Metrika manuscript No.
(will be inserted by the editor)

Four Simple Axioms of Dependence Measures

Tamás F. Móri · Gábor J. Székely

Received: date / Accepted: date

Abstract Recently new methods for measuring and testing dependence have appeared in the literature. One way to evaluate and compare these measures with each other and with classical ones is to consider what are reasonable and natural axioms that should hold for any measure of dependence. We propose four natural axioms for dependence measures and establish which axioms hold or fail to hold for several widely applied methods. All of the proposed axioms are satisfied by distance correlation. We prove that if a dependence measure is defined for all bounded nonconstant real valued random variables and is invariant with respect to all one-to-one measurable transformations of the real line, then the dependence measure cannot be weakly continuous. This implies that the classical maximal correlation cannot be continuous and thus its application is problematic. The recently introduced maximal information coefficient has the same disadvantage. The lack of weak continuity means that as the sample size increases the empirical values of a dependence measure do not necessarily converge to the population value.

Keywords correlation · distance correlation · maximal correlation · maximal information coefficient · invariance

T. F. Móri was supported by the Hungarian National Research, Development and Innovation Office NKFIH – Grant No. K125569. Part of this research was based on work supported by the National Science Foundation, while the second author was working at the Foundation. G. J. Székely is grateful for many interesting discussions with Yakir and David Reshef, Abram M. Kagan, and Gábor Tusnády.

T. F. Móri
Department of Probability Theory and Statistics, ELTE Eötvös Loránd University, Pázmány P. s. 1/C, H-1117 Budapest, Hungary
E-mail: mori@math.elte.hu

G. J. Székely
National Science Foundation, 2415 Eisenhower Avenue, Alexandria, VA 22314, and Rényi Institute of Mathematics, Hungarian Academy of Sciences, Reáltanoda u. 13–15, H-1053 Budapest, Hungary.
E-mail: gszekely@nsf.gov

1 Introduction: Rényi's axioms

It is hard to overestimate the importance of dependence measures in statistics and in science. When we try to find the cause X that is (partly) responsible for an effect Y then it is a natural first step to find out if X and Y are statistically dependent. Thus it is not surprising that Pearson's linear correlation $\rho(X,Y)$ is responsible for many important causal discoveries like smoking and lung cancer. Unfortunately $\rho(X,Y)=0$ does not mean that X and Y are independent (the converse is true). Thus if we measure the dependence of X and Y by $\rho(X,Y)$ and it happens to be 0 then we might suspect that there is no causal relationship between X and Y even when there is. This is a typical problem when the relationship between the variables is highly nonlinear, not even monotonic. A good example is Volokh (2015) where the title of the article is 'Zero correlation between state homicide and state gun laws'.

A well-known remedy is to consider *maximal correlation*, namely $\sup_{f,g}\rho(f(X),g(Y))$ where f,g are Borel-measurable functions. Maximal correlation is zero if and only if X and Y are independent. But this fact itself does not make maximal correlation an ideal measure of dependence. In this paper we explain our concerns and suggest a solution.

Rényi (1959) proposed seven important properties of dependence measures Δ as axioms. Rényi's axioms are as follows. Let X and Y be real valued random variables.

(A) $\Delta(X,Y)$ is defined for all random variables X and Y, neither of them being constant with probability 1.
(B) $\Delta(X,Y)=\Delta(Y,X)$ (symmetry).
(C) $0\le\Delta(X,Y)\le 1$.
(D) $\Delta(X,Y)=0$ if and only if X and Y are independent.
(E) $\Delta(X,Y)=1$ if there is a strict dependence between X and Y; that is, either $X=g(Y)$ or $Y=f(X)$, where $g(x)$ and $f(x)$ are Borel measurable functions.
(F) If the Borel measurable functions $f(x)$ and $g(x)$ map the real axis in a one-to-one way onto itself, $\Delta(f(X),g(Y))=\Delta(X,Y)$.
(G) If the joint distribution of X and Y is normal, then $\Delta(X,Y)=|\rho(X,Y)|$ where $\rho(X,Y)$ is the correlation coefficient of X and Y.

Maximal correlation satisfies all of the above axioms (Rényi, 1959). Rényi's axioms collect some of the most important properties of a dependence measure, but not all of these properties are essential for a good measure of dependence. On the other hand, not even this list of strong restrictions characterizes maximal correlation, as shown by Linfoot's information-theoretical measure (Linfoot, 1957). So one might wonder which of these axioms are critically important and whether this list contains all critically important properties of dependence measures as axioms.

Our goal here is to find a "minimalist" system of axioms that we can expect to be satisfied by all acceptable dependency measures Δ. First of all, we do not want to define Δ for all random variables (that are not constant with

probability 1) because not even Pearson's correlation is defined for random variables with infinite variance. Even if we define Δ for random variables with finite variances only, the absolute value of Pearson's correlation ρ does not satisfy (E), (F). We will replace them with a weaker version where 1–1 invariance is replaced by similarity invariance satisfied by $|\rho|$. There is another reason for not assuming 1–1 invariance of Δ. The 1–1 invariance would imply the existence of many *uncorrelated* random variables X, Y for which $\Delta(X, Y) = 1$, which is counterintuitive. It is not surprising that there exist perfectly dependent random variables with zero Pearson's correlation. For a related statement see Kimeldorf and Sampson (1978). The following proposition shows that with very few exceptions for *all random variables* X one can find a 1–1 real function f such that X and $f(X)$ are uncorrelated.

Proposition 1 *Let X be a square integrable random variable defined on an arbitrary probability space. Suppose the distribution of X is not concentrated on three or less points. Then there exists a measurable injective function $f : \mathbb{R} \to \mathbb{R}$ such that X and $f(X)$ are uncorrelated. This f can be chosen piecewise linear.*

Such an f cannot exist if X takes on exactly two values, because in this case uncorrelatedness is equivalent to independence. When the distribution of X is supported on exactly 3 points then a necessary and sufficient condition for f to exist is $P(X = EX) = 0$.

For an elementary proof see the Appendix.

If this proposition were not enough justification for weakening (E) and (F), in what follows we will see that such a strong invariance is not compatible with our new axiom of continuity (axiom (iv) below). This axiom of continuity is not there among Rényi's axioms because then the system would be contradictory. But why is continuity so natural that one should suppose it as an axiom? If there is a tiny little change/perturbation in the distribution of (X, Y) and this tiny little perturbation changes $\Delta(X, Y)$ dramatically, e.g., changes it from 1 to 0 then Δ has no stability. We cannot rely our statistical inference on such an unstable Δ because a minor perturbation, no matter how small it is, can result in a completely different statistical inference. This can be viewed as a violation of distributional robustness. If we replace weak convergence by stronger forms of convergence then of course this would allow more measures of dependence to be continuous but these measures might violate distributional robustness. We do not need to disregard all nonrobust measures but we need to be aware of this deficiency.

Recall that Euclidean geometry is characterized by invariances with respect to the Euclidean group of transformations (translations, rotations, and reflections). Similarity geometry deals with geometrical objects with the same shape. We can obtain one object from another by scaling (enlarging or shrinking). Similarity transformations consist of all Euclidean transformations and all (nonzero) scaling; that is, changing the measurement units. Instead of 1–1 invariance, in our axioms we suppose similarity invariance only. Similarity invariance is something we do not want to weaken because changing the scale,

(that is, changing the measurement unit), should not affect the degree of dependence. Luckily, similarity invariance does not contradict continuity. This is shown by the example of distance correlation explained below.

The classical correlation ratio does not satisfy (B) so we dropped this axiom, too. We will see that axiom (G) is also unnecessarily restrictive and, among others, would disqualify distance correlation. For more details see below. See also Lehmann (1966), Schweizer and Wolff (1981), Dedecker and Prieur (2005), and Reimherr and Nicolae (2013) for more comments on dependence measures.

2 New axioms

Let S be a nonempty set of pairs of nondegenerate random variables X, Y taking values in Euclidean spaces or in real, separable Hilbert spaces H. (Nondegenerate means that the random variable is not constant with probability 1.) Then $\Delta(X,Y) : S \to [0,1]$ is called a dependence measure on S if the following four axioms hold. In the axioms below we need similarity transformations of H. Similarity of H is defined as a bijection (1–1 correspondence) from H onto itself that multiplies all distances by the same positive real number (scale). Similarities are known to be compositions of a translation, an orthogonal linear mapping, and a uniform scaling. We assume that if $(X,Y) \in S$ then $(LX, MY) \in S$ for all similarity transformations L, M of H.

(i) $\Delta(X,Y) = 0$ if and only if X and Y are independent.

(ii) $\Delta(X,Y)$ is invariant with respect to all similarity transformations of H; that is, $\Delta(LX, MY) = \Delta(X,Y)$ where L, M are similarity transformations of H.

(iii) $\Delta(X,Y) = 1$ if and only if $Y = LX$ with probability 1, where L is a similarity transformation of H.

(iv) $\Delta(X,Y)$ is continuous; that is, if $(X_n, Y_n) \in S$, $n = 1, 2, \dots$ such that for some positive constant K we have $E\big(|X_n|^2 + |Y_n|^2\big) \le K$, $n = 1, 2, \dots$ and (X_n, Y_n) converges weakly (converges in distribution) to (X,Y) then $\Delta(X_n, Y_n) \to \Delta(X,Y)$. (The condition on the boundedness of second moments can be replaced by any other condition that guarantees the convergence of expectations: $E(X_n) \to E(X)$ and $E(Y_n) \to E(Y)$; such a condition is the uniform integrability of X_n, Y_n which follows from the boundedness of second moments.)

Remark 1 (a) Functions of independent random variables are independent, thus property $\Delta(X,Y) = 0$ is invariant with respect to all 1–1 Borel measurable transformations of H. On the other hand we do not suppose this 1–1 invariance for other values of Δ. As we shall see, such a strong condition would contradict axiom (iv). In axiom (ii) and (iii) one can try to replace the invariance with respect to similarities by other groups of invariances, particularly, when the statistical problem in question exhibits symmetries/invariances in the sense of (Lehmann and Romano, 2005, Chapter 6), see also Eaton (1989).

It is up to the statistician to choose the right level of invariance. Too much invariance is not necessarily good. Even if a very strong invariance of Δ does not contradict other important axioms it might decrease the power of Δ in testing independence. If $H = \mathbb{R}$, the real line, affine transformations coincide with similarities. In higher dimensions, however, affine invariance for all bounded nonconstant random variables contradicts axiom (iv) as it is proved in Theorem 1. This makes the choice of similarity invariance in our axioms even more natural.

(b) Rényi did not assume axiom (iv). Theorem 1 below explains that if he did then no dependence measure would have satisfied all his axioms.

(c) Why did we suppose that S does not contain random variables that are constant with probability 1? Because if Y is such a random variable then it is independent of all random variables X and thus by axiom (i) we have $\Delta(X, Y) = 0$. On the other hand, for all $X \in S$ axiom (iii) implies $\Delta(X, X/n) = 1$ for $n = 1, 2, \ldots$. But for bounded random variables X the limit of X/n is 0 and $\Delta(X, 0) = 0$ which contradicts axiom (iv). In axiom (A) Rényi also assumes that the random variables X and Y are not constant with probability 1, i.e., their distributions are nondegenerate. This assumption guarantees that Δ cannot be discontinuous at degenerate distributions because Δ is simply not defined there. Thus Rényi did not overlook the importance of weak continuity of Δ, he just could not assume it because it would have been inconsistent with his other axioms.

Let us see that our system of new axioms is not contradictory when S is the set of all nondegenerate random variables with finite expectation. For this it is sufficient to define a dependence measure that satisfies the four axioms. Such a measure is *distance correlation*, which was introduced in Székely et al. (2007).

First of all recall the definition of the sample distance correlation. Take all pairwise distances between sample values of one variable, and do the same for the second variable. Rigid motion invariance is automatically guaranteed if instead of sample elements we work with their distances. Another advantage of working with distances is that they are always real numbers even when the data are vectors of possibly different dimensions. Once we have computed the distance matrices of both samples, double-center them (so each has column and row means equal to zero). Then average the entries of the matrix which holds componentwise products of the two centered distance matrices. This is the square of the sample distance covariance. If we denote the centered distances by A_{ij}, $i, j = 1, \ldots, n$ and B_{ij}, $i, j = 1, \ldots, n$ where n is the sample size, then the squared sample distance covariance is

$$\frac{1}{n^2} \sum_{i,j=n}^{n} A_{i,j} B_{i,j}.$$

This definition is very similar to, and almost equally simple as, the definition of Pearson's covariance, except that here we have double indices.

The population squared distance covariance can be reduced to the following form (Székely et al., 2007) if $E|X|^2$ and $E|Y|^2$ are finite. Let (X, Y), (X, Y'), (X'', Y'') be independent and identically distributed then the distance covariance is the square root of

$$\begin{aligned}\mathrm{dCov}^2(X,Y) := E(|X-X'|\,|Y-Y'|) + E(|X-X'|)E(|Y-Y'|) \\ - E(|X-X'|\,|Y-Y''|) - E(|X-X''|\,|Y-Y'|).\end{aligned}$$

In the above referred paper we proved that $\mathrm{dCov}(X, Y)$ is a metric, and the distance variance, $\mathrm{dCov}(X, X)$ is zero if and only if X is constant with probability 1. Once we defined distance covariance and distance variance we can define distance correlation the same way as we defined correlation with the help of covariance and variance. If the random variables X, Y have finite expected values and they are not constant with probability 1 then the definition of *population distance correlation* is the following:

$$\mathcal{R}(X,Y) := \frac{\mathrm{dCov}(X,Y)}{\sqrt{\mathrm{dCov}(X,X)\,\mathrm{dCov}(Y,Y)}}\,.$$

If $\mathrm{dCov}(X, X)\,\mathrm{dCov}(Y, Y) = 0$ then define $\mathcal{R}(X, Y) = 0$. Distance correlation equals zero if and only if the variables are independent, whatever be the underlying distributions and whatever be the dimension of the two variables (for a transparent explanation see below). This fact and the simplicity of the statistic make distance correlation an attractive candidate for measuring dependence. For generalizations to metric spaces see Lyons (2013) and Jakobsen (2017).

In Székely et al. (2007) an alternative formula for $\mathrm{dCov}^2(X, Y)$ was given in terms of characteristic functions $f_{X,Y}$, f_X and f_Y of (X, Y), X, and Y respectively. If the random variable X takes values in a p-dimensional Euclidean space $\mathbb{R}^p$ and Y takes values in $\mathbb{R}^q$ and both variables have finite expectations we have

$$\mathrm{dCov}^2(X,Y) := \frac{1}{c_p c_q}\int_{R^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}\,|s|_q^{1+q}}\,dt\,ds.$$

where c_p and c_q are constants. This formula clearly shows that independence of X and Y is equivalent to $\mathrm{dCov}(X, Y) = 0$. It is interesting to note that in Hoeffding's dissertation (Hoeffding, 1940) it is proved that for real valued X and Y with finite variance, Pearson's covariance

$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [F_{X,Y}(x,y) - F_X(x)F_Y(y)]\,dx\,dy,$$

where F denotes the cumulative distribution functions. Thus we might want to define a sign or rather a direction of distance covariance and distance correlation as the argument of the complex number

$$z := \int_{\mathbb{R}^{p+q}} [f_{X,Y}(t,s) - f_X(t)f_Y(s)]\,w(t,s)\,dt\,ds,$$

where $w(t,s)$ is a suitable weight function. In the most natural case of $w(-t,-s) = w(t,s)$, this z is always real, so its direction is not more than a sign. Unfortunately in the most natural choice for w when $w(s,t) = \left(|t|_p^{1+p}|s|_q^{1+q}\right)^{-1}$, it is not trivial that z exists at all. We plan to return to this problem in another paper. We also note that in Hoeffding (1948) a test of independence was introduced, based on

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2\, dF_{X,Y}(x,y).$$

If the expectations of X, Y do not exist, we can generalize distance correlation for random variables with finite $\alpha > 0$ moments. See Székely et al. (2007); Székely and Rizzo (2009). It is easy to see that the population distance correlation, $\mathcal{R}(X,Y)$, satisfies axioms (ii) and (iv). For the proof that $\mathcal{R}(X,Y)$ satisfies (i) and (iii), see Székely et al. (2007).

In the special case when (X,Y) are jointly distributed as bivariate normal, distance correlation $\mathcal{R}$ is a deterministic function of Pearson correlation $\rho = \rho(X,Y)$ (Székely et al., 2007, Theorem 7), namely,

$$\mathcal{R}^2(X,Y) = \frac{\rho\arcsin\rho + \sqrt{1-\rho^2} - \rho\arcsin(\rho/2) - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}}.$$

Note that this is a strictly increasing, convex function of $|\rho|$, $\mathcal{R}(X,Y) \le |\rho(X,Y)|$ with equality when $\rho = 0$ or $\rho = \pm 1$. Thus $\mathcal{R}(X,Y)$ does not satisfy Rényi's axiom (G). It is also clear that if Δ satisfies our four axioms then $h(\Delta)$ also satisfies them whenever h is a strictly increasing, continuous function, $h(0) = 0$, $h(1) = 1$, and $0 < h(x) < 1$ for $0 < x < 1$. In the definition of partial distance correlation (Székely and Rizzo, 2014) $h(x) = x^2$ is applied. In this case the distance standard deviations of the random variables X, Y are measured in the same units as the X distances and Y distances. If we insisted on axiom (G) we would disqualify distance correlation and also its square and instead would have accepted a complicated function of distance correlation as "legitimate".

An important generalization of distance correlation is Sejdinovic et al. (2013). This is related to a generalized distance correlation where the distance is a more general metric than the Euclidean one. These generalizations under some natural conditions like scale invariance also satisfy our axioms.

With the new system of axioms our goal was not to characterize a single dependence measure. The new system of axioms is "minimalist" in the sense that all good dependence measures can be expected to satisfy them. We show that even this "minimalist" system of axioms can disqualify several classical measures and also some recently introduced measures of dependence. For example, we will see that neither the maximal correlation coefficient nor the recently introduced maximal information coefficient satisfy axiom (iv). The same axiom fails to hold for the correlation ratio as shown below.

3 Important dependence measures

In this section we give a list of important dependent measures and discuss some of their their connections. Then we discuss if they satisfy our new axioms.

Example 1 (Pearson's correlation ρ) Let S be the set of bivariate Gaussian random variables (X, Y). The absolute value of Pearson's classical correlation $\rho(X, Y)$ satisfies axioms (i) – (iv). On the history of ρ see Pearson (1920) and Stigler (1989). We know that $\rho = 0$ if and only if X, Y are independent, and $|\rho| = 1$ if and only if there is a linear relationship between X and Y.

For a multivariate version see Escoufier (1973) and Josse and Holmes (2014). It is well-known that Pearson's correlation does not satisfy axiom (i) for general random variables. This problem is partially addressed in the next example.

Example 2 (Spearman's ρ and Kendall's τ) If H is the real line then the invariance of Δ with respect to all monotone transformations means that Δ is independent of the marginal distributions of X, Y. Monotone invariance implies that instead of the joint cdf $F_{X,Y}$ we can focus on the copula $C(u; v) = F_{X,Y}(F_X^{-1}(u); F_Y^{-1}(v))$ where F_X^{-1}, F_Y^{-1} denote the generalized inverse functions of the cdf's F_X and F_Y of X and Y, respectively. The copula can also be viewed as the joint distribution of two uniform $(0, 1)$ variables. We have $C(F_X(x), F_Y(y)) = F_{X,Y}(x, y)$. Two random variables are independent if and only if $C(u, v) = uv$.

With the copula function $C(u, v)$ Spearman's ρ (Spearman, 1904) and Kendall's τ (Kendall, 1938) can be defined as follows:

$$\rho_S(X, Y) := 12 \int_0^1 \int_0^1 (C(u, v) - uv)\, du\, dv,$$

and

$$\tau(X, Y) := 4 \int_{[0,1]^2} C(u, v)\, dC(u, v) - 1,$$

respectively. An equivalent definition is

$$\tau := P((X - X')(Y - Y') > 0) - P((X - X')(Y - Y') < 0),$$

where (X', Y') is an iid copy of (X, Y).

The absolute values, $|\rho_S|$ and $|\tau|$, satisfy (i) – (iv) for positive quadrant dependent or for negative quadrant dependent random variables: these properties mean that $C(u, v) \geq uv$ or $C(u, v) \leq uv$, respectively for all $0 \leq u, v \leq 1$. For general random variables X, Y axiom (i) typically does not hold.

Example 3 (Affine and monotone invariant distance correlation) Distance correlation applied to standardized random variables is obviously affine invariant. It is defined for random vectors X and Y with nonsingular covariance matrices Σ_X and Σ_Y, resp., as

$$\Delta(X, Y) = \mathcal{R}\big(\Sigma_X^{-1/2} X,\, \Sigma_Y^{-1/2} Y\big).$$

For interesting consequences see Dueck et al. (2014). This affine invariant distance correlation is continuous, because so is the standardization on the set of bounded random variables with nonsingular covariance matrices. This fact does not contradict Theorem 1 in Section 4 because of the condition of nonsingularity.

If we apply distance correlation to the copula $C(u;v) = F_{X,Y}(F_X^{-1}(u); F_Y^{-1}(v))$ where F_X^{-1}, F_Y^{-1} denote generalized inverse functions then we get a monotone invariant version of $\mathcal{R}$. For the sample distance correlation this means that we compute the distance correlation of the ranks.

Example 4 (Maximal correlation) Maximal correlation is defined as $\sup_{f,g} \rho(f(X), g(Y))$ where f, g are Borel-measurable functions. It was first introduced by Hirschfeld (1935) and Gebelein (1941), and then studied by Rényi (1959). Recently it has become increasingly popular, see Papadatos and Xifara (2013), Papadatos (2014), López Blázquez and Salamanca Miño (2014), Huang and Zhu (2016). Maximal correlation satisfies (i), (ii), and (iii), but, as we shall see in Theorem 1, it cannot satisfy (iv) because maximal correlation is invariant with respect to all 1–1 Borel functions on the real line.

Example 5 (Correlation ratio) The correlation ratio was introduced by Karl Pearson as part of analysis of variance. The definition of the correlation ratio is the following. $\Delta(X,Y) := \operatorname{Var}(E(X|Y))/\operatorname{Var}(X)$ provided that $\operatorname{Var}(X)$ exists and is positive. This measure is clearly not symmetric in X, Y. On the other hand it is easy to show that the symmetric maximal correlation is the same as the square root of the maximal correlation ratio.

Proposition 2 *The maximal correlation of X, Y,*

$$\sup_{f,g} \rho(f(X), g(Y)),$$

where f, g are Borel-measurable functions, is equal to the square root of the supremum of the correlation ratio, which is the square root of

$$\sup_f \{\operatorname{Var}\{E(f(X)|Y) : \operatorname{Var} f(X) = 1\}.$$

Thus the supremum of a nonsymmetric measure of dependence became symmetric.

Proof For the proof we can suppose without loss of generality that f, g are such that $\operatorname{Var} f(X) = \operatorname{Var} g(Y) = 1$. Then by the Cauchy-Schwarz inequality

$$\begin{aligned}
\operatorname{cov}(f(X), g(Y)) &= E(\operatorname{cov}(f(X), g(Y)|Y) + \operatorname{cov}(E(f(X)|Y), E(g(Y)|Y)) \\
&= 0 + \operatorname{cov}(E(f(X)|Y), g(Y)) \\
&\le \sqrt{\operatorname{Var} E(f(X)|Y)}\sqrt{\operatorname{Var} g(Y)} \\
&= \sqrt{\operatorname{Var} E(f(X)|Y)}.
\end{aligned}$$

Equality holds if $g(Y) = aE(f(X)|Y)+b$ where a, b are constants, $a \neq 0$. Thus

$$\begin{aligned}\operatorname{maxcorr}^2(X,Y) &= \sup_{f,g} \rho^2(f(X), g(Y)) \\ &= \sup\{\operatorname{Var} E(f(X)|Y) : \operatorname{Var} f(X) \operatorname{Var} g(Y) = 1\}.\end{aligned}$$

The correlation ratio satisfies axiom (ii), but it does not satisfy (i), (iii), and (iv), see Proposition 3. On a multivariate generalization of the correlation ratio see Sampson (1984).

Example 6 (Maximal information coefficient) Denote by $I(X,Y)$ the mutual information between two discrete random variables X and Y taking finitely many values (x, y):

$$I(X,Y) := \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)},$$

where $p(x, y)$ is the probability that $(X, Y) = (x, y)$, $p(x) = P(X = x)$, and $p(y) = P(Y = y)$. The population maximal information coefficient (MIC_*) of a pair (X, Y) of random variables is defined as

$$\mathrm{MIC}_*(X,Y) = \sup_G \frac{I((X,Y)|_G)}{\log ||G||},$$

where

- G is a rectangular grid imposed on the support of (X, Y),
- $(X, Y)|_G$ denotes the discrete distribution induced by (X, Y) on the cells of G, and
- $||G||$ denotes the minimum of the number of rows and the number of columns of G.

See Reshef et al. (2016).

MIC_* is the population value of the maximal information coefficient statistics (MIC) introduced in Reshef et al. (2011). For comments see Speed (2011), Simon and Tibshirani (2011).

MIC_* is not invariant with respect to all measurable 1–1 functions of H. Unfortunately, even axiom (iii) may be violated if the cdf of the random variable X is not continuous. In addition, axiom (iv) is not satified (see Proposition 4).

4 Dependence measures and the new axioms

Our main result is the following.

Theorem 1 *Suppose S is a set of pairs of non-constant random variables and if $(X, Y) \in S$ then $(LX, MY) \in S$ for all affine transformations L, M of H. If the dependence measure $\Delta(X, Y)$ on S is invariant with respect to all affine transformations L, M of H where $\dim H > 1$ then axiom (iv) cannot hold.*

If $\dim H = 1$ *then affinity is the same as similarity and in this case distance correlation is affine invariant. On the other hand, if* $\Delta(X, Y)$ *is invariant with respect to all* 1–1 *Borel measurable functions of* H *then even if* $\dim H = 1$*, axiom (iv) cannot hold.*

Proof Suppose $\dim H > 1$. We will show that every continuous and affine invariant dependence measure must be constant, hence violating axiom (i).

Let X, Y, X^*, and Y^* be arbitrary H-valued random variables, bounded and nonconstant. We will show that $\Delta(X, Y) = \Delta(X^*, Y^*)$. We can suppose that X^* and Y^* do not have constant coordinates at all. Then, by scale invariance, for every real number $c \neq 0$ we have

$$\Delta\big((X_1, X_2, \dots), Y\big) = \Delta\big((cX_1, X_2, \dots), Y\big),$$

and by continuity, this remains true for $c = 0$. Similarly, we get the same if X_1 is replaced by X_1^*. Thus, X_1 can be changed to X_1^* with no effect on Δ. Gradually, all coordinates of X can be replaced by those of X^*, and then the same can be done with Y. Consequently, $\Delta(X, Y) = \Delta(X^*, Y^*)$. During these changes of coordinates we have to avoid making any of the random variables constant by changing one of their coordinates. This can be achieved if we first replace the constant coordinates by the corresponding nonconstant ones.

If $H = \mathbb{R}$, the real line, such a result cannot be true because on the real line affine transformations coincide with similarities. But if we require invariance with respect to all 1–1 Borel measurable functions, then, as a first step, we can map our scalar random variables to $\mathbb{R}^2$ with the help of a 1–1 Borel measurable function (Gouvêa, 2011), then the reasoning above can be applied.

Let us note that for real valued random variables monotone invariance does not contradict continuity, this can be seen from Kimeldorf and Sampson (1978) or from distance correlation applied to ranks.

The next result is a corollary of Theorem 1, but because of its importance we state and prove this corollary separately.

Corollary 1 *The maximal correlation coefficient does not satisfy axiom (iv).*

In fact, it can happen that the maximal correlation coefficient of X_n, Y_n is 1 for $n = 1, 2, \dots$, but in their weak limit (X, Y) the random variables X and Y are independent, and by this their maximal correlation is 0.

Proof Suppose that the real valued random variables (X_n, Y_n) are such that $X_n = Y_n$ with probability $1/n$ and with the remaining $1-1/n$ probability $X_n = X$, $Y_n = Y$, where X and Y are independent. Then the maximal correlation of X_n, Y_n is 1 while in the limit they are independent.

Remark 2 More invariance of Δ is not necessarily better. For example maximal correlation satisfies axiom (F), i.e., maximal correlation is invariant with respect to all 1–1 transformations of the real line onto itself, but then this implies that the empirical maximal correlation is essentially always 1 because if

$X_1, X_2, \ldots, X_n$ are distinct real numbers then we can always find a 1–1 transformation f of the real line such that $Y_i = f(X_i), i = 1, 2, \ldots, n$. On the other hand, because distance correlation is not invariant for "most" 1–1 transformations we can apply $\sup_{f,g} \mathcal{R}(f(X), g(Y))$ where f, g are arbitrary functions for which $\mathcal{R}(f(X), g(Y))$ exists to detect "hidden dependencies" between X and Y. This can easily happen when X, Y are high dimensional vectors, most of their coordinates are independent and thus $\mathcal{R}(X, Y)$ is very small, but e.g. the first coordinate of X is always the same as the first coordinate of Y. This strong lower dimensional dependency is masked by the independence of other coordinates. Maximal distance correlation, i.e. $\sup_{f,g} \mathcal{R}(f(X), g(Y))$, can reveal this hidden dependency. Even if we maximize the sample distance correlation with respect to linear functions only (linear combinations of the coordinates) of (high dimensional) X and Y, we get a powerful dimension reduction tool, a "distance" counterpart of canonical correlation analysis.

Proposition 3 *The correlation ratio satisfies axiom (ii), but it does not satisfy (i), (iii), and (iv).*

Proof The correlation ratio does not satify axiom (i). Although it is zero when X and Y are independent, it can be zero in other cases, too; for example, when the conditional distribution of X given Y is symmetric. Axiom (*ii*) clearly holds, because on the real line similarities coincide with linear transformations. On the other hand, axiom (iii) does not hold because for the correlation ratio $\Delta(X, Y) = 1$ if and only if X is almost surely equal to a Borel measurable function of Y. Indeed, since $1 - \Delta(X, Y) = E[\mathrm{Var}(X|Y)]/\mathrm{Var}(X)$, we have $\Delta(X, Y) = 1$ if and only if $\mathrm{Var}(X|Y) = 0$; that is, X is a Borel measurable function os Y with probability 1. Axiom (iv) does not hold either as shown by the following example. Let Y be a nondegenerate, bounded, integer valued random variable and X be uniformly distributed on the interval $(0, 1)$ such that X is independent of Y. Define $Y_n = Y + \frac{1}{n}X$. Then $X = n\{Y_n\}$, where $\{.\}$ stands for fractional part, hence $\Delta(X, Y_n) = 1$. On the other hand, (X, Y_n) tends to (X, Y) everywhere, not only in distribution, and $\Delta(X, Y) = 0$.

Proposition 4 *The population maximal information coefficient* MIC_* *satisfies axioms (i) and (ii) but does not satisfy axioms (iii) and (iv).* MIC_* *is invariant with respect to all monotone transformations but not to all measurable* 1–1 *functions, hence Theorem 1 does not apply to* MIC_*.

Proof It is clear that $\mathrm{MIC}_*(X, Y) = 0$ if X and Y are independent. On the other hand, $\mathrm{MIC}_*(X, Y) = 0$ means that $I((X, Y)|_G) = 0$ for every grid G, which implies that the discretized by G versions of X and Y are independent. Particularly we obtain that the joint distribution of (X, Y) coincides with a product measure on rectangles, hence on all bidimensional Borel sets, too.

Since monotone transformations of the coordinates map grids into grids, MIC_* is invariant with respect to them. Particularly, it satisfies axiom (ii), because every affine transformation on $\mathbb{R}$ is monotone.

If X is discrete, then $\mathrm{MIC}_*(X, X) = 1$ if and only if there exists a partition of the real line into at least 2 parts such that X falls into every partition interval

with the same probability. For example, suppose that $P(X = 0) = p \neq 1/2$, and $P(X = 1) = 1 - p$. Then $\mathrm{MIC}_*(X, X) = -p \log p - (1 - p) \log(1 - p) < 1$. Thus, axiom (iii) is hurt.

Let X take the values $0, 1, 2$ with probabilities $1/4$, $1/2$, $1/4$, respectively. For computing $\mathrm{MIC}_*(X, X)$ it is easy to see that there exist altogether three essentially different grids: 2×2, 2×3, and 3×3. Consequently,

$$\begin{aligned}\mathrm{MIC}_*(X, X) &= \max\left\{ \frac{\frac{1}{4}\log 4 + \frac{3}{4}\log\frac{4}{3}}{\log 2}\,,\ \frac{\frac{1}{4}\log 4 + \frac{1}{2}\log 2 + \frac{1}{4}\log 4}{\log 3} \right\} \\ &= 0.946\cdots < 1.\end{aligned}$$

Let $f(x) = 3 - x$, if $1 \leq x \leq 2$, and $f(x) = x$ otherwise. This f interchanges 1 and 2, and does not move 0. It is a piecewise continuous 1–1 function, and $\mathrm{MIC}_*(f(X), f(X)) = 1$, which is obtained by considering the partition $\mathbb{R} = (-\infty, 3/2) \cup [3/2, +\infty)$. Thus, MIC_* is not invariant with respect to measurable 1–1 transformations.

If one prefers a counterexample with continuous joint distribution, let (X, Y) be uniformly distributed over the black squares of a 4×4 checkerboard. Then the maximal information coefficient of that distribution is $1/2$, but if we permute the rows and columns of the checkerboard, we can turn it into a 2×2 checkerboard with bigger squares, and the maximal information coefficient of that distribution is 1. This permutation can be performed using piecewise linear functions. Details are left to the reader.

Finally, we show that MIC_* is not weakly continuous, thus axiom (iv) is not satisfied.

Let X be uniformly distributed on the interval $[0, 1]$, and define Y as the fractional part of nX. We will show that $\mathrm{MIC}_*(X, Y) = 1$ for every positive integer n.

Clearly, Y is also uniformly distributed on $[0, 1]$. Let $k \geq 2$ be arbitrary, and impose an $nk \times k$ equidistant grid on the unit square. Then the distribution $(X, Y)|_G$ is discrete uniform of size nk, with discrete uniform marginals of size nk and k, respectively. Hence

$$I((X, Y)|_G) = \log(nk) + \log k - \log(nk) = \log k,$$

while $||G|| = k$. (In information theory, logarithm is meant on base 2, but the base does not matter here.) Thus,

$$\mathrm{MIC}_*(X, Y) = \frac{I((X, Y)|_G)}{\log ||G||} = 1.$$

Now, as $n \to \infty$, the joint distribution of (X, Y) converges weakly to the uniform distribution on the unit square, which, having independent marginals, yields $\mathrm{MIC}_* = 0$.

In light of Corollary 1 and Proposition 4 it is unlikely that the maximal correlation or the maximal information coefficient will be the correlation for the 21st century; see Speed (2011). Distance correlation on the other hand turned

out to be a new powerful tool for detection of associations between data sets, see the summary of a plenary talk at the Joint Mathematical Meeting in 2017 (Richards, 2017).

5 Conclusion

There are many examples of dependence measures that satisfy our new axioms (i) – (iv) for different important sets S but if we want S to contain all pairs of bounded (nonconstant) random variables in an arbitrary Euclidean space or separable Hilbert space, then of the well-known dependence measures, the distance correlation seems to be the simplest and most appealing one that satisfies all axioms (i) – (iv).

6 Appendix

Proof of Proposition 1 Without loss of generality assume that $E[X] = 0$. Let Q denote the distribution of X on the Borel sets of $\mathbb{R}$. We have to find a 1–1 function f such that $\int xf(x)\,dQ = 0$.

By assumption, there exist real numbers $t_1 < t_2 < t_3$ such that each of the intervals $(-\infty, t_1]$, $(t_1, t_2]$, $(t_2, t_3]$, $(t_3, +\infty)$ has positive measure (w.r.t. Q). Let δ be a suitably small positive number (the meaning of "suitably" will be made clear later). One can find $t_0 < t_1$ and $t_4 > t_3$ such that both $Q(-\infty, t_0]$ and $Q(t_4, +\infty)$ are less than δ (possibly 0).

Let the intervals $(-\infty, t_0]$, $(t_0.t_1]$, $(t_1, t_2]$, $(t_2, t_3]$, $(t_3, t_4]$, and $(t_4, +\infty)$ be denoted by A_0, A_1, A_2, A_3, A_4, and A_5, respectively. Introduce

$$\mu_i = \int_{A_i} x\,dQ, \quad \sigma_i^2 = \int_{A_i} x^2 dQ, \quad 0 \le i \le 5.$$

Then $\mu_0 + \cdots + \mu_5 = 0$.

It is not hard to see that there exist real constants a_1, a_2, a_3, a_4, all different, such that

$$a_1(\mu_0 + \mu_1) + a_2\mu_2 + a_3\mu_3 + a_4(\mu_4 + \mu_5) = 0. \tag{1}$$

Indeed, consider the hyperplane $\mathcal{L}$ of all vectors $(a_1, a_2, a_3, a_4) \in \mathbb{R}^4$ satisfying (1). $\mathcal{L}$ cannot coincide with the hyperplane $\mathcal{L}_{1,2} = \{a_1 = a_2\}$, because the $\mathcal{L}_{1,2}$ is orthogonal to the vector $(1, -1, 0, 0)$, which is not parallel to $(\mu_0 + \mu_1, \mu_2, \mu_3, \mu_4 + \mu_5)$, since the latter can have at most one 0 coordinate. Thus, $\dim(\mathcal{L} \cap \mathcal{L}_{1,2}) = 2$. The same holds for $\mathcal{L}_{i,j}$, the hyperplane defined by equality $a_i = a_j$ $(i \neq j)$. Since $\mathcal{L}$ cannot be covered by six of its lower dimensional subspaces, the existence of a vector in $\mathcal{L}$ with different coordinates follows.

Let $K > \max_{1 \le i \le 4} |a_i|$. By continuity, if δ is small enough, one can find constants b_1, b_2, b_3, b_4 all different, such that $\max_{1 \le i \le 4} |b_i| < K$, and

$$-K\mu_0 + b_1\mu_1 + b_2\mu_2 + b_3\mu_3 + b_4\mu_4 + K\mu_5 = 0.$$

Finally, choose $c_0, c_1, \ldots, c_5$ in such a way that none of them are equal to 0, c_0 and c_5 are positive, and $\sum_{i=0}^{5} c_i \sigma_i^2 = 0$. This can be done, because there are at least 3 positive among the quantities σ_i^2.

Now, let $b_0 = -K$, $b_5 = K$, and $f(x) = b_i + \varepsilon c_i x$ if $x \in A_i$, $0 \le i \le 5$. Then f is injective provided ε is a sufficiently small positive number, and

$$\int_{\mathbb{R}} x f(x)\, dQ = \sum_{i=0}^{5} (b_i \mu_i + \varepsilon c_i \sigma_i^2) = 0,$$

as needed.

Such an f cannot exist if X can take on exactly two values, because in that case uncorrelatedness is equivalent to independence.

When the distribution of X is concentrated on exactly 3 points, and X is supposed to have mean 0, then such an f exists if and only if zero is not among the possible values of X. (If $E[X] = 0$ is not supposed, the necessary and sufficient condition for f to exist is $P(X = E[X]) = 0$.) Indeed, let $x_1 < x_2 < x_3$ be the possible values of X, with probabilities q_1, q_2, q_3, respectively. Then $q_1 x_1 + q_2 x_2 + q_3 x_3 = 0$, and $x_1 < 0 < x_3$. We are looking for real numbers f_1, f_2, f_3 such that $q_1 x_1 f_1 + q_2 x_2 f_2 + q_3 x_3 f_3 = 0$. If $x_2 = 0$, then it can only achieved with $f_1 = f_3$. In the complementary case $f_1 = -1$, $f_3 = 1$ and $f_2 = (q_1 x_1 - q_3 x_3)/(q_2 x_2)$ will do, because $f_2 = 1$ would imply $-q_1 x_1 + q_2 x_2 + q_3 x_3 = 0$, hence $q_1 x_1 = 0$, which is not allowed, and similarly, $f_2 = -1$ would imply $q_3 x_3 = 0$.

References

Bickel PJ, Xu Y (2009) Discussion of: Brownian Distance Covariance. Ann Appl Stat 3:1266–1269. https://doi.org/10.1214/09-AOAS312A

Dedecker J, Prieur C (2005) New Dependence Coefficients. Examples and Applications to Statistics. Probab Theory Relat Fields 132:203–236. https://doi.org/10.1007/s00440-004-0394-3

Dueck J, Edelmann D, Gneiting T, Richards, D (2014) The Affinely Invariant Distance Correlation. Bernoulli 20:2305–2330. https://doi.org/10.3150/13-BEJ558.

Eaton ML (1989) Group Invariance. Applications in Statistics, NSF-CBMS Regional Conference Series in Probability and Statistics 1. IMS, Hayward, CA.

Escoufier Y (1973) Le Traitement des Variables Vectorielles. Biometrics 29:751–760. https://doi.org/10.2307/2529140

Feuerverger A (1993) A Consistent Test for Bivariate Dependence. Int Stat Rew 61:419–433. https://doi.org/10.2307/1403753

Gebelein, H (1941) Das statistische Problem der Korrelation als Variations- und Eigenwert-problem und sein Zusammenhang mit der Ausgleichungsrechnung. Z Angew Math Mech 21:364–379. https://doi.org/10.1002/zamm.19410210604

Gouvêa FQ (2011) Was Cantor Surprised? Am Math Mon 118:198–209. https://doi.org/10.4169/amer.math.monthly.118.03.198
Hirschfeld HO (1935) A Connection Between Correlation and Contingency. Math Proc Camb Philos Soc 31:520–524. https://doi.org/10.1017/S0305004100013517
Hoeffding W (1940) Masstabinvariante Korrelationstherie. Schr Math Inst und Inst Angew Math Univ Berlin 5:181–233.
Hoeffding W (1948) A Non-Parametric Test of Independence. Ann Math Stat 19:546–557. https://doi.org/10.1214/aoms/1177730150
Huang Q, Zhu Y (2016) Model-Free Sure Screening Via Maximum Correlation. J Multivar Anal 148:89–106. 10.1016/j.jmva.2016.02.014
Jakobsen ME (2017) Distance Covariance in Metric Spaces: Non-Parametric Independence Testing in Metric Spaces. https://arxiv.org/pdf/1706.03490. Accessed 9 Jan 2018
Josse J, Holmes S (2014) Tests of Independence and Beyond. https://arxiv.org/pdf/1307.7383v3. Accessed 9 Jan 2018
Kendall MG (1938) A New Measure of Rank Correlation. Biometrika 30:81–93. https://doi.org/10.2307/2332226
Kimeldorf G, Sampson AR (1978) Monotone Dependence. Ann Stat 6:895–903. https://doi.org/10.1214/aos/1176344262
Lehmann EL (1966) Some Concepts of Dependence. Ann Math Stat 37:1137–1153. https://doi.org/10.1214/aoms/1177699260
Lehmann EL, Romano JP (2005) Testing Statistical Hypotheses (3rd ed). Springer, New York. https://doi.org/10.1007/0-387-27605-X
Linfoot EH (1957) An Informational Measure of Correlation. Inf Control 1:85–89. https://doi.org/10.1016/S0019-9958(57)90116-X
López Blázquez F, Salamanca Miño B (2014) Maximal Correlation in a Non-Diagonal Case. J Multivar Anal 131:265–278. https://doi.org/10.1016/j.jmva.2014.07.008
Lyons R (2013) Distance Covariance in Metric Spaces. Ann Probab 41:3284–3305. https://doi.org/10.1214/12-AOP803
Papadatos N (2014) Some Counterexamples Concerning Maximal Correlation and Linear Regression. J Multivar Anal 126:114–117. https://doi.org/10.1016/j.jmva.2013.12.008
Papadatos N, Xifara T (2013) A Simple Method for Obtaining the Maximal Correlation Coefficient and Related Characterizations. J Multivar Anal 118:102–114. https://doi.org/10.1016/j.jmva.2013.03.017
Pearson K (1920) Notes on the History of Correlation. Biometrika 13:25–45. https://doi.org/10.2307/2331722
Reimherr M, Nicolae DL (2013) On Quantifying Dependence: A Framework For Developing Interpretable Measures. Stat Sci 28:116–130. https://doi.org/10.1214/12-STS405
Rényi A (1959) On Measures of Dependence. Acta Mat Acad Sci Hung 10:441–451. https://doi.org/10.1007/BF02024507
Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Detect-

ing Novel Associations in Large Data Sets. Science 334/6062:1518–1524. https://doi.org/10.1126/science.1205438

Reshef YA, Reshef DN, Finucane HK, Sabeti PC, Mitzenmacher M (2016) Measuring Dependence Powerfully and Equitably. J Mach Learn Res 17(212):1–63.

Richards DStP (2017) Distance Correlation: A New Tool for Detecting Association and Measuring Correlation Between Data Sets. Plenary talk at the Joint Mathematics Meeting, Atlanta, 2017. Not Am Math Soc 64:16–18. https://doi.org/10.1090/noti1457

Sampson AR (1984) A Multivariate Correlation Ratio. Stat Probab Lett 2:77–81. https://doi.org/10.1016/0167-7152(84)90054-3

Schweizer B, Wolff EF (1981) On Nonparametric Measures of Dependence for Random Variables. Ann Stat 9:879–885. https://doi.org/10.1214/aos/1176345528

Sejdinovic D, Sriperumbudur B, Gretton A, Fukumiyu K (2013) Equivalence of Distance-Based and RKHS-based Statistics in Hypothesis Testing. Ann Stat 41:2263–2291. https://doi.org/10.1214/13-AOS1140

Simon N, Tibshirani R (2011) Comment on "Detecting Novel Associations in Large Data Set" by Reshef et al, Science Dec 16, 2011. https://arxiv.org/pdf/1401.7645v1. Accessed 9 Jan 2018

Spearman C (1904) A Proof and Measurement of Association Between Two Things. Am J Psychol 15:72–101. https://doi.org/10.2307/1412159

Speed T (2011) A Correlation for the 21st Century. Science 334/6062:1502–1503. https://doi.org/10.1126/science.1215894

Stigler S (1989) Francis Galton's Account of the Invention of Correlation. Stat Sci 4:73–79. https://doi.org/10.1214/ss/1177012580

Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and Testing Independence by Correlation of Distances. Ann Stat 35:2769–2794. https://doi.org/10.1214/009053607000000505

Székely GJ, Rizzo ML (2009) Brownian Distance Covariance. Ann Appl Stat 3:1236–1265. https://doi.org/10.1214/09-AOAS312

Székely GJ, Rizzo ML (2014) Partial Distance Correlation With Methods for Dissimilarities. Ann Stat 42:2382–2412. https://doi.org/10.1214/14-AOS1255

Volokh E (2015) Zero Correlation Between State Homicide and State Gun Laws. The Washington Post, October 6, 2015