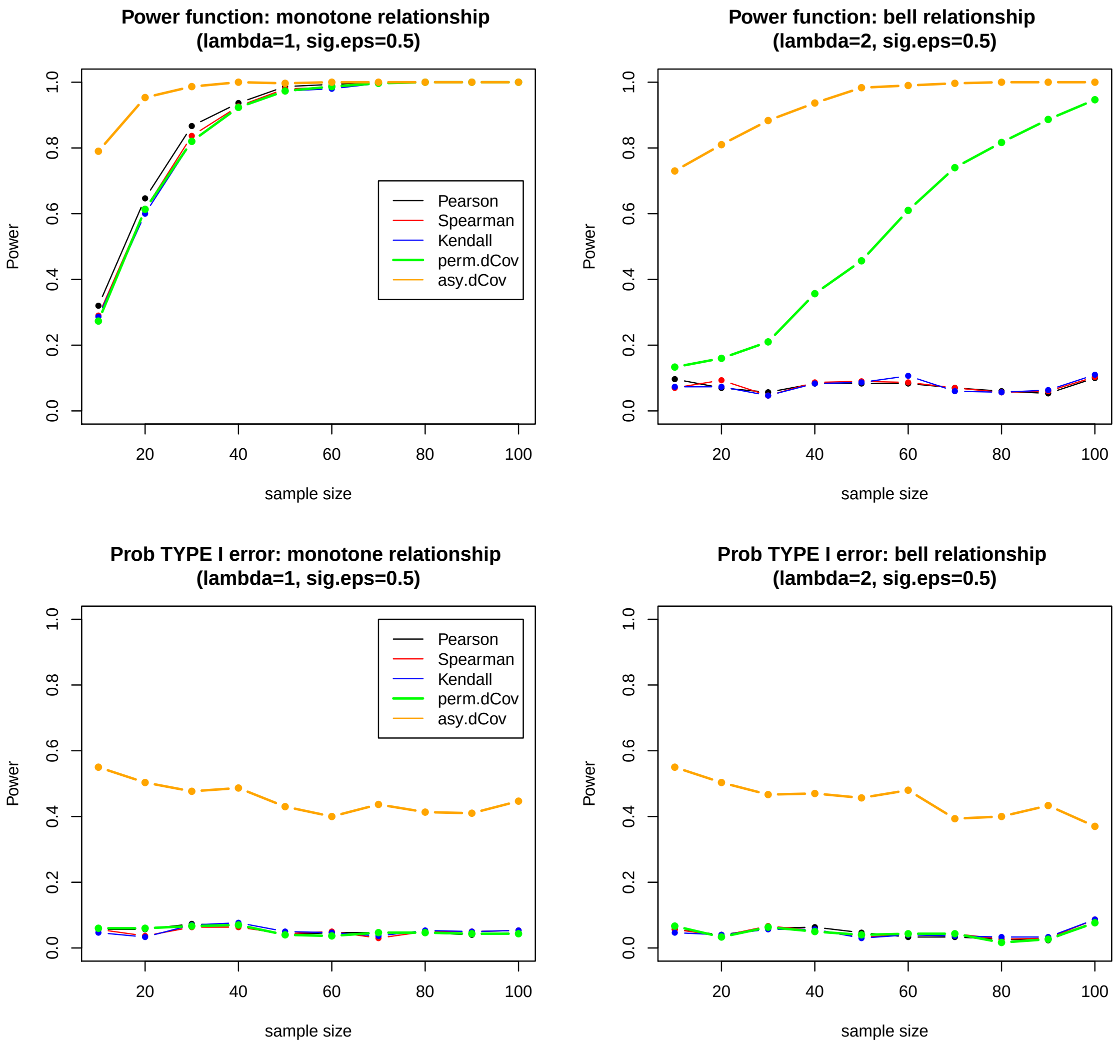

Figure 15: Power and probability of TYPE I error of the tests as functions of the sample size. Top left: power of the tests when there is a monotone relationship between the two variables; all the tests perform well. Top right: power of the tests when there is a nonmonotone (bell) relationship between the two variables; only the distance covariance test shows increasing power in detecting the nonmonotone dependence. Bottom: correspondent probabilities of TYPE I error. We can see that the asymptotic distance covariance test shows a too large probability of TYPE I error, while for all the others that probability is garanteed to be smaller than the test level (5%).

regression curve, the analysis of the residuals (e.g. checking for heteroskedasticity) could be supported by a test of independence based on the distance covariance.

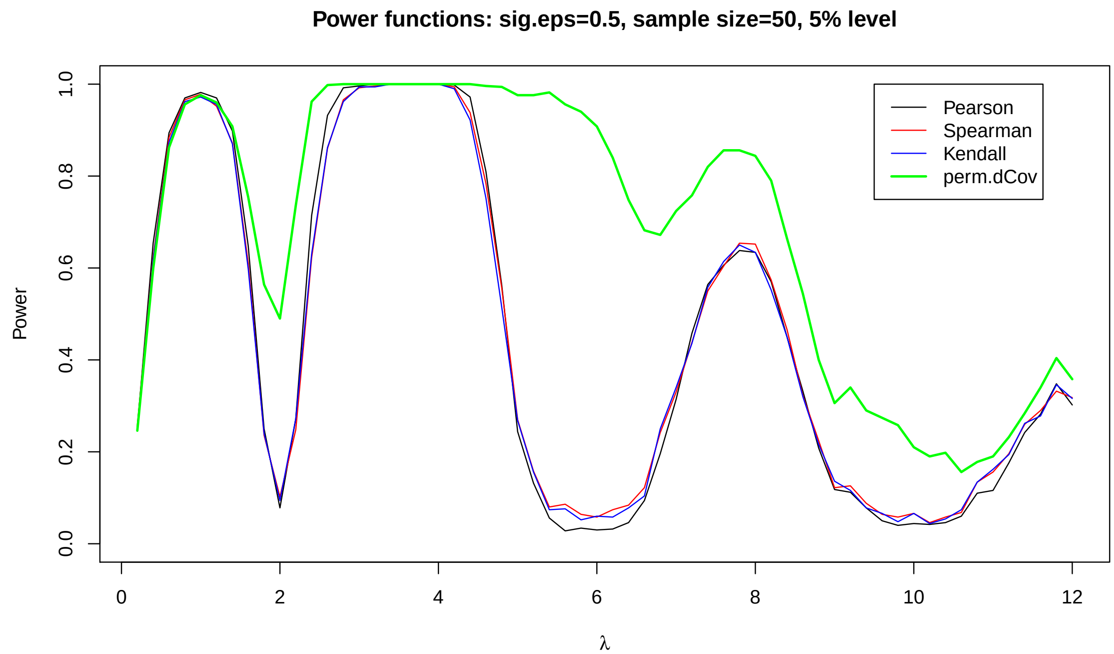

Figure 16: Power of the tests as function of $\lambda \in (0, 12)$ and fixed sample size $n = 50$. The distance covariance test performs generally better than any other test.

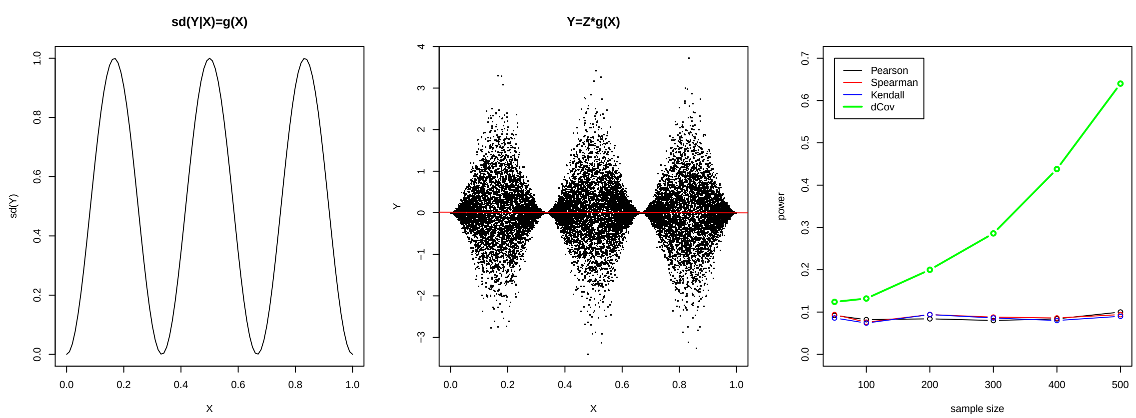

Figure 17: $Y = Zg(X)$, where $g(X)$ is some function of $X \sim \text{Uniform}(0, 1)$, and $Z \sim N(0, 1)$. Left: standard deviation of $Y|X = x$ against x. Center: a large sample of (X, Y) to describe the shape of their joint distribution. Right: power functions of tests Pearson, Spearman, Kendall, and distance covariance. Distance covariance test is the only one that shows improvements in power for increasing sample size.

2 Discussion

- Recall that distance correlation is unsigned with values in $[0, 1]$. The main advantages of this measure are: it is equal to zero if and only if the two considered random vectors of arbitrary dimensions are independent; it has a straightforward formula; its sample version is easy to compute since it is based on distance matrices of the sample points; the equivalence of distance covariance with Brownian covariance leads to some useful intuitions about why a test of independence based on it can be used to detect any kind of relationship. One disadvantage of the distance correlation, compared to Pearson's correlation, is that there is little interpretation to the value of the distance correlation. One must simply ask if the distance correlation is significant or not; one cannot interpret if the correlation comes from linear or non-linear interactions. On the other hand, Pearson's correlation gives a sense of the slope of the regressed line between two variables. Thus, one may be able to garner more information about the dependencies between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ by first fitting a linear regression model between $\mathbf{X}$ and $\mathbf{Y}$, and then testing for independence between the residuals of the fitted model. A significant dependence between the residuals would reveal that the data have non-linear interactions that could not be captured by the linear model.

- The Brownian covariance of two random vectors $\mathbf{X} \in \mathbb{R}^p, \mathbf{Y} \in \mathbb{R}^q$ is obtained by averaging over all the possible realizations of two independent Wiener fields and with respect to the joint distribution of $(\mathbf{X}, \mathbf{Y})$. Since a Wiener field has the property of having continuous trajectories a.s., we can see the Brownian covariance as an approximate average over all the possible continuous functions $f : \mathbb{R}^p \to \mathbb{R}$, $g : \mathbb{R}^q \to \mathbb{R}$ evaluated at random points $(\mathbf{X}, \mathbf{Y})$. This intuition, combined with Theorem 1, can help to understand why tests based on the distance covariance $\mathcal{V}(\mathbf{X}, \mathbf{Y})$ can detect any kind of relationship between $\mathbf{X}$ and $\mathbf{Y}$. We can see also that averaging with respect to Wiener processes is much easier than maximizing with respect to all the possible functions, as it is instead required to compute the maximal correlation (Renyi 1951).

- The distribution of the test statistic $n\mathcal{V}_n^2/T_2$ can be approximated via permutation or via χ_1^2 asymptotic approximation. As seen in the simulations, the first method should be generally preferred since the probability of TYPE I error is well controlled, while for the second method it might not, especially for not so large sample sizes. However, for very large sample sizes, we expect that the computational time due to permutations could make the use the χ_1^2 approximation convenient. Yet, the behaviour of the probability of TYPE I error for the χ_1^2 approximation can vary not only with respect to the sample size (i.e. as $n \to \infty$, this probability will be upper bounded by the desired level of the test), but it can also be affected by the shapes of the marginal distributions of the random vectors considered.

- Because a particular weight function was chosen in which the distance covariance arose, it is unclear if other weight functions could lead to similar dependency measures. In a recent paper, [7] made a connection between the distance covariance and maximum mean discrepancies (MMD), which is based on assessing distances between embedded distributions to reproducing kernel Hilbert spaces. The paper showed that the distance covariance can be computed with a special kernel of a class of distance-induced kernels. In fact, other parametric kernels exist that can yield more powerful tests than the distance covariance. Another advantage to using kernels is that we are no longer restricted to the Euclidean domain, and can test variables in applications such as text strings and graphs.

References

[1] Feuerverger, Andrey. “A consistent test for bivariate dependence.” International Statistical Review/Revue Internationale de Statistique (1993): 419-433.

[2] Rényi, Alfrd. “On measures of dependence.” Acta mathematica hungarica 10, no. 3 (1959): 441-451.

[3] Blum, J. R., J. Kiefer, and M. Rosenblatt. “Distribution free tests of independence based on the sample distribution function.” The annals of mathematical statistics (1961): 485-498.

[4] Rosenblatt, Murray. “A quadratic measure of deviation of two-dimensional density estimates and a test of independence.” The Annals of Statistics (1975): 1-14.

[5] Székely, Gbor J., Maria L. Rizzo, and Nail K. Bakirov. “Measuring and testing dependence by correlation of distances.” The Annals of Statistics 35, no. 6 (2007): 2769-2794.

[6] Székely, Gbor J., and Maria L. Rizzo. “Brownian distance covariance.” The annals of applied statistics 3, no. 4 (2009): 1236-1265.

[7] Sejdinovic, Dino, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. ”Equivalence of distance-based and RKHS-based statistics in hypothesis testing.” The Annals of Statistics 41, no. 5 (2013): 2263-2291.