J. Appl. Probab. **60,** 1115–1135 (2023)
*doi:*10.1017/jpr.2022.118

THE BERKELMANS–PRIES DEPENDENCY FUNCTION: A GENERIC MEASURE OF DEPENDENCE BETWEEN RANDOM VARIABLES

GUUS BERKELMANS,[*] [**] *Centrum Wiskunde & Informatica*
SANDJAI BHULAI,[***] *Vrije Universiteit (VU)*
ROB VAN DER MEI,[****] *Centrum Wiskunde & Informatica AND Vrije Universiteit*
JORIS PRIES,[*] [*****] *Centrum Wiskunde & Informatica*

Abstract

Measuring and quantifying dependencies between random variables (RVs) can give critical insights into a dataset. Typical questions are: 'Do underlying relationships exist?', 'Are some variables redundant?', and 'Is some target variable Y highly or weakly dependent on variable X?' Interestingly, despite the evident need for a general-purpose measure of dependency between RVs, common practice is that most data analysts use the Pearson correlation coefficient to quantify dependence between RVs, while it is recognized that the correlation coefficient is essentially a measure for *linear* dependency only. Although many attempts have been made to define more generic dependency measures, there is no consensus yet on a standard, general-purpose dependency function. In fact, several ideal properties of a dependency function have been proposed, but without much argumentation. Motivated by this, we discuss and revise the list of desired properties and propose a new dependency function that meets all these requirements. This general-purpose dependency function provides data analysts with a powerful means to quantify the level of dependence between variables. To this end, we also provide Python code to determine the dependency function for use in practice.

Keywords: Probability theory; measure theory; distributions; association; correlation

2020 Mathematics Subject Classification: Primary 62H20
Secondary 60A10; 62H05

1. Introduction

As early as 1958, Kruskal [14] stated that 'There are infinitely many possible measures of association, and it sometimes seems that almost as many have been proposed at one time or another'. Many years later, even more dependency measures have been suggested. Yet, and rather surprisingly, there still does not exist consensus on a general dependency function. Often the statement 'Y is dependent on X' means that Y is not independent of X. However, there are

Received 4 February 2022; revision received 14 October 2022.
[*] Postal address: Department of Stochastics, P.O. Box 94079, 1090 GB Amsterdam, Netherlands
[**] Email address: gberkelmans@cwi.nl
[***] Postal address: Department of Mathematics, De Boelelaan 1111, 1081 HV Amsterdam, Netherlands. Email: s.bhulai@vu.nl
[****] Postal address: Department of Stochastics, Science Park 123, 1098 XG Amsterdam, Netherlands. Email: mei@cwi.nl
[*****] Email address: joris.pries@cwi.nl

different levels of dependency. For example, random variable (RV) Y can be fully determined by RV X (i.e. $Y(\omega) = f(X(\omega))$ for all $\omega \in \Omega$ and for a measurable function f), or only partially.

But how should we quantify how much Y is dependent on X? Intuitively, and assuming that the dependency measure is normalized to the interval [0, 1], we would say that if Y is fully determined by X then the dependency of Y with respect to X is as strong as possible, and so the dependency measure should be 1. On the other side of the spectrum, if X and Y are independent, then the dependency measure should be 0; and, vice versa, it is desirable that dependence 0 *implies* that X and Y are stochastically independent. Note that the commonly used *Pearson correlation coefficient* does not meet these requirements. In fact, many examples exist where Y is fully determined by X while the correlation is zero.

Taking a step back, why is it actually useful to examine dependencies in a dataset? Measuring dependencies between the variables can lead to critical insights, which will lead to improved data analysis. First of all, it can reveal important explanatory relationships. How do certain variables interact? If catching a specific disease is highly dependent on the feature value of variable X, research should be done to investigate if this information can be exploited to reduce the number of patients with this disease. For example, if hospitalization time is dependent on a healthy lifestyle, measures can be taken to try to improve the overall fitness of a population. Dependencies can therefore function as an actionable steering rod. It is, however, important to keep in mind that dependency does not always mean causality. Dependency relations can also occur due to mere coincidence or as a by-product of another process.

Dependencies can also be used for dimensionality reduction. If Y is highly dependent on X, not much information is lost when only X is used in the dataset. In this way, redundant variables or variables that provide little additional information can be removed to reduce the dimensionality of the dataset. With fewer dimensions, models can be trained more efficiently.

In these situations a dependency function can be very useful. However, finding the proper dependency function can be hard, as many attempts have already been made. In fact, most of us have a 'gut feeling' for what a dependency function should entail. To make this feeling more mathematically sound, Rényi [18] proposed a list of ideal properties for a dependency function. A long list of follow-up papers (see the references in Table 1) use this list as the basis for a wish list, making only minor changes to it, adding or removing some properties.

In view of the above, the contribution of this paper is threefold:

- We determine a new list of ideal properties for a dependency function.
- We present a new dependency function and show that it fulfills all requirements.
- We provide Python code to determine the dependency function for the discrete and continuous case (https://github.com/joris-pries/BP-Dependency).

The remainder of this paper is organized as follows. In Section 2, we summarize which ideal properties have been stated in previous literature. By critically assessing these properties, we derive a new list of ideal properties for a dependency function (see Table 2) that lays the foundation for a new search for a general-purpose dependency function. In Section 3, the properties are checked for existing methods, and we conclude that there does not yet exist a dependency function that has all the desired properties. Faced by this, in Section 4 we define a new dependency function and show in Section 5 that this function meets all the desired properties. Finally, Section 6 outlines the general findings and addresses possible future research opportunities.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

TABLE 1. A summary of desirable properties for a dependency function as stated in previous literature.

Property group	Property		Article(s)
Range	**I.1**	$0 \leq \mathrm{Dep}(X, Y) \leq 1$	[1, 4, 7, 8, 12, 18–21]
	I.2	$\mathrm{Dep}(X, Y) = 0 \Leftarrow X$ and Y are independent	[7, 12, 19]
	I.3	$\mathrm{Dep}(X, Y) = 0 \Rightarrow X$ and Y are independent	[21]
	I.4	$\mathrm{Dep}(X, Y) = 0 \Leftrightarrow X$ and Y are independent	[1, 4, 8, 16, 18, 20]
	I.5	$\mathrm{Dep}(X, Y) = 1 \Leftrightarrow Y = LX$ with probability 1, where L is a similarity transformation	[16]
	I.6	$\mathrm{Dep}(X, Y) = 1 \Leftarrow X$ and Y are strictly dependent	[1, 7, 18, 19]
	I.7	$\mathrm{Dep}(X, Y) = 1 \Leftrightarrow X$ and Y are comonotonic or countermonotonic	[4]
	I.8	$\mathrm{Dep}(X, Y) = 1 \Leftrightarrow X$ and Y are strictly dependent	[8]
General	**I.9**	$\mathrm{Dep}(X, Y)$ is defined for any X, Y where both are not constant	[8, 16, 18]
	I.10	Well-defined for both continuous and discrete variables	[7]
	I.11	Defined for both categorical and continuous variables, and for ordinal categorical variables for which there may be underlying continuous variables	[12]
	I.12	There is a close relationship between the measure for the continuous variables and the measure for the discretization of the variables	[12]
Symmetric	**I.13**	$\mathrm{Dep}(X, Y) = \mathrm{Dep}(Y, X)$	[1, 4, 18–20]
Applying function to argument	**I.14**	$\mathrm{Dep}(f(X), g(Y)) = \mathrm{Dep}(X, Y)$ with f, g strictly monotonic functions	[1]
	I.15	$\mathrm{Dep}(f(X), Y) = \mathrm{Dep}(X, Y)$ with $f : \mathbb{R} \to \mathbb{R}$ strictly monotonic on the range of X	[4]
	I.16	$\mathrm{Dep}(f(X), f(Y)) = \mathrm{Dep}(X, Y)$ with f continuous and strictly increasing	[7, 20]
	I.17	$\mathrm{Dep}(f(X), g(Y)) = \mathrm{Dep}(X, Y)$ if $f(\cdot)$, $g(\cdot)$ map the real axis in a one-to-one way onto itself	[12, 18]
	I.18	$\mathrm{Dep}(X, Y)$ is invariant with respect to all similarity transformations	[16]

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

TABLE 1. Continued.

Property group	Property		Article(s)
	I.19	$\mathrm{Dep}(X, Y)$ is invariant with respect to translation and scaling	[20]
	I.20	$\mathrm{Dep}(X, Y)$ is scale invariant	[21]
	I.21	$\mathrm{Dep}(X, Y)$ is a function of the Pearson correlation if the joint distribution of X and Y is normal	[1, 7, 21]
Behavior normal distribution	**I.22**	$\mathrm{Dep}(X, Y) = \lvert\rho(X, Y)\rvert$ if the joint distribution of X and Y is normal, where ρ is the Pearson correlation	[12, 18]

TABLE 2. New list of desirable properties for a dependency function.

Property group	Property	
Asymmetric	**II.1**	There exist RVs X, Y such that $\mathrm{Dep}(Y \mid X) \neq \mathrm{Dep}(X \mid Y)$
	II.2	$0 \leq \mathrm{Dep}(Y \mid X) \leq 1$ for all RVs X and Y
	II.3	$\mathrm{Dep}(Y \mid X) = 0 \Leftrightarrow X$ and Y are independent
	II.4	$\mathrm{Dep}(Y \mid X) = 1 \Leftarrow Y$ is strictly dependent on X
Intuitive	**II.5**	If $Y_1, Y_2, \ldots, Y_N, S$ are independent with $\mathbb{P}(S \in [N]) = 1$, $\mathbb{P}(S = i) = p_i$, and $X = Y_S$ then $\mathrm{Dep}(Y_i \mid X) = p_i$ must hold
General	**II.6**	Applicable for any combination of continuous, discrete, and categorical RVs X, Y, where Y is not almost surely (a.s.) constant
	II.7	$\mathrm{Dep}(g(Y) \mid f(X)) = \mathrm{Dep}(Y \mid X)$ for any isomorphisms f, g
Functions	**II.8**	$\mathrm{Dep}(Y \mid f(X)) \leq \mathrm{Dep}(Y \mid X)$ for any measurable function f

2. Desired properties of a dependency function

What properties should an ideal dependency function have? In this section we summarize previously suggested properties. Often, these characteristics are posed without much argumentation. Therefore, we analyze and discuss which properties are actually ideal and which properties are believed to be not relevant, or even wrong.

In Table 1, a summary is given of 22 'ideal properties' found in previous literature, grouped into five different categories. These properties are denoted by I.1–22. From these properties we derive a new set of desirable properties denoted by II.1–8; see Table 2. Next, we discuss the properties suggested in previous literature and how the new list is derived from them.

Desired property II.1. *(Asymmetry.)* At first glance, it seems obvious that a dependency function should adhere to property I.13 and be symmetric. However, this is a common misconception for the dependency function. Y can be fully dependent on X, but this does not mean that X is fully dependent on Y. Lancaster [15] indirectly touched upon this same point by

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

defining *mutual complete dependence*. First it is stated that Y is *completely dependent* on X if $Y = f(X)$. X and Y are called *mutually completely dependent* if X is completely dependent on Y and vice versa. Thus, this indirectly shows that dependence should not necessarily be symmetric, otherwise the extra definition would be redundant. In [15] the following great asymmetric example was given.

Example 2.1. Let $X \sim \mathcal{U}(0, 1)$ be uniformly distributed and let $Y = -1$ if $X \leq \frac{1}{2}$ and $Y = 1$ if $X > \frac{1}{2}$. Here, Y is fully dependent on X, but not vice versa.

To drive the point home even more, we give another asymmetric example.

Example 2.2. X is uniformly randomly drawn from $\{1, 2, 3, 4\}$, and $Y := X \,(\mathrm{mod}\, 2)$. Y is fully dependent on X, because given X the value of Y is deterministically known. On the other hand, X is not completely known given Y. Note that $Y = 1$ still leaves the possibility for $X = 1$ or $X = 3$. Thus, when assessing the dependency between variable X and variable Y, Y is fully dependent on X, whereas X is not fully dependent on Y. In other words, $\mathrm{Dep}(X, Y) \neq \mathrm{Dep}(Y, X)$.

In conclusion, *an ideal dependency function should not always be symmetric*. To emphasize this point even further, we change the notation of the dependency function. Instead of $\mathrm{Dep}(X, Y)$, we will write $\mathrm{Dep}(Y \mid X)$ for how much Y is dependent on X. Based on this, property I.13 is changed into II.1.

Desired property II.2. *(Range.)* An ideal dependency function should be scaled to the interval [0, 1]. Otherwise, it can be very hard to draw meaningful conclusions from a dependency score without a known maximum or minimum. What would a score of 4.23 mean without any information about the possible range? Therefore, property I.1 is retained. A special note on the range for the well-known Pearson correlation coefficient [17], which is $[-1, 1]$: The negative or positive sign denotes the direction of the linear correlation. When examining more complex relationships, it is unclear what 'direction' entails. We believe that a dependency function should measure by *how much* variable Y is dependent on X, and not necessarily in which way. In summary, we require $0 \leq \mathrm{Dep}(Y \mid X) \leq 1$.

Desired property II.3. *(Independence and dependency* 0*.)* If Y is independent of X, it should hold that the dependency achieves the lowest possible value, namely zero. Otherwise, it is vague what a dependency score lower than the dependency between two independent variables means. A major issue of the commonly used Pearson correlation coefficient is that zero correlation does not imply independence. This makes it complicated to derive conclusions from a correlation score. Furthermore, note that if Y is independent of X, it should automatically hold that X is also independent of Y. In this case, X and Y are independent, because otherwise some dependency relation should exist. Thus, we require $\mathrm{Dep}(Y \mid X) = 0 \iff X$ and Y are independent.

Desired property II.4. *(Functional dependence and dependency* 1*.)* If Y is strictly dependent on X (and thus fully determined by X), the highest possible value should be attained. It is otherwise unclear what a higher dependency would mean. However, it is too restrictive to demand that the dependency is only 1 if Y is strictly dependent on X. Rényi [18] stated 'It seems at the first sight natural to postulate that $\delta(\xi, \eta) = 1$ only if there is a strict dependence of the mentioned type between ξ and η, but this condition is rather restrictive, and it is better to leave it out'. Take, for example, $Y \sim \mathcal{U}(-1, 1)$ and $X := Y^2$. Knowing X reduces the infinite

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

set of possible values for Y to only two ($\pm\sqrt{X}$), whereas it would reduce to one if Y was fully determined by X. It would be very restrictive to enforce $\mathrm{Dep}(Y \mid X) < 1$, as there is only an infinitesimal difference compared to the strictly dependent case. Summarizing, we require $Y = f(X) \to \mathrm{Dep}(Y \mid X) = 1$.

Desired property II.5. *(Unambiguity.)* Kruskal [14] stated 'It is important to recognize that the question "Which single measure of association should I use?" is often unimportant. There may be no reason why two or more measures should not be used; the point I stress is that, whichever ones are used, they should have clear-cut population interpretations.' It is very important that a dependency score leaves no room for ambiguity. The results should meet our natural expectations. Therefore, we introduce a new requirement based on a simple example. Suppose we have a number of independent RVs and observe one of these at random. The dependency of each random variable on the observed variable should be equal to the probability it is picked. More formally, let $Y_1, Y_2, \ldots, Y_N$, and S be independent variables, with S a selection variable such that $\mathbb{P}(S = i) = p_i$ and $\sum_{i=1}^{N} p_i = 1$. When X is defined as $X = \sum_{i=1}^{N} \mathbf{1}_{S=i} \cdot Y_i$, it should hold that $\mathrm{Dep}(Y_i \mid X) = p_i$ for all $i \in \{1, \ldots, N\}$. Stated simply, the dependency function should give the desired results in specific situations where we can argue what the outcome should be. This is one of these cases.

Desired property II.6. *(Generally applicable.)* Our aim is to find a general dependency function, which we denote by $Dep(X \mid Y)$. This function must be able to handle all kinds of variables: *continuous*, *discrete*, and *categorical* (even nominal). These types of variables occur frequently in a dataset. A general dependency function should be able to measure the dependency of a categorical variable Y on a continuous variable X. Stricter than I.9–12, we want a single dependency function that is applicable to any combination of these variables.

There is one exception to this generality. In the case that Y is almost surely constant, it is completely independent as well as completely determined by X. Arguing what the value of a dependency function should be in this case is similar to arguing about the value of $\frac{0}{0}$. Therefore, we argue that in this case it should be either undefined or return some value that represents the fact that Y is almost surely constant (for example -1, since this cannot be normally attained).

Desired property II.7. *(Invariance under isomorphisms.)* Properties I.14–20 discuss when the dependency function should be invariant. Most are only meant for variables with an ordering, as 'strictly increasing', 'translation', and 'scaling' are otherwise ill-defined. As the dependency function should be able to handle nominal variables, we assume that the dependency is invariant under isomorphisms, see II.7. Note that this is a stronger assumption than I.14–20. Compare Example 2.2 with Example 2.3. It should hold that $\mathrm{Dep}(Y \mid X) = \mathrm{Dep}(Y' \mid X')$ and $\mathrm{Dep}(X \mid Y) = \mathrm{Dep}(X' \mid Y')$, as the relationship between the variables is the same (only altered using isomorphisms). So, for any isomorphisms f and g we require $\mathrm{Dep}(g(Y) \mid f(X)) = \mathrm{Dep}(Y \mid X)$.

Example 2.3. Let X' be uniformly randomly drawn from $\{\circ, \triangle, \square, \Diamond\}$, and $Y' = \clubsuit$ if $X' \in \{\circ, \square\}$ and $Y' = \spadesuit$ if $X' \in \{\triangle, \Diamond\}$.

Desired property II.8. *(Non-increasing under functions of X.)* Additionally, $\mathrm{Dep}(Y \mid X)$ should not increase if a measurable function f is applied to X since any dependence on $f(X)$ corresponds to a dependence on X (but not necessarily the other way around). The information gained from knowing X can only be reduced, never increased by applying a function.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

However, though it might be natural to expect the same for functions applied to Y, consider once again Example 2.2 (but with X and Y switched around) and the following two functions: $f_1(Y) := Y \pmod 2$ and $f_2(Y) := \lceil (Y/2) \rceil$. Then $f_1(Y)$ is completely predicted by X and should therefore have a dependency of 1, while $f_2(Y)$ is independent of X and should therefore have a dependency of 0. So the dependency should be free to increase or decrease for functions applied to Y. To conclude, for any measurable function f we require $\mathrm{Dep}(Y \mid f(X)) \leq \mathrm{Dep}(Y \mid X)$.

2.1. Exclusion of Pearson correlation coefficient as a special case

According to properties I.21 and I.-22, when X and Y are normally distributed the dependency function should coincide with or be a function of the Pearson correlation coefficient. However, these properties lack good reasoning for why this would be ideal. It is not obvious why this would be a necessary condition. Moreover, there are many known problems and pitfalls with the correlation coefficient [4, 11], so it seems undesirable to force an ideal dependency function to reduce to a function of the correlation coefficient when the variables are normally distributed. This is why we exclude these properties.

3. Assessment of the desired properties for existing dependency measures

In this section we assess whether existing dependency functions have the properties listed above. In doing so, we limit ourselves to the most commonly used dependency measures. Table 3 shows which properties each investigated measure adheres to.

Although the desired properties listed in Table 2 seem not too restrictive, many dependency measures fail to have many of these properties. One of the most commonly used dependency measures, the Pearson correlation coefficient, does not even satisfy one of the desirable properties. Furthermore, almost all measures are not asymmetric. The one measure that comes closest to fulfilling all the requirements is the *uncertainty coefficient* [17]. This is a normalized asymmetric variant of the *mutual information* measure [17], where the discrete variant is defined as

$$C_{XY} = \frac{I(X, Y)}{H(Y)} = \frac{\sum_{x,y} p_{X,Y}(x, y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x) \cdot p_Y(y)}\right)}{-\sum_y p_Y(y) \log\left(p_Y(y)\right)},$$

where $H(Y)$ is the entropy of Y and $I(X, Y)$ is the mutual information of X and Y. Note that throughout the paper we use the following notation: $p_X(x) = \mathbb{P}(X = x)$, $p_Y(y) = \mathbb{P}(Y = y)$, and $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$. In addition, for a set H we define $p_X(H) = \mathbb{P}(X \in H)$ (and similarly for p_Y and $p_{X,Y}$).

However, the uncertainty coefficient does not satisfy properties II.5 and II.6. For example, if $Y \sim \mathcal{U}(0, 1)$ is uniformly drawn, the entropy of Y becomes

$$H(Y) = -\int_0^1 f_Y(y) \ln\left(f_Y(y)\right) \mathrm{d}y = -\int_0^1 1 \cdot \ln(1) \, \mathrm{d}y = 0.$$

Thus, for any X, the uncertainty coefficient is now undefined (division by zero). Therefore, the uncertainty coefficient is not as generally applicable as property II.6 requires.

Two other measures that satisfy many (but not all) properties are *mutual dependence* [1] and *maximal correlation* [5]. Mutual dependence is defined as the Hellinger distance [9]

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

TABLE 3. Properties of previous dependencies functions (✗ = property not satisfied, ✓ = property satisfied)

	Asymmetric	Intuitive				General	Functions	
Measure	**II.1**	**II.2**	**II.3**	**II.4**	**II.5**	**II.6**	**II.7**	**II.8**
Pearson correlation coefficient [17]	✗	✗	✗	✗	✗	✗	✗	✗
Spearman's rank correlation coefficient [17]	✗	✗	✗	✗	✗	✗	✗	✗
Kendall rank correlation coefficient [17]	✗	✗	✗	✗	✗	✗	✗	✗
Mutual information [17]	✗	✗	✓	✗	✗	✗	✓	✓
Uncertainty coefficient [17]	✓	✓	✓	✓	✗	✗	✓	✓
Total correlation [22]	✗	✗	✓	✗	✗	✗	✓	✓
Mutual dependence [1]	✗	✓	✓	✗	✗	✓	✓	✓
Δ_{L_1} [3]	✗	✗	✓	✗	✗	✓	✓	✓
Δ_{SD} [3]	✗	✗	✓	✗	✗	✓	✗	✗
Δ_{ST} [3]	✗	✗	✗	✗	✗	✓	✗	✗
Monotone correlation [13]	✗	✓	✓	✗	✗	✗	✗	✗
Maximal correlation [5]	✗	✓	✓	✓	✗	✓	✓	✓
Distance correlation [21]	✗	✓	✓	✗	✗	✗	✗	✗
Maximum canonical correlation (first) [10]	✗	✓	✗	✗	✗	✗	✗	✗
Strong mixing coefficient [2]	✗	✓	✓	✗	✗	✓	✓	✓
β-mixing coefficient [2]	✗	✓	✓	✗	✗	✓	✓	✓

d_h between the joint distribution and the product of the marginal distributions, defined as (cf. [1])

$$d(X, Y) \triangleq d_h(f_{XY}(x, y), f_X(x) \cdot f_Y(y)). \tag{3.1}$$

Maximal correlation is defined as (cf. [18])

$$S(X, Y) = \sup_{f,g} R(f(X), g(Y)), \tag{3.2}$$

where R is the Pearson correlation coefficient, and where f, g are Borel-measurable functions such that $R(f(X), g(Y))$ is defined [18].

Clearly, (3.1) and (3.2) are symmetric. The joint distribution and the product of the marginal distributions does not change by switching X and Y. Furthermore, the Pearson correlation coefficient is symmetric, making the maximal correlation also symmetric. Therefore, neither measure has property II.1.

There are two more measures (one of which is a variation of the other) which satisfy many (but not all) properties, and additionally closely resemble the measure we intend to propose. Namely, the *strong mixing coefficient* [2],

$$\alpha(X, Y) = \sup_{A \in \mathcal{E}_X, B \in \mathcal{E}_Y} \{|\mu_{X,Y}(A \times B) - \mu_X(A)\mu_Y(B)|\},$$

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

and its relaxation, the *β-mixing coefficient* [2],

$$\beta(X, Y) = \sup \left\{ \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |(\mu_{X,Y}(A_i \times B_j) - \mu_X(A_i)\mu_Y(B_j))| \right\},$$

where the supremum is taken over all finite partitions $(A_1, A_2, \ldots, A_I)$ and $(B_1, B_2, \ldots, B_J)$ of E_X and E_Y with $A_i \in \mathcal{E}_X$ and $B_j \in \mathcal{E}_Y$. However, these measures fail the properties II.1, II.4, and II.5.

4. The Berkelmans–Pries dependency function

After devising a new list of ideal properties (see Table 2) and showing that these properties are not fulfilled by existing dependency functions (see Table 3), we will now introduce a new dependency function that will meet all requirements. Throughout, we refer to this function as the *Berkelmans–Pries* (BP) dependency function.

The key question surely is: What is dependency? Although this question deserves an elaborate philosophical study, we believe that measuring the dependency of Y on X is essentially measuring how much the distribution of Y changes on average based on the knowledge of X, divided by the maximum possible change. This is the key insight on which the BP dependency function is based. To measure this, we first have to determine the difference between the distribution of Y *with* and *without* conditioning on the value of X times the probability that X takes on this value (Section 4.1). Secondly, we have to measure what the maximum possible change in probability mass is, which is used to properly scale the dependency function and make it asymmetric (see Section 4.2).

4.1. Definition of the expected absolute change in distribution

We start by measuring the *expected absolute change in distribution* (UD), which is the difference between the distribution of Y *with* and *without* conditioning on the value of X times the probability that X takes on this value. For discrete RVs, we obtain the following definition.

Definition 4.1. *(Discrete UD.)* For any discrete RVs X and Y,

$$\mathrm{UD}(X, Y) := \sum_x p_X(x) \cdot \sum_y |p_{Y|X=x}(y) - p_Y(y)|.$$

More explicit formulations of UD for specific combinations of RVs are given in Appendix B. For example, when X and Y remain discrete and take values in E_X and E_Y, respectively, it can equivalently be defined as

$$\mathrm{UD}(X, Y) := 2 \sup_{A \subset E_X \times E_Y} \left\{ \sum_{(x,y) \in A} (p_{X,Y}(x, y) - p_X(x) \cdot p_Y(y)) \right\}.$$

Similarly, for continuous RVs, we obtain the following definition for UD.

Definition 4.2. *(Continuous UD.)* For any continuous RVs X and Y,

$$\mathrm{UD}(X, Y) := \int_{\mathbb{R}} \int_{\mathbb{R}} |f_{X,Y}(x, y) - f_X(x) f_Y(y)| \, \mathrm{d}y \, \mathrm{d}x.$$

Note that this is the same as Δ_{L_1} [3].

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

In the general case, UD is defined in the following manner.

Definition 4.3. *(General UD.)* For any $X : (\Omega, \mathcal{F}, \mu) \to (E_X, \mathcal{E}(X))$ and $Y : (\Omega, \mathcal{F}, \mu) \to (E_Y, \mathcal{E}(Y))$, UD is defined as

$$\mathrm{UD}(X, Y) := 2 \sup_{A \in \mathcal{E}(X) \bigotimes \mathcal{E}(Y)} \{\mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A)\},$$

where $\mathcal{E}(X) \bigotimes \mathcal{E}(Y)$ is the σ-algebra generated by the sets $C \times D$ with $C \in \mathcal{E}(X)$ and $D \in \mathcal{E}(Y)$. Furthermore, $\mu_{(X,Y)}$ denotes the joint probability measure on $\mathcal{E}(X) \bigotimes \mathcal{E}(Y)$, and $\mu_X \times \mu_Y$ is the product measure.

4.2. Maximum UD given *Y*

Next, we have to determine the maximum of UD for a fixed Y in order to scale the dependency function to [0, 1]. To this end, we prove that, for a given Y, X fully determines $Y \Rightarrow \mathrm{UD}(X, Y) \geq \mathrm{UD}(X', Y)$ for any RV X'.

The full proof for the general case is given in Appendix C.4, which uses the upper bound determined in Appendix C.3. However, we show the discrete case here to give some intuition about the proof. Let $C_y = \{x \mid p_{X,Y}(x, y) \geq p_X(x) \cdot p_Y(y)\}$; then

$$\begin{aligned}
\mathrm{UD}(X, Y) &= 2 \sum_y (p_{X,Y}(C_y \times \{y\}) - p_X(C_y) \cdot p_Y(y)) \\
&\leq 2 \sum_y (\min\{p_X(C_y), p_Y(y)\} - p_X(C_y) \cdot p_Y(y)) \\
&= 2 \sum_y (\min\{p_X(C_y) \cdot (1 - p_Y(y)), (1 - p_X(C_y)) \cdot p_Y(y)\}) \\
&\leq 2 \sum_y (p_Y(y) \cdot (1 - p_Y(y))) \\
&= 2 \sum_y (p_Y(y) - p_Y(y)^2) = 2\left(1 - \sum_y p_Y(y)^2\right),
\end{aligned}$$

with equality if and only if both inequalities are equalities. Which occurs if and only if $p_{X,Y}(C_y \times \{y\}) = p_X(C_y) = p_Y(y)$ for all y. So we have equality when, for all y, the set C_y has the property that $x \in C_y$ if and only if $Y = y$. Or equivalently, $Y = f(X)$ for some function f. Thus,

$$UD(X, Y) \leq 2\left(1 - \sum_y p_Y(y)^2\right),$$

with equality if and only if $Y = f(X)$ for some function f.

Note that this holds for every X that fully determines Y. In particular, for $X := Y$ it now follows that $\mathrm{UD}(Y, Y) = 2 \cdot \left(1 - \sum_y p_Y(y)^2\right) \geq \mathrm{UD}(X', Y)$ for any RV X'.

4.3. Definition of the Berkelmans–Pries dependency function

Finally, we can define the BP dependency function to measure how much Y is dependent on X.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

Definition 4.4. *(BP dependency function.)* For any RVs X and Y, the *Berkelmans–Pries dependency function* is defined as

$$\mathrm{Dep}(Y \mid X) := \begin{cases} \frac{\mathrm{UD}(X,Y)}{\mathrm{UD}(Y,Y)} & \text{if } Y \text{ is not a.s. constant,} \\ \text{undefined} & \text{if } Y \text{ is trivial (has an atom of size 1).} \end{cases}$$

This is the difference between the distribution of Y *with* and *without* conditioning on the value of X times the probability that X takes on this value, divided by the largest possible difference for an arbitrary X'. Note that $\mathrm{UD}(Y, Y) = 0$ if and only if Y is almost surely constant (see Appendix C.4), which leads to division by zero. However, we previously argued in Section 2 that if Y is almost surely constant, it is completely independent as well as completely determined by X. It should therefore be undefined.

5. Properties of the Berkelmans–Pries dependency function

Next, we show that our new BP dependency function satisfies all the requirements from Table 2. To this end, we use properties of UD (see Appendix C) to derive properties II.1–8.

5.1. Property II.1 (Asymmetry)

In Example 2.1 we have $\mathrm{UD}(X, Y) = 1$, $\mathrm{UD}(X, X) = 2$, and $\mathrm{UD}(Y, Y) = 1$. Thus,

$$\mathrm{Dep}(Y \mid X) = \frac{\mathrm{UD}(X, Y)}{\mathrm{UD}(Y, Y)} = 1,$$
$$\mathrm{Dep}(X \mid Y) = \frac{\mathrm{UD}(X, Y)}{\mathrm{UD}(X, X)} = \frac{1}{2}.$$

Therefore, we see that $\mathrm{Dep}(Y \mid X) \neq \mathrm{Dep}(X \mid Y)$ for this example, thus making the BP dependency asymmetric.

5.2. Property II.2 (Range)

In Appendix C.2, we show that, for every X, Y, $\mathrm{UD}(X, Y) \geq 0$. Furthermore, in Appendix C.3 we prove that $\mathrm{UD}(X, Y) \leq 2\left(1 - \sum_{y \in d_Y} \mu_Y(\{y\})^2\right)$ for all RVs X. In Appendix C.4 we show for almost all cases that this bound is tight for $\mathrm{UD}(Y, Y)$. Thus, it must hold that $0 \leq \mathrm{UD}(X, Y) \leq \mathrm{UD}(Y, Y)$, and it then immediately follows that $0 \leq \mathrm{Dep}(Y \mid X) \leq 1$.

5.3. Property II.3 (Independence and dependency 0)

In Appendix C.2 we prove that $\mathrm{UD}(X, Y) = 0 \Leftrightarrow X$ and Y are independent. Furthermore, note that $\mathrm{Dep}(Y \mid X) = 0$ if and only if $\mathrm{UD}(X, Y) = 0$. Thus, $\mathrm{Dep}(Y \mid X) = 0 \Leftrightarrow X$ and Y are independent.

5.4. Property II.4 (Functional dependence and dependency 1)

In Section C.4, we show that if X fully determines Y, and X' is any RV, we have $\mathrm{UD}(X, Y) \geq \mathrm{UD}(X', Y)$. This holds in particular for $X := Y$. Thus, if X fully determines Y it follows that $\mathrm{UD}(X, Y) = \mathrm{UD}(Y, Y)$, so $\mathrm{Dep}(Y \mid X) = \mathrm{UD}(X, Y)/\mathrm{UD}(Y, Y) = 1$. In conclusion, if there exists a measurable function f such that $Y = f(X)$, then $\mathrm{Dep}(Y \mid X) = 1$.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

5.5. Property II.5 (Unambiguity)

We show the result for discrete RVs here; for the proof of the general case see Appendix C.5. Let E be the range of the independent $Y_1, Y_2, \ldots, Y_N$. By definition, $\mathbb{P}(X=x) = \sum_j \mathbb{P}(Y_j=x)\cdot\mathbb{P}(S=j)$, so, for all $i \in \{1, \ldots N\}$,

$$\begin{aligned}
\mathrm{UD}(X, Y_i) &= 2 \sup_{A\subset E\times E} \left\{ \sum_{(x,y)\in A} (\mathbb{P}(X=x, Y_i=y) - \mathbb{P}(X=x)\mathbb{P}(Y_i=y)) \right\} \\
&= 2 \sup_{A\subset E\times E} \left\{ \sum_{(x,y)\in A} \left(\sum_j \mathbb{P}(Y_j=x, Y_i=y, S=j) - \mathbb{P}(X=x)\mathbb{P}(Y_i=y) \right) \right\} \\
&= 2 \sup_{A\subset E\times E} \left\{ \sum_{(x,y)\in A} \left(\sum_{j\neq i} \mathbb{P}(Y_j=x)\mathbb{P}(Y_i=y)\mathbb{P}(S=j) \right.\right. \\
&\quad \left.\left. + \mathbb{P}(Y_i=x, Y_i=y)\mathbb{P}(S=i) - \sum_j \mathbb{P}(Y_j=x)\mathbb{P}(S=j)\mathbb{P}(Y_i=y) \right) \right\} \\
&= 2 \sup_{A\subset E\times E} \left\{ \sum_{(x,y)\in A} (p_i\mathbb{P}(Y_i=x, Y_i=y) - p_i\mathbb{P}(Y_i=x)\mathbb{P}(Y_i=y)) \right\} \\
&= p_i \cdot \mathrm{UD}(Y_i, Y_i).
\end{aligned}$$

This leads to

$$\mathrm{Dep}(Y_i \mid X) = \frac{\mathrm{UD}(X, Y_i)}{\mathrm{UD}(Y_i, Y_i)} = \frac{p_i \cdot \mathrm{UD}(Y_i, Y_i)}{\mathrm{UD}(Y_i, Y_i)} = p_i.$$

Therefore, we can conclude that property II.5 holds.

5.6. Property II.6 (Generally applicable)

The BP dependency measure can be applied for any combination of continuous, discrete, and categorical variables. It can handle arbitrarily many RVs as input by combining them. Thus, the BP dependency function is generally applicable.

5.7. Property II.7 (Invariance under isomorphisms)

In Appendix C.6 we prove that applying a measurable function to X or Y does not increase UD. Thus, it must hold for all isomorphisms f, g that

$$\mathrm{UD}(X, Y) = \mathrm{UD}(f^{-1}(f(X)), g^{-1}(g(Y))) \leq \mathrm{UD}(f(X), g(Y)) \leq \mathrm{UD}(X, Y).$$

Therefore, all inequalities are actually equalities. In other words, $\mathrm{UD}(f(X), g(Y)) = \mathrm{UD}(X, Y)$.

It now immediately follows for the BP dependency measure that

$$\mathrm{Dep}(g(Y) \mid f(X)) = \frac{\mathrm{UD}(f(X), g(Y))}{\mathrm{UD}(g(Y), g(Y))} = \frac{\mathrm{UD}(X, Y)}{\mathrm{UD}(Y, Y)} = \mathrm{Dep}(Y \mid X),$$

and thus property II.7 is satisfied.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

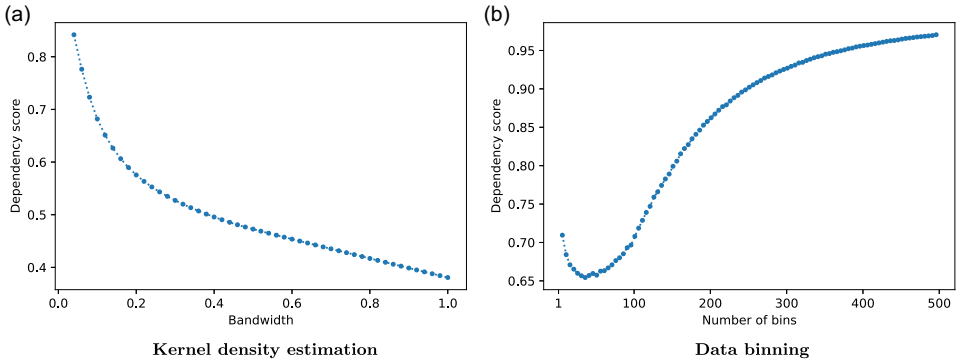

FIGURE 1. Influence of the chosen (a) bandwidth and (b) number of bins on the dependency score $\mathrm{Dep}(Y \mid X)$ with 5000 samples of $X \sim \mathcal{U}(0, 1)$ and $Y = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.1)$.

5.8. Property II.8 (Non-increasing under functions of X)

In Appendix C.6 we prove that transforming X or Y using a measurable function does not increase UD. In other words, for any measurable function f, $\mathrm{UD}(f(X), Y) \leq \mathrm{UD}(X, Y)$. Consequently, property II.8 holds for the BP dependency function, as

$$\mathrm{Dep}(Y \mid f(X)) = \frac{\mathrm{UD}(f(X), Y)}{\mathrm{UD}(Y, Y)} \leq \frac{\mathrm{UD}(X, Y)}{\mathrm{UD}(Y, Y)} = \mathrm{Dep}(Y \mid X).$$

6. Discussion and further research

Motivated by the need to measure and quantify the level dependence between random variables, we have proposed a general-purpose dependency function. The function meets an extensive list of important and desired properties, and can be viewed as a powerful alternative to the classical Pearson correlation coefficient, which is often used by data analysts today.

While it is recommended to use our new dependency function, it is important to understand the limitations and potential pitfalls of the new dependency function; we now discuss these aspects.

The underlying probability density function of an RV is often unknown in practice; instead, a set of outcomes is observed. These samples can then be used (in a simple manner) to approximate any discrete distribution. However, this is generally not the case for continuous variables. There are two main categories for dealing with continuous variables: either (i) the observed samples are combined using kernel functions into a continuous function (*kernel density estimation* [6]), or (2) the continuous variable is reduced to a discrete variable using *data binning*. The new dependency measure can be applied thereafter.

A main issue is that the dependency measure is dependent on the parameter choices of either kernel density estimation or data binning. To illustrate this, we conduct the following experiment. Let $X \sim \mathcal{U}(0, 1)$, and define $Y = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.1)$. Next, we draw 5000 samples of X and ϵ and determine each corresponding Y. For kernel density estimation we use Gaussian kernels with constant bandwidth. The result of varying the bandwidth on the dependency score can be seen in Figure 1(a). With data binning, both X and Y are binned using bins with fixed size. Increasing or decreasing the number of bins changes the size of the bins. The impact of changing the number of bins on the dependency score can be seen in Figure 1(b).

The main observation from Figures 1(a) and 1(b) is that the selection of the parameters is important. In the case of kernel density estimation, we see the traditional trade-off between

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

over-fitting when the bandwidth is too small and under-fitting when the bandwidth is too large. On the other hand, with data binning, we see different behavior: having too few bins seems to overestimate the dependency score, and as the number of bins increases the estimator of the dependency score decreases up to a certain point, after which it starts increasing again. The bottom of the curve seems to be marginally higher than the true dependency score of 0.621.

This observation raises a range of interesting questions for future research. For example, are the dependency scores estimated by binning consistently higher than the true dependency? Is there a correction that can be applied to get an unbiased estimator? Is the minimum of this curve an asymptotically consistent estimator? Which binning algorithms give the closest approximation to the true dependency?

An interesting observation with respect to kernel density estimation is that it appears that at a bandwidth of 0.1 the estimator of the dependency score is close to the true dependency score of approximately 0.621. However, this parameter choice could only be made if the underlying probability process was known a priori.

Yet, there is another challenge with kernel density estimation, when X consists of many variables or feature values. Each time Y is conditioned on a different value of X, either the density needs to be estimated again or the estimation of the joint distribution needs to be integrated. Both can rapidly become very time-consuming. When using data binning, it suffices to bin the data once. Furthermore, no integration is required, making it much faster. Therefore, our current recommendation would be to bin the data and not use kernel density estimation.

Another exciting research avenue would be to fundamentally explore the set of functions that satisfy all desired dependency properties. Is the BP dependency the only measure that fulfills all conditions? If two solutions exist, can we derive a new solution by smartly combining them? Without property II.5, any order-preserving bijection of [0, 1] with itself would preserve all properties when applied to a solution. However, property II.5 does restrict the solution space. It remains an open problem whether this is restrictive enough to result in a unique solution: the BP dependency.

Appendix A. Notation

The following general notation is used throughout the appendices. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_X, \mathcal{E}_X)$ and $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_Y, \mathcal{E}_Y)$ be RVs. Secondly, let $\mu_X(A) = \mathbb{P}(X^{-1}(A))$, $\mu_Y(A) = \mathbb{P}(Y^{-1}(A))$ be measures induced by X and Y on $(E_X, \mathcal{E}_X)$ and $(E_Y, \mathcal{E}_Y)$, respectively. Furthermore, $\mu_{X,Y}(A) = \mathbb{P}(\{\omega \in \Omega \mid (X(\omega), Y(\omega)) \in A\})$ is the joint measure and $\mu_X \times \mu_Y$ is the product measure on $(E_X \times E_Y, \mathcal{E}_X \bigotimes \mathcal{E}_Y)$ generated by $(\mu_X \times \mu_Y)(A \times B) = \mu_X(A)\mu_Y(B)$.

Appendix B. Formulations of UD

In this appendix we give multiple formulations of UD. Depending on the type of RVs, the following formulations can be used.

B.1. General case

For any X, Y, UD is defined as

$$\begin{aligned} \mathrm{UD}(X, Y) := & \sup_{A \in \mathcal{E}(X) \bigotimes \mathcal{E}(Y)} \{\mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A)\} \\ & + \sup_{B \in \mathcal{E}(X) \bigotimes \mathcal{E}(Y)} \{(\mu_X \times \mu_Y)(B) - \mu_{(X,Y)}(B)\} \\ & = 2 \sup_{A \in \mathcal{E}(X) \bigotimes \mathcal{E}(Y)} \{\mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A)\}. \end{aligned} \tag{B.1}$$

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

B.2. Discrete RVs only

When X, Y are discrete RVs, (B.1) simplifies into

$$\mathrm{UD}(X, Y) := \sum_{x,y} |p_{X,Y}(x, y) - p_X(x) \cdot p_Y(y)|,$$

or, equivalently,

$$\mathrm{UD}(X, Y) := \sum_{x} p_X(x) \cdot \sum_{y} |p_{Y|X=x}(y) - p_Y(y)|.$$

Similarly, when X and Y take values in E_X and E_Y, respectively, (B.1) becomes

$$\begin{aligned}\mathrm{UD}(X, Y) := &\sup_{A \subset E_X \times E_Y} \left\{ \sum_{(x,y)\in A} (p_{X,Y}(x, y) - p_X(x)p_Y(y)) \right\} \\ &+ \sup_{A \subset E_X \times E_Y} \left\{ \sum_{(x,y)\in A} (p_X(x)p_Y(y) - p_{X,Y}(x, y)) \right\} \\ &= 2 \sup_{A \subset E_X \times E_Y} \left\{ \sum_{(x,y)\in A} (p_{X,Y}(x, y) - p_X(x)p_Y(y)) \right\}.\end{aligned}$$

B.3. Continuous RVs only

When X, Y are continuous RVs, (B.1) becomes

$$\mathrm{UD}(X, Y) := \int_{\mathbb{R}} \int_{\mathbb{R}} |f_{X,Y}(x, y) - f_X(x)f_Y(y)| \, \mathrm{d}y \, \mathrm{d}x,$$

or, equivalently,

$$\mathrm{UD}(X, Y) := \int_{\mathbb{R}} f_X(x) \int_{\mathbb{R}} |f_{Y|X=x}(y) - f_Y(y)| \, \mathrm{d}y \, \mathrm{d}x.$$

Another formulation (more measure theoretical) would be:

$$\mathrm{UD}(X, Y) := 2 \cdot \sup_{A \in \mathcal{B}(\mathbb{R}^2)} \left\{ \int_A (f_{X,Y}(x, y) - f_X(x)f_Y(y)) \, \mathrm{d}y \, \mathrm{d}x \right\}.$$

B.4. Mix of discrete and continuous

When X is discrete and Y is continuous, (B.1) reduces to

$$\mathrm{UD}(X, Y) := \sum_{x} p_X(x) \int_{y} |f_{Y|X=x}(y) - f_Y(y)| \, \mathrm{d}y.$$

Vice versa, if X is continuous and Y is discrete, (B.1) becomes

$$\mathrm{UD}(X, Y) := \int_{x} f_X(x) \sum_{y} |p_{Y|X=x}(y) - p_Y(y)| \, \mathrm{x}.$$

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

Appendix C. UD Properties

In this appendix we prove properties of UD that are used in Section 5 to show that the BP dependency measure satisfies all the properties in Table 2.

C.1. Symmetry

For the proofs below it is useful to show that $\mathrm{UD}(X, Y)$ is symmetric, i.e. $\mathrm{UD}(X, Y) = \mathrm{UD}(Y, X)$ for every X, Y. This directly follows from the definition, as

$$\begin{aligned}\mathrm{UD}(X, Y) &= 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \{\mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A)\} \\ &= 2 \sup_{A \in \mathcal{E}_Y \bigotimes \mathcal{E}_X} \{\mu_{(Y,X)}(A) - (\mu_Y \times \mu_X)(A)\} \\ &= \mathrm{UD}(Y, X).\end{aligned}$$

C.2. Independence and UD = 0

Since we are considering a measure of dependence, it is useful to know what the conditions for independence are. Below we show that we have independence of X and Y if and only if $\mathrm{UD}(X, Y) = 0$.

Note that

$$\begin{aligned}\mathrm{UD}(X, Y) = &\sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \{\mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A)\} \\ &+ \sup_{B \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \{(\mu_X \times \mu_Y)(B) - \mu_{(X,Y)}(B)\} \\ \geq\ &(\mu_{(X,Y)}(E_X \times E_Y) - (\mu_X \times \mu_Y)(E_X \times E_Y)) \\ &+ ((\mu_X \times \mu_Y)(E_X \times E_Y) - \mu_{(X,Y)}(E_X \times E_Y)) \\ =\ &0,\end{aligned}$$

with equality if and only if $\mu_{(X,Y)} = \mu_X \times \mu_Y$ on $\mathcal{E}_X \bigotimes \mathcal{E}_Y$, so if and only if X and Y are independent. So, in conclusion, the following properties are equivalent:

- X and Y are independent random variables.
- $\mathrm{UD}(X, Y) = 0$.

C.3. Upper bound for a given Y

To scale the dependency function it is useful to know what the range of $\mathrm{UD}(X, Y)$ is for a given random variable Y. We already know it is bounded below by 0 (see Appendix C.2). However, we have not yet established an upper bound. What follows is a derivation of the upper bound.

A μ_Y-atom A is a set such that $\mu_Y(A) > 0$ and, for any $B \subset A$, $\mu_Y(B) \in \{0, \mu_Y(A)\}$. Consider the equivalence relation $\sim$ on μ_Y-atoms characterized by $S \sim T$ if and only if $\mu_Y(S \triangle T) = 0$. Then let I be a set containing exactly one representative from each equivalence class. Note that I is countable, so we can enumerate the elements $A_1, A_2, A_3, \ldots$ Additionally, for any $A, B \in I$ we have $\mu_Y(A \cap B) = 0$.

Next, we define $B_i := A_i \setminus \bigcup_{j=1}^{i-1} A_j$ to obtain a set of disjoint μ_Y-atoms. In what follows we assume I to be infinite, but the proof works exactly the same for finite I when you replace ∞ with $|I|$.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

Let $E_Y^* := E_Y \setminus \bigcup_{j=1}^{\infty} B_j$, so that the B_j and the E_Y^* form a partition of E_Y. Furthermore, let $b_j := \mu_Y(B_j)$ be the probabilities of being in the individual atoms in I (and therefore the sizes corresponding to the equivalence classes of atoms). We now have, for any RV X,

$$\begin{aligned}\mathrm{UD}(X,Y) &= 2 \sup_{A\in\mathcal{E}_X\bigotimes\mathcal{E}_Y} \{\mu_{X,Y}(A) - (\mu_X\times\mu_Y)(A)\} \\ &\le 2 \sup_{A\in\mathcal{E}_X\bigotimes\mathcal{E}_Y} \{\mu_{X,Y}(A\cap(E_X\times E_Y^*)) - (\mu_X\times\mu_Y)(A\cap(E_X\times E_Y^*))\} \\ &\quad + 2 \sup_{A\in\mathcal{E}_X\bigotimes\mathcal{E}_Y} \left\{ \sum_{j=1}^{\infty} (\mu_{X,Y}(A\cap(E_X\times B_j)) \right. \\ &\qquad \left. - (\mu_X\times\mu_Y)(A\cap(E_X\times B_j))) \right\}. \end{aligned} \tag{C.1}$$

Now note that the first term is at most $\mu_Y(E_Y^*) = 1 - \sum_{i=1}^{\infty} b_i$. To bound the second term, we examine each individual term of the summation. First, we note that the set of finite unions of 'rectangles' (Cartesian products of elements in $\mathcal{E}_X$ and $\mathcal{E}_Y$)

$$R := \left\{ C\in\mathcal{E}_X\bigotimes\mathcal{E}_{\mathcal{Y}} \mid \text{ there exists } k\in\mathbb{N} \text{ such that } C = \bigcup_{i=1}^{k} (A_i\times B_i), \text{ with, for all } i, A_i\in\mathcal{E}_X \wedge B_i\in\mathcal{E}_Y \right\}$$

is an algebra. Therefore, for any $D\in\mathcal{E}_X\bigotimes\mathcal{E}_Y$ and $\epsilon>0$, there exists a $D_\epsilon\in R$ such that $\nu(D_\epsilon\triangle D)<\epsilon$, where $\nu := \mu_{X,Y} + (\mu_X\times\mu_Y)$. Specifically, for $A\cap(E_X\times B_j)$ and $\epsilon>0$ there exists a $B_{j,\epsilon}\in R$ such that $\nu(B_{j,\epsilon}\triangle A\cap(E_X\times B_j))<\epsilon$ and $B_{j,\epsilon}\subset E_X\times B_j$ holds, since intersecting with this set only decreases the expression while remaining in R.

Thus, we have

$$|\mu_{X,Y}(A\cap(E_X\times B_j)) - \mu_{X,Y}(B_{j,\epsilon})| + |(\mu_X\times\mu_Y)(A\cap(E_X\times B_j)) - (\mu_X\times\mu_Y)(B_{j,\epsilon})| < \epsilon.$$

Therefore, it must hold that

$$\mu_{X,Y}(A\cap(E_X\times B_j)) - (\mu_X\times\mu_Y)(A\cap(E_X\times B_j)) \le \mu_{X,Y}(B_{j,\epsilon}) - (\mu_X\times\mu_Y)(B_{j,\epsilon}) + \epsilon.$$

Since $B_{j,\epsilon}$ is a finite union of 'rectangles', we can also write it as a finite union of k disjoint 'rectangles' such that $B_{j,\epsilon} = \bigcup_{i=1}^{k} S_i\times T_i$ with $S_i\in\mathcal{E}_X$ and $T_i\in\mathcal{E}_Y$ for all i. It now follows that

$$\mu_{X,Y}(B_{j,\epsilon}) - (\mu_X\times\mu_Y)(B_{j,\epsilon}) + \epsilon = \epsilon + \sum_{i=1}^{k} \mu_{X,Y}(S_i\times T_i) - (\mu_X\times\mu_Y)(S_i\times T_i).$$

For all i we have $T_i\subset B_j$ such that either $\mu_Y(T_i)=0$ or $\mu_Y(T_i)=b_j$, since B_j is an atom of size b_j. This allows us to separate the sum:

$$\begin{aligned}\epsilon + \sum_{i=1}^{k} \mu_{X,Y}(S_i\times T_i) - (\mu_X\times\mu_Y)(S_i\times T_i) &= \epsilon \\ &\quad + \sum_{i:\mu_Y(T_i)=0} (\mu_{X,Y}(S_i\times T_i) - (\mu_X(S_i)\times\mu_Y(T_i)) \\ &\quad + \sum_{i:\mu_Y(T_i)=b_j} (\mu_{X,Y}(S_i\times T_i) - (\mu_X(S_i)\times\mu_Y(T_i)) \\ &= \star.\end{aligned}$$

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

The first sum is equal to zero, since $\mu_{X,Y}(S_i \times T_i) \leq \mu_Y(T_i) = 0$. The second sum is bounded above by $\mu_{X,Y}(S_i \times T_i) \leq \mu_{X,Y}(S_i \times B_j)$. By defining $S' = \bigcup_{i:\mu_Y(T_i)=b_j} S_i$, we obtain

$$\begin{aligned}
\star &\leq \epsilon + 0 + \sum_{i:\mu_Y(T_i)=b_j} (\mu_{X,Y}(S_i \times B_j) - b_j \cdot \mu_X(S_i)) \\
&= \epsilon + \mu_{X,Y}(S' \times B_j) - b_j \cdot \mu_x(S') \\
&\leq \epsilon + \min\left\{(1-b_j)\cdot \mu_X(S'), b_j \cdot (1-\mu_X(S')\right\} \\
&\leq \epsilon + b_j - b_j^2.
\end{aligned}$$

But, since this is true for any $\epsilon > 0$, we have

$$\mu_{X,Y}(A \cap (E_X \times B_j)) - (\mu_X \times \mu_Y)(A \cap (E_X \times B_j)) \leq b_j - b_j^2.$$

Plugging this back into (C.1) gives

$$\begin{aligned}
\mathrm{UD}(X, Y) &\leq 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \{\mu_{X,Y}(A \cap (E_X \times E_Y^*)) - (\mu_X \times \mu_Y)(A \cap (E_X \times E_Y^*))\} \\
&\quad + 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} \left\{ \sum_{j=1}^{\infty} (\mu_{X,Y}(A \cap (E_X \times B_j)) - (\mu_X \times \mu_Y)(A \cap (E_X \times B_j))) \right\} \\
&\leq 2\left(1 - \sum_{i=1}^{\infty} b_i\right) + 2 \cdot \sum_{j=1}^{\infty} (b_j - b_j^2) \\
&= 2\left(1 - \sum_{i=1}^{\infty} b_i^2\right).
\end{aligned}$$

Note that in the continuous case the summation is equal to 0, so the upper bound simply becomes 2. In the discrete case, where E_Y is the set in which Y takes its values, the expression becomes $\mathrm{UD}(X, Y) \leq 2\big(1 - \sum_{i \in E_Y} \mathbb{P}(Y = i)^2\big)$.

C.4. Functional dependence attains maximum UD

Since we established an upper bound in Appendix C.3, the next step is to check whether this bound is actually attainable. What follows is a proof that this bound is achieved for any random variable X for which $Y = f(X)$ for some measurable function f.

Let $Y = f(X)$ for some measurable function f; then $\mu_X(f^{-1}(C)) = \mu_Y(C)$ for all $C \in \mathcal{E}_Y$. Let the μ_Y-atoms B_j and E_Y^* be the same as in Appendix C.3. Since E_Y^* contains no atoms, for every $\epsilon > 0$ there exists a partition $T_1, \ldots, T_k$ for some $k \in \mathbb{N}$ such that $\mu_Y(T_i) < \epsilon$ for all i. Then, consider the set $K = \big(\bigcup_i (f^{-1}(T_i) \times T_i)\big) \cup \big(\bigcup_j (f^{-1}(B_j) \times B_j)\big)$. It now follows that

$$\begin{aligned}
\mathrm{UD}(X, Y) &= 2 \sup_{A \in \mathcal{E}_Y \bigotimes \mathcal{E}_Y} \mu_{X,Y}(A) - (\mu_X \times \mu_Y)(A) \\
&\geq 2\mu_{X,Y}(K) - (\mu_X \times \mu_Y)(K)
\end{aligned}$$

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

$$
\begin{aligned}
&= 2\Bigg(\sum_i (\mu_{X,Y}(f^{-1}(T_i) \times T_i) - \mu_X(f^{-1}(T_i))\mu_Y(T_i)) \\
&\quad + \sum_j (\mu_{X,Y}(f^{-1}(B_j) \times B_j) - \mu_X(f^{-1}(B_j))\mu_Y(B_j)) \Bigg) \\
&\geq 2\Bigg(\sum_i (\mu_Y(T_i) - \epsilon * \mu_Y(T_i)) + \sum_j (b_j - b_j^2) \Bigg) \\
&= 2\Bigg(\Bigg(1 - \sum_j b_j\Bigg) - \epsilon \Bigg(1 - \sum_j b_j\Bigg) + \sum_j (b_j - b_j^2) \Bigg).
\end{aligned}
$$

But, since this holds for any $\epsilon > 0$, we have $\mathrm{UD}(X, Y) \geq 2\big(1 - \sum_j b_j^2\big)$. As this is also the upper bound from Appendix C.3, equality must hold. Thus, we can conclude that $\mathrm{UD}(X, Y)$ is maximal for Y if $Y = f(X)$ (so, in particular, if $X = Y$). As a result, for any RVs X_1, X_2, Y with $Y = f(X_1)$ for some measurable function f, we have $\mathrm{UD}(X_1, Y) \geq \mathrm{UD}(X_2, Y)$. Note that a corollary of this proof is that $\mathrm{UD}(Y, Y) = 0$ if and only if there exists a μ_Y-atom B_i with $\mu_Y(B_i) = 1$.

C.5. Unambiguity

In Section 5, we show for discrete RVs that property II.5 holds. In this section, we prove the general case. Let $Y_1, \ldots, Y_N$ and S be independent RVs where S takes values in $1, \ldots, N$ with $\mathbb{P}(S = i) = p_i$. Finally, define $X := Y_S$. Then we will show that $\mathrm{Dep}(Y_i \mid X) = p_i$.

Let $\mathcal{E}$ be the σ-algebra on which the independent Y_i are defined. Then we have $\mu_{X,Y_i,S}(A \times \{j\}) = \mu_{Y_j,Y_i}(A)\mu_S(\{j\}) = p_j\mu_{Y_j,Y_i}(A)$ for all j. Additionally, we have $\mu_X(A) = \sum_j p_j\mu_{Y_j}(A)$. Lastly, due to independence for $i \neq j$, we have $\mu_{Y_j,Y_i} = \mu_{Y_j} \times \mu_{Y_i}$. Combining all this gives

$$
\begin{aligned}
\mathrm{UD}(X, Y_i) &= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \big\{\mu_{X,Y_i}(A) - (\mu_X \times \mu_{Y_i})(A)\big\} \\
&= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \left\{ \sum_j \mu_{X,Y_i,S}(A \times \{j\}) - \sum_j p_j(\mu_{Y_j} \times \mu_{Y_i})(A) \right\} \\
&= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \left\{ \sum_j p_j(\mu_{Y_j,Y_i}(A) - (\mu_{Y_j} \times \mu_{Y_i})(A)) \right\} \\
&= 2 \sup_{A \in \mathcal{E} \times \mathcal{E}} \{p_i(\mu_{Y_i,Y_i}(A) - (\mu_{Y_i} \times \mu_{Y_i})(A))\} \\
&= p_i \cdot \mathrm{UD}(Y_i, Y_i).
\end{aligned}
$$

C.6. Measurable functions never increase UD

Next, we prove another useful property of UD: applying a measurable function to one of the variables does not increase the UD. Let $f : (E_X, \mathcal{E}_X) \to (E_{X'}, \mathcal{E}_{X'})$ be a measurable function. Then $h : E_X \times E_Y \to E_{X'} \times E_Y$ with $h(x, y) = (f(x), y)$ is measurable. Now it follows that

$$
\begin{aligned}
\mathrm{UD}(f(X), Y) &= 2 \sup_{A \in \mathcal{E}_{X'} \otimes \mathcal{E}_Y} \{\mu_{(f(X),Y)}(A) - (\mu_{f(X)} \times \mu_Y)(A)\} \\
&= 2 \sup_{A \in \mathcal{E}_{X'} \otimes \mathcal{E}_Y} \{\mu_{(X,Y)}(h^{-1}(A)) - (\mu_X \times \mu_Y)(h^{-1}(A))\},
\end{aligned}
$$

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

with $h^{-1}(A) \in \mathcal{E}_X \bigotimes \mathcal{E}_Y$. Thus,

$$\begin{aligned}\mathrm{UD}(f(X), Y) &\leq 2 \sup_{A \in \mathcal{E}_X \bigotimes \mathcal{E}_Y} (\mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A)) \\ &= \mathrm{UD}(X, Y).\end{aligned}$$

Appendix C.1 proved that UD is symmetric. Therefore, for $g : E_Y \to E_{Y'}$, $\mathrm{UD}(X, g(Y)) \leq \mathrm{UD}(X, Y)$.

Acknowledgement

The authors wish to thank the anonymous referees for their useful comments, which led to a significant improvement of the readability and quality of the paper.

Funding information

There are no funding bodies to thank relating to the creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Agarwal, R., Sacre, P. and Sarma, S. V. (2015). Mutual dependence: A novel method for computing dependencies between random variables. Preprint, arXiv:1506.00673.
[2] Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. Prob. *Surv.* **2**, 107–144.
[3] Capitani, L., Bagnato, L. and Punzo, A. (2014). Testing serial independence via density-based measures of divergence. *Methodology Comput. Appl. Prob.* **16**, 627–641.
[4] Embrechts, P., McNeil, A. J. and Straumann, D. (2002). *Correlation and Dependence in Risk Management: Properties and Pitfalls*. Cambridge University Press, pp. 176–223.
[5] Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *J. Appl. Math. Mech.* **21**, 364–379.
[6] Gramacki, A. (2017). *Nonparametric Kernel Density Estimation and Its Computational Aspects*, 1st edn. Springer, New York.
[7] Granger, C. W., Maasoumi, E. and Racine, J. (2004). A dependence metric for possibly nonlinear processes. *J. Time Series Anal.* **25**, 649–669.
[8] Gretton, A., Herbrich, R., Smola, A., Bousquet, O. and Schölkopf, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6**, 2075–2129.
[9] Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *J. reine angew. Math.* **1909**, 210–271.
[10] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–377.
[11] Janse, R. J., Hoekstra, T., Jager, K. J., Zoccali, C., Tripepi, G., Dekker, F. W. and van Diepen, M. (2021). Conducting correlation analysis: Important limitations and pitfalls. *Clinical Kidney J.* **14**, 2332–2337.
[12] Joe, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.* **84**, 157–164.
[13] Kimeldorf, G. and Sampson, A. R. (1978). Monotone dependence. *Ann. Statist.* **6**, 895–903.
[14] Kruskal, W. H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.* **53**, 814–861.
[15] Lancaster, H. O. (1963). Correlation and complete dependence of random variables. *Ann. Math. Statist.* **34**, 1315–1321.
[16] Móri, T. F. and Székely, G. J. (2019). Four simple axioms of dependence measures. *Metrika* **82**, 1–16.
[17] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007). *Numerical Recipes*, 3 edn. Cambridge University Press.
[18] Rényi, A. (1959). On measures of dependence. *Acta Math. Acad. Scient. Hungar.* **10**, 441–451.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press

[19] RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., MCVEAN, G., TURNBAUGH, P. J., LANDER, E. S., MITZENMACHER, M. AND SABETI, P. C. (2011). Detecting novel associations in large data sets. *Science* **334**, 1518–1524.
[20] SUGIYAMA, M. AND BORGWARDT, K. M. (2013). Measuring statistical dependence via the mutual information dimension. In *Proc. Twenty-Third Int. Joint Conf. Artificial Intelligence*, IJCAI '13. AAAI Press, Beijing, pp. 1692–1698.
[21] SZÉKELY, G. J. AND RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Statist.* **3**, 1236–1265.
[22] WATANABE, S. (1960). Information theoretical analysis of multivariate correlation. *IBM J. Res. Devel.* **4**, 66–82.

https://doi.org/10.1017/jpr.2022.118 Published online by Cambridge University Press