

**THE BP DEPENDENCY FUNCTION: A GENERIC MEASURE
OF DEPENDENCE BETWEEN RANDOM VARIABLES**

GUUS BERKELMANS ^{*}, *Centrum Wiskunde & Informatica (CWI)*

SANDJAI BHULAI,^{**} *Vrije Universiteit (VU)*

ROB VAN DER MEI,^{***} *Centrum Wiskunde & Informatica (CWI)*

JORIS PRIES ^{****}, *Centrum Wiskunde & Informatica (CWI)*

arXiv:2203.12329v1 [stat.ML] 23 Mar 2022

* Postal address: Department of Stochastics, P.O. Box 94079, 1090 GB Amsterdam, Netherlands

** Postal address: Department of Mathematics, De Boelelaan 1111, 1081 HV Amsterdam, Netherlands

*** Postal address: Department of Stochastics, Science Park 123, 1098 XG Amsterdam, Netherlands

**** Postal address: Department of Stochastics, P.O. Box 94079, 1090 GB Amsterdam, Netherlands

* Email address: gberkelmans@cw.nl

** Email address: s.bhulai@vu.nl

*** Email address: mei@cw.nl

**** Email address: joris.pries@cw.nl

: Equal contribution and corresponding authors

Abstract

Measuring and quantifying dependencies between random variables (RV's) can give critical insights into a data-set. 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 RV's, common practice of data analysis is that most data analysts use the *Pearson correlation coefficient* (PCC) to quantify dependence between RV's, while it is well-recognized that the PCC is essentially a measure for *linear* dependency only. Although many attempts have been made to define more generic dependency measures, there is yet no consensus 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, in this paper we will 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 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

In as early as 1958, Kruskal [12] 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 different levels of dependency. For example, 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 $[0,1]$, one would say that if Y is fully determined by X then the dependency of Y w.r.t. 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 PCC does not meet these requirements. In fact, many examples exist where Y is fully determined by X while $\text{PCC}=0$.

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 byproduct 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 data-set. In this way, redundant variables or variables that provide little additional information, can be removed to reduce the dimensionality of the data-set. 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 [16] proposed a list of ideal properties for a dependency function. A long list of follow-up papers (see the references in Table 1 below) 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.

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), which 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 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.

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 to be believed not relevant, or even wrong.

In Table 1 below, a summary is given of (twenty-two) '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 [13] indirectly touched upon this same point by 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 [13] 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 $X > \frac{1}{2}$.

Then, 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 out of $\{1, 2, 3, 4\}$ and $Y := X \bmod 2$.

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, $\text{Dep}(X, Y) \neq \text{Dep}(Y, X)$.

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 $\text{Dep}(X, Y)$, we will denote $\text{Dep}(Y|X)$ for how much Y is dependent on X . Based by 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's correlation coefficient* [15], 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.

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's 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.

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 [16] 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 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 $\text{Dep}(Y|X) < 1$, as there is only an infinitesimal difference compared to the strictly dependent case.

Desired property II.5 (Unambiguity):

Kruskal [12] 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 stroke with our natural expectation. Therefore, we introduce a new requirement based on a simple example: suppose we have a number of independent RV's 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, \dots, Y_N, S be

independent variables with S a selection variable s.t. $\mathbb{P}(S = i) = p_i$ 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 $\text{Dep}(Y_i|X) = p_i$ for all $i \in \{1, \dots, N\}$.

Desired property II.6 (Generally applicable):

Our aim is to find a general dependency function, which we denote by $\text{Dep}(X|Y)$. 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 a bit similar to arguing 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 the following example.

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

It should hold that $\text{Dep}(Y|X) = \text{Dep}(Y'|X')$ and $\text{Dep}(X|Y) = \text{Dep}(X'|Y')$, as the relationship between the variables is the same (only altered using isomorphisms).

Desired property II.8 (Non-increasing under functions of X):

Additionally, $\text{Dep}(Y|X)$ 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.

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 2 functions: $f_1(Y) := Y \bmod 2$ and $f_2(Y) := \lceil \frac{Y}{2} \rceil$. 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 .

Exclusion of Pearson’s correlation coefficient as a special case:

According to properties I.21-22, when X and Y are normally distributed the dependency function should coincide with or be a function of the *Pearson’s correlation coefficient*. However, these properties lack a good argumentation for why this would be ideal. It is not obvious why this would be a necessary condition. Even more, there are many known problems and pitfalls with the correlation coefficient [3, 9], 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 leave these properties out.

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 this section 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 any

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

Property group	Property	Article(s)
Range	I.1. $0 \leq \text{Dep}(X, Y) \leq 1$	[16, 1, 18, 6, 10, 7, 19, 17, 3]
	I.2. $\text{Dep}(X, Y) = 0 \Leftrightarrow X$ and Y are independent	[10, 6, 17]
	I.3. $\text{Dep}(X, Y) = 0 \Rightarrow X$ and Y are independent	[19]
	I.4. $\text{Dep}(X, Y) = 0 \Leftrightarrow X$ and Y are independent	[16, 1, 18, 7, 14, 3]
	I.5. $\text{Dep}(X, Y) = 1 \Leftrightarrow Y = LX$ with probability 1, where L is a similarity transformation	[14]
	I.6. $\text{Dep}(X, Y) = 1 \Leftrightarrow X$ and Y are strictly dependent	[16, 1, 6, 17]
	I.7. $\text{Dep}(X, Y) = 1 \Leftrightarrow X$ and Y are comonotonic or countermonotonic	[3]
	I.8. $\text{Dep}(X, Y) = 1 \Leftrightarrow X$ and Y are strictly dependent	[7]
General	I.9. $\text{Dep}(X, Y)$ is defined for any X, Y where both are not constant	[16, 7, 14]
	I.10. Well-defined for both continuous and discrete variables	[6]
	I.11. Defined for both categorical and continuous variables; and for ordinal categorical variables for which there may be underlying continuous variables	[10]
	I.12. There is a close relationship between the measure for the continuous variables and the measure for the discretization of the variables	[10]
Symmetric	I.13. $\text{Dep}(X, Y) = \text{Dep}(Y, X)$	[16, 1, 18, 17, 3]
	I.14. $\text{Dep}(f(X), g(Y)) = \text{Dep}(X, Y)$ with f, g strictly monotonic functions	[1]
Applying function to argument	I.15. $\text{Dep}(f(X), Y) = \text{Dep}(X, Y)$ with $f : \mathbb{R} \rightarrow \mathbb{R}$ strictly monotonic on the range of X	[3]
	I.16. $\text{Dep}(f(X), f(Y)) = \text{Dep}(X, Y)$ with f continuous and strictly increasing	[18, 6]
	I.17. $\text{Dep}(f(X), g(Y)) = \text{Dep}(X, Y)$ if $f(\cdot), g(\cdot)$ map the real axis in a one-to-one way onto itself	[16, 10]
	I.18. $\text{Dep}(X, Y)$ is invariant with respect to all similarity transformations	[14]
	I.19. $\text{Dep}(X, Y)$ is invariant with respect to translation and scaling	[18]
	I.20. $\text{Dep}(X, Y)$ is scale invariant	[19]
Behavior normal distribution	I.21. $\text{Dep}(X, Y)$ is a function of the Pearson's correlation if the joint distribution of X and Y is normal	[1, 6, 19]
	I.22. $\text{Dep}(X, Y) = \rho(X, Y) $ if the joint distribution of X and Y is normal, where ρ is the Pearson's correlation	[16, 10]

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

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

one of the desirable properties. Furthermore, almost all measures are not asymmetric. The one measure that comes closest to fulfilling all requirements, is the *uncertainty coefficient* [15]. This is a normalized asymmetric variant of the *mutual information* [15], where the discrete variant is defined as

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

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

$$\begin{aligned} H(Y) &= - \int_0^1 f_Y(y) \ln(f_Y(y)) dy \\ &= - \int_0^1 1 \cdot \ln(1) dy \\ &= 0. \end{aligned}$$

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.5 requires.

TABLE 3: Properties of previous dependencies functions (\times = property not satisfied, \checkmark = property satisfied)

Measure	Asymmetric		Intuitive			General		Functions	
	II.1	II.2	II.3	II.4	II.5	II.6	II.7	II.8	
Pearson correlation coefficient [15]	\times	\times	\times	\times	\times	\times	\times	\times	
Spearman's rank correlation coefficient [15]	\times	\times	\times	\times	\times	\times	\times	\times	
Kendall rank correlation coefficient [15]	\times	\times	\times	\times	\times	\times	\times	\times	
Mutual information [15]	\times	\times	\checkmark	\times	\times	\checkmark	\checkmark	\checkmark	
Uncertainty coefficient [15]	\checkmark	\checkmark	\checkmark	\checkmark	\times	\times	\checkmark	\checkmark	
Total correlation [20]	\times	\times	\checkmark	\times	\times	\checkmark	\checkmark	\checkmark	
Mutual dependence [1]	\times	\checkmark	\checkmark	\checkmark	\times	\checkmark	\checkmark	\checkmark	
Δ_{L_1} [2]	\times	\times	\checkmark	\times	\times	\checkmark	\checkmark	\checkmark	
Δ_{SD} [2]	\times	\times	\checkmark	\times	\times	\checkmark	\times	\times	
Δ_{ST} [2]	\times	\times	\times	\times	\times	\checkmark	\times	\times	
Monotone correlation [11]	\times	\checkmark	\checkmark	\times	\times	\times	\times	\times	
Maximal correlation [4]	\times	\checkmark	\checkmark	\checkmark	\times	\checkmark	\checkmark	\checkmark	
Distance correlation [19]	\times	\checkmark	\checkmark	\times	\times	\times	\times	\times	

Two other measures that satisfy many (but not all) properties are *mutual dependence* [1] and *maximal correlation* [4]. Mutual dependence is defined as the Hellinger distance [8] d_h between the joint distribution and the product of the marginal distributions, defined as follows (cf. [1]):

$$d(X, Y) \triangleq d_h(f_{XY}(x, y), f_X(x) \cdot f_Y(y)). \quad (1)$$

Maximal correlation is defined as (cf. [16]):

$$S(X, Y) = \sup_{f, g} R(f(X), g(Y)), \quad (2)$$

where R is the Pearson correlation coefficient, and where f, g are Borel-measurable functions, such that $R(f(X), g(Y))$ has a sense [16].

Clearly, Equations (1) and (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, both measures do not have property II.1

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 will 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, where the *BP* dependency function is based on. 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 in 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 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. To this end, for two discrete RV's X and Y , UD is defined as follows:

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

More explicit formulations of UD for specific combinations of RV's are given in Appendix A.1. For example, when X and Y remain discrete and take values in E_X and E_Y respectively, it can equivalently be defined as:

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

Similarly, if X and Y are two continuous RV's, UD can be written as

$$\text{UD}(X, Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} |f_{X,Y}(x, y) - f_X(x)f_Y(y)| dy dx,$$

which is the same as Δ_{L_1} [2].

In general, for $X : (\Omega, \mathcal{F}, \mu) \rightarrow (E_X, \mathcal{E}(X))$ and $Y : (\Omega, \mathcal{F}, \mu) \rightarrow (E_Y, \mathcal{E}(Y))$ UD is defined as

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

where $\mathcal{E}(X) \otimes \mathcal{E}(Y)$ 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 $\mathcal{E}(X) \otimes \mathcal{E}(Y)$ 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 :

$$X \text{ fully determines } Y \Rightarrow \text{UD}(X, Y) = \max_{X'} \{ \text{UD}(X', Y) \}.$$

The full proof for the general case is given in Appendix A.2.4, which uses the upper bound determined in Appendix A.2.3. However, we will show the discrete case here to give some intuition about the proof. Let $C_y = \{x | p_{X,Y}(x, y) \geq p_X(x) \cdot p_Y(y)\}$, then

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

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

Note that this holds for every X that fully determines Y . In particular, for $X := Y$ it now follows that

$$UD(Y, Y) = \max_{X'} \{UD(X', Y)\} = 2 \cdot \left(1 - \sum_y p_Y(y)^2 \right).$$

4.3. Definition Berkelmans-Pries Dependency Function

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

$$\text{Dep}(Y|X) = \begin{cases} \frac{UD(X,Y)}{UD(Y,Y)} & \text{if } Y \text{ is not a.s. constant,} \\ \text{undefined} & \text{if } Y \text{ is a.s. constant.} \end{cases}$$

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 $UD(Y, Y) = 0$ if and only if Y is almost surely constant (see Appendix A.2.4), 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 requirements from Table 2. To this end, we use properties of UD (see Appendix A.2) to derive properties II.1-8.

Property II.1 (Asymmetry): It holds for Example 2.1 that $\text{UD}(X, Y) = 1$, $\text{UD}(X, X) = 2$, and $\text{UD}(Y, Y) = 1$. Thus,

$$\begin{aligned}\text{Dep}(Y|X) &= \frac{\text{UD}(X, Y)}{\text{UD}(Y, Y)} = 1, \\ \text{Dep}(X|Y) &= \frac{\text{UD}(X, Y)}{\text{UD}(X, X)} = \frac{1}{2}.\end{aligned}$$

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

Property II.2 (Range): In Appendix A.2.2, we show that for every X, Y it holds that $\text{UD}(X, Y) \geq 0$. Furthermore, in Appendix A.2.3 we prove that $\sup_{X'} \{\text{UD}(X, Y)\} \leq 2 \left(1 - \sum_{y \in d_Y} \mu_Y(\{y\})^2\right)$. In Appendix A.2.4 we show for almost all cases that this bound is tight for $\text{UD}(Y, Y)$. Thus, it must hold that $0 \leq \text{UD}(X, Y) \leq \text{UD}(Y, Y)$ and it then immediately follows that $0 \leq \text{Dep}(Y|X) \leq 1$. The result for the highly specific remaining cases remains an open problem, the formulation (and conjecture) of which can be found in the Appendix A.2.4.

Property II.3 (Independence and dependency 0): In Appendix A.2.2, we prove that

$$\text{UD}(X, Y) = 0 \Leftrightarrow X \text{ and } Y \text{ are independent.}$$

Furthermore, note that $\text{Dep}(Y|X) = 0$ if and only if $\text{UD}(X, Y) = 0$. Thus,

$$\text{Dep}(Y|X) = 0 \Leftrightarrow X \text{ and } Y \text{ are independent.}$$

Property II.4 (Functional dependence and dependency 1): In Section A.2.4, we show that if X fully determines Y it holds that $\text{UD}(X, Y) = \max_{X'} \{\text{UD}(X', Y)\}$, in almost all cases. This holds in particular for $X := Y$. Thus, if X fully determines Y it follows that

$$\text{Dep}(Y|X) = \frac{\text{UD}(X, Y)}{\text{UD}(Y, Y)} = \frac{\max_{X'} \text{UD}(X', Y)}{\max_{X'} \text{UD}(X', Y)} = 1.$$

The result for the highly specific remaining cases remains an open problem, the formulation (and conjecture) of which can be found in the Section A.2.4.

Property II.5 (Unambiguity): Let \mathcal{E} denote the σ -algebra, where Y_1, Y_2, \dots, Y_N are defined. By definition, it holds that $\mathbb{P}(X = x) = \sum_j \mathbb{P}(Y_j = x) \cdot \mathbb{P}(S = j)$, so for

all $i \in \{1, \dots, N\}$

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

This leads to

$$\text{Dep}(Y_i|X) = \frac{\text{UD}(X, Y_i)}{\text{UD}(Y_i, Y_i)} = \frac{p_i \text{UD}(Y_i, Y_i)}{\text{UD}(Y_i, Y_i)} = p_i.$$

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

Property II.6 (Generally applicable): The *BP* dependency measure can be applied for any combination of continuous, discrete and categorical variables. It can handle arbitrary many RV's as input by combining them. Thus, the *BP* dependency function is generally applicable.

Property II.7 (Invariance under isomorphisms): In Appendix A.2.6, we show that UD is invariant under isomorphisms. In other words, for any isomorphisms f, g it holds that

$$\text{UD}(X, Y) = \text{UD}(f(X), g(Y)).$$

It follows for the *BP* dependency measure that

$$\begin{aligned} \text{Dep}(g(Y)|f(X)) &= \frac{\text{UD}(f(X), g(Y))}{\text{UD}(g(Y), g(Y))} = \frac{\text{UD}(X, Y)}{\text{UD}(Y, Y)} \\ &= \text{Dep}(Y|X), \end{aligned}$$

thus Property II.7 is satisfied.

Desired property II.8 (Non-increasing under functions of X): In Appendix A.2.5, we prove that transforming X or Y using a measurable function does not increase UD. In other words, for any measurable function f , it holds that

$$\text{UD}(f(X), Y) \leq \text{UD}(X, Y).$$

Consequently, Property II.8 holds for the *BP* dependency function, as

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

6. Discussion and Further Research

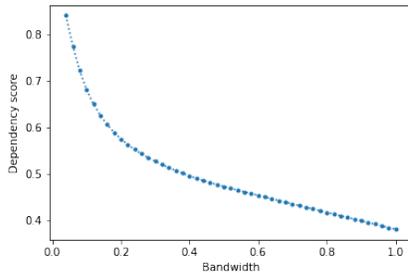
Motivated by the need to measure and quantify the level dependence between random variables, we have proposed a general-purpose dependency function. The function meets an extensive list of important and desired properties, and can be viewed as a powerful alternative to the classical PCC, which is often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

Whilst it is recommended to use our new dependency function, it is important to understand the limitations and potential pitfalls of the new dependency function. Below we elaborate on these aspects.

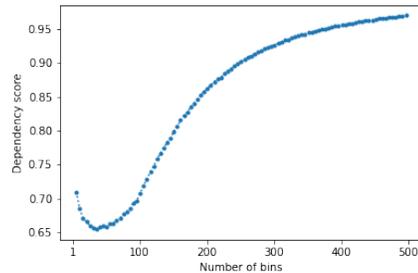
The underlying probability density function of a RV is often unknown in practice; instead, a set of outcomes is observed. These samples can then be used (in a simple manner) to approximate any discrete distribution. However, this is generally not the case for continuous variables. There are mainly two categories for dealing with continuous variables: either (1) the observed samples are combined using kernel functions into a continuous function (*kernel density estimation* [5]), 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 of 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 5,000 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 1a. 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 1b.



(a) **Kernel density estimation**



(b) **Data binning**

FIGURE 1: Influence of chosen bandwidth (a) / number of bins (b) on the dependency score $\text{Dep}(Y|X)$ with 5,000 samples of $X \sim \mathcal{U}(0, 1)$ and $Y = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.1)$.

The main observation from Figures 1a and 1b is that the selection of the parameters is important. In the case of the *kernel density estimation*, we see the traditional *trade-off* between *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 behaviour: Having too few bins seems to overestimate the dependency score and as bins increase the estimator of the dependency score decreases up to a certain point, where-after 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 of 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.

Appendix A.

Notation

The following general notation is used throughout this appendix. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (E_X, \mathcal{E}_X)$ and $Y : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (E_Y, \mathcal{E}_Y)$ be RV's. 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, $\mu_{X,Y}(A) = \mathbb{P}(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in A\})$ is the joint measure and $\mu_X \times \mu_Y$ is the product measure on $(E_X \times E_Y, \mathcal{E}_X \otimes \mathcal{E}_Y)$ generated by $(\mu_X \times \mu_Y)(A \times B) = \mu_X(A)\mu_Y(B)$.

A.1. Formulations of UD

In this appendix, we give multiple formulations of the *expected absolute change in distribution* (UD). Depending on the type of RV's, these formulations can be used.

A.1.1. **General case** For any X, Y the UD is defined as

$$\begin{aligned} \text{UD}(X, Y) &= \sup_{A \in \mathcal{E}(X) \otimes \mathcal{E}(Y)} \{ \mu_{(X, Y)}(A) - (\mu_X \times \mu_Y)(A) \} \\ &+ \sup_{B \in \mathcal{E}(X) \otimes \mathcal{E}(Y)} \{ (\mu_X \times \mu_Y)(B) - \mu_{(X, Y)}(B) \} \\ &= 2 \sup_{A \in \mathcal{E}(X) \otimes \mathcal{E}(Y)} \{ \mu_{(X, Y)}(A) - (\mu_X \times \mu_Y)(A) \}, \end{aligned} \quad (3)$$

A.1.2. **Discrete RV's only** When X, Y are discrete RV's, Equation 3 simplifies into

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

or equivalently,

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

Similarly, when X and Y take values in E_X and E_Y respectively, Equation 3 becomes

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

A.1.3. **Continuous RV's only** When X, Y are continuous RV's, Equation 3 becomes:

$$\text{UD}(X, Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} |f_{X, Y}(x, y) - f_X(x)f_Y(y)| dy dx,$$

or equivalently

$$\text{UD}(X, Y) = \int_{\mathbb{R}} f_X(x) \int_{\mathbb{R}} |f_{Y|X=x}(y) - f_Y(y)| dy dx.$$

Another formulation (more measure theoretical) would be:

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

A.1.4. Mix of discrete and continuous When X is discrete and Y is continuous, Equation 3 reduces to:

$$\text{UD}(X, Y) = \sum_x p_X(x) \int_y |f_{Y|X=x}(y) - f_Y(y)| dy.$$

Vice versa, if X is continuous and Y is discrete, Equation 3 becomes:

$$\text{UD}(X, Y) = \int_x f_X(x) \sum_y |p_{Y|X=x}(y) - p_Y(y)| dx.$$

A.2. Properties expected absolute change in distribution (UD)

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

A.2.1. Symmetry UD UD is symmetric i.e. $\text{UD}(X, Y) = \text{UD}(Y, X)$ for every X, Y as

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

A.2.2. Independence and UD = 0 Note that

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

with equality if and only if $\mu_{(X,Y)} = \mu_X \times \mu_Y$ on $\mathcal{E}_X \otimes \mathcal{E}_Y$, so if and only if X and Y are independent.

A.2.3. Upper bound given Y To scale the dependency function, we need to determine $\sup_{X'} \{\text{UD}(X', Y)\}$ for a given Y . Let $d_Y = \{y \in E_Y \mid \mu_Y(\{y\}) > 0\}$ be the set of all singletons with positive probability and let $c_Y = E_Y \setminus d_Y$. Furthermore, let $D_Y = E_X \times d_Y$ and $C_Y = E_X \times c_Y$. Then,

$$\begin{aligned} \text{UD}(X, Y) &= 2 \sup_{A \in \mathcal{E}_X \otimes \mathcal{E}_Y} \{\mu_{(X, Y)}(A) - (\mu_X \times \mu_Y)(A)\} \\ &= 2 \sup_{A \in \mathcal{E}_X \otimes \mathcal{E}_Y} \{\mu_{(X, Y)}(A \cap C_Y) - (\mu_X \times \mu_Y)(A \cap C_Y)\} \\ &\quad + 2 \sup_{A \in \mathcal{E}_X \otimes \mathcal{E}_Y} \{\mu_{(X, Y)}(A \cap D_Y) - (\mu_X \times \mu_Y)(A \cap D_Y)\}. \end{aligned}$$

The first term is upper-bounded by

$$\begin{aligned} &2 \sup_{A \in \mathcal{E}_X \otimes \mathcal{E}_Y} \{\mu_{(X, Y)}(A \cap C_Y) - (\mu_X \times \mu_Y)(A \cap C_Y)\} \\ &\leq 2\mu_{(X, Y)}(C_Y) - 0 \\ &= 2\mu_Y(c_Y) \\ &= 2(1 - \sum_{y \in d_Y} \mu_Y(\{y\})), \end{aligned}$$

with equality if and only if there exists a set A such that $\mu_{(X, Y)}(A \cap C_Y) = \mu_Y(c_Y)$ and $\mu_X \times \mu_Y(A \cap C_Y) = 0$.

The second term is upper-bounded by

$$\begin{aligned} &2 \sup_{A \in \mathcal{E}_X \otimes \mathcal{E}_Y} \{\mu_{(X, Y)}(A \cap D_Y) - (\mu_X \times \mu_Y)(A \cap D_Y)\} \\ &= 2 \sum_{y \in d_Y} \sup_{B \in \mathcal{E}_X} \{\mu_{(X, Y)}(B \times \{y\}) - (\mu_X \times \mu_Y)(B \times \{y\})\} \\ &\leq 2 \sum_{y \in d_Y} \sup_{B \in \mathcal{E}_X} \{\min\{\mu_X(B), \mu_Y(\{y\})\} - \mu_X(B) \cdot \mu_Y(\{y\})\} \\ &= 2 \sum_{y \in d_Y} \sup_{B \in \mathcal{E}_X} \{\min\{\mu_X(B) \cdot (1 - \mu_Y(\{y\})), (1 - \mu_X(B)) \cdot \mu_Y(\{y\})\}\} \\ &\leq 2 \sum_{y \in d_Y} (\mu_Y(\{y\}) - \mu_Y(\{y\})^2), \end{aligned}$$

with equality if and only if for all $y \in d_Y$ we have that $\sup_{B \in \mathcal{E}_X} \{\mu_X(B) = \mu_Y(\{y\})\} = \sup_{B \in \mathcal{E}_X} \{\mu_{(X, Y)}(B \times \{y\})\}$.

Combining these two upper-bounds gives

$$\begin{aligned} \text{UD}(X, Y) &\leq 2\left(1 - \sum_{y \in d_Y} \mu_Y(\{y\})\right) + 2 \sum_{y \in d_Y} (\mu_Y(\{y\}) - \mu_Y(\{y\})^2) \\ &= 2 \left(1 - \sum_{y \in d_Y} \mu_Y(\{y\})^2\right). \end{aligned}$$

A.2.4. Functional dependence attains maximum UD If $Y = f(X)$ for a measurable function $f : E_X \rightarrow \mathbb{R}^m$, let $d_Y = \{y \in \mathbb{R}^m \mid \mathbb{P}(Y = y) > 0\}$ be the set of elements y with positive probability (which is a countable set) and let $c_Y = \mathbb{R}^m \setminus d_Y$. Then for any $\epsilon > 0$ there exists a partition $T_{1,\epsilon}, T_{2,\epsilon}, \dots, T_{k_\epsilon,\epsilon}$ of c_Y such that $\mathbb{P}(Y \in T_{i,\epsilon}) < \epsilon$, thus $\mu_X(f^{-1}(T_{i,\epsilon})) = \mathbb{P}(X \in f^{-1}(T_{i,\epsilon})) < \epsilon$. Now we define the set

$$B_\epsilon = \left(\bigcup_{i=1}^{k_\epsilon} (f^{-1}(T_{i,\epsilon}) \times T_{i,\epsilon}) \right) \cup \left(\bigcup_{y \in d_Y} (f^{-1}(\{y\}) \times \{y\}) \right),$$

then

$$\begin{aligned} \text{UD}(X, Y) &= 2 \sup_{A \in \mathcal{E}_X \otimes \mathcal{E}_Y} \{ \mu_{(X,Y)}(A) - (\mu_X \times \mu_Y)(A) \} \\ &\geq \sup_{\epsilon} \{ 2(\mu_{(X,Y)}(B_\epsilon) - (\mu_X \times \mu_Y)(B_\epsilon)) \} \\ &\geq \sup_{\epsilon} \left\{ 2\left(1 - \sum_{i=1}^{k_\epsilon} (\mu_X(f^{-1}(T_{i,\epsilon}))\mu_Y(T_{i,\epsilon})) - \sum_{y \in d_Y} \mu_X(f^{-1}(\{y\}))\mu_Y(\{y\}))\right) \right\} \\ &\geq \sup_{\epsilon} \left\{ 2\left(1 - \epsilon\mu_Y(c_Y) - \sum_{y \in d_Y} \mu_Y(\{y\})^2\right) \right\} \\ &= 2\left(1 - \sum_{y \in d_Y} \mu_Y(\{y\})^2\right). \end{aligned}$$

In Appendix A.2.3, we have determined the upper bound, which is also equal to $2(1 - \sum_{y \in d_Y} \mu_Y(\{y\})^2)$. Thus, UD is maximized. As corollary to this result, we find that $\text{UD}(Y, Y) = 0$ iff there exists y such that $\mathbb{P}(Y = y) = 1$ so iff Y is almost surely constant.

Note that the only property of Y being constrained to \mathbb{R}^m we really need is the fact that all atoms can be split into singletons and a null-set. So specifically, once we remove singletons with positive probability we are left with a non-atomic measurable

space. This therefore allows us to define partitions. The proof can therefore be easily transferred to any space where this condition holds.

Only for highly specific situations, we were unable to prove that the upper bound is tight and achieved by $\text{UD}(Y, Y)$. It specifically concerns the cases where there exist sets A with $\mu_Y(A) > 0$ and for all subsets $A' \subset A$ it holds that $\mu_Y(A') \in \{\mu_Y(A), 0\}$ and additionally for all elements $a \in A$ it holds that $\mu_Y(\{a\}) = 0$. We call these 'non-trivial atoms'. In practice non-trivial atoms are highly irregular. It is mostly interesting from a theoretical point of view (for the sake of completeness).

We conjecture that for this case the upper bound would be equal to $\text{UD}(X, Y) \leq 2(1 - \sum_{A \in B} \mu_Y(A)^2)$ where B is a set of atoms of μ_Y with one representative per equivalence class (where $A_1 \sim A_2$ if they differ by a null-set) and that this bound is attained for X for which $Y = f(X)$ (so in particular $X = Y$). However, as mentioned above this problem remains open.

A.2.5. Measurable functions never increase UD Let $f : (E_X, \mathcal{E}_X) \rightarrow (E_{X'}, \mathcal{E}_{X'})$ be a measurable function. Then $h : E_X \times E_Y \rightarrow E_{X'} \times E_Y$ with $h(x, y) = (f(x), y)$ is measurable. Now it follows that

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

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

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

In Appendix A.2.1, it is proven that UD is symmetric. Therefore, it also holds for $g : E_Y \rightarrow E_{Y'}$, that

$$\text{UD}(X, g(Y)) \leq \text{UD}(X, Y).$$

A.2.6. UD invariant under isomorphisms Using Appendix A.2.5, it must hold for all isomorphisms f, g that

$$\begin{aligned} \text{UD}(X, Y) &= \text{UD}(f^{-1}(f(X)), g^{-1}(g(Y))) \\ &\leq \text{UD}(f(X), g(Y)) \\ &\leq \text{UD}(X, Y). \end{aligned}$$

Therefore, all inequalities are actually equalities. In other words,

$$\text{UD}(f(X), g(Y)) = \text{UD}(X, Y).$$

References

- [1] AGARWAL, R., SACRE, P. AND SARMA, S. V. Mutual dependence: A novel method for computing dependencies between random variables 2015. 9, 11
- [2] CAPITANI, L., BAGNATO, L. AND PUNZO, A. (2014). Testing serial independence via density-based measures of divergence. *Methodology And Computing In Applied Probability* **16**, 627–641. 11, 13
- [3] EMBRECHTS, P., MCNEIL, A. J. AND STRAUMANN, D. (2002). *Correlation and Dependence in Risk Management: Properties and Pitfalls*. Cambridge University Press, Cambridge. p. 176–223. 8, 9
- [4] GEBELEIN, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* **21**, 364–379. 11
- [5] GRAMACKI, A. (2017). *Nonparametric Kernel Density Estimation and Its Computational Aspects* 1st ed. Springer Publishing Company, Incorporated, New York. 17
- [6] GRANGER, C. W., MAASOUMI, E. AND RACINE, J. (2004). A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis* **25**, 649–669. 9
- [7] GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. AND SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research* **6**, 2075–2129. 9
- [8] HELLINGER, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik* **1909**, 210–271. 11

- [9] 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 Journal* **14**, 2332–2337. 8
- [10] JOE, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* **84**, 157–164. 9
- [11] KIMELDORF, G. AND SAMPSON, A. R. (1978). Monotone dependence. *The Annals of Statistics* **6**, 895–903. 11
- [12] KRUSKAL, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association* **53**, 814–861. 2, 6
- [13] LANCASTER, H. O. (1963). Correlation and complete dependence of random variables. *The Annals of Mathematical Statistics* **34**, 1315–1321. 4, 5
- [14] MÓRI, T. F. AND SZÉKELY, G. J. (2019). Four simple axioms of dependence measures. *Metrika* **82**, 1–16. 9
- [15] PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. AND FLANNERY, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing* 3 ed. Cambridge University Press, USA. 5, 10, 11
- [16] RÉNYI, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica* **10**, 441–451. 3, 6, 9, 11
- [17] 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. 9
- [18] SUGIYAMA, M. AND BORGWARDT, K. M. (2013). Measuring statistical dependence via the mutual information dimension. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. IJCAI '13. AAAI Press, Beijing. p. 1692–1698. 9
- [19] SZÉKELY, G. J. AND RIZZO, M. L. (2009). Brownian distance covariance. <https://doi.org/10.1214/09-AOAS312> **3**, 1236–1265. 9, 11
- [20] WATANABE, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development* **4**, 66–82. 11