The Classic Cross-Correlation and the Real-Valued Jaccard and Coincidence Indices

Luciano da Fontoura Costa

The Classic Cross-Correlation and the Real-Valued Jaccard and Coincidence Indices

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

14th Nov 2021

Abstract

In this work we describe and compare the classic inner product and Pearson correlation coefficient as well as the recently introduced real-valued Jaccard and coincidence indices. Special attention is given to diverse schemes for taking into account the signs of the operands, as well as on the study of the geometry of the scalar field surface related to the generalized multiset binary operations underling the considered similarity indices. The possibility to split the classic inner product, cross-correlation, and Pearson correlation coefficient is also described.

'At the horizon line, where the sea meets the sky, countless sails.'

LdaFC

1 Introduction

Though not often realized, the correlation, as well as the closely related convolution binary operators (in the mathematical sense of taking two arguments) are among the most frequently employed operations in science and technology. Basically, the correlation between two functions $f(x)$ and $g(x)$ can be understood in terms of the *inner product*, which is a functional acting over the whole extent of both functions.

More specifically, the inner product can be understood as being related to the product of one of the vectors f (or functions) by the projection of the other onto f. Provided the magnitudes of the two vectors are kept constant, the inner product will also quantify the *similarity* between the two vectors, as gauged from the smallest angle between them.

Two other similarity approaches, namely the real-valued Jaccard and coincidence indices, have been recently proposed [1, 2, 3, 4], mainly based on extensions of the multiset theory to take into account real, possibly negative values.

In the present work, we aim at studying in some detail the structure and geometry of these three considered indices, namely the inner product as well as the real-valued Jaccard and coincidence indices. We start by presenting some basic concepts related to the inner product and data standardization, which is often applied to datasets and which implied negative respective values. Then, we revise and present several schemes that can be adopted to express the sign alignment between two real values (i.e. $xy > 0$ or $xy < 0$). The several new multiset binary operations (in the sense of taking two arguments) are then revised, which are involved in the considered similarity indices.

The real-valued Jaccard and coincidence indices are presented next, including an interesting result relating the classic inner product with two generalized multiset operations. The geometry and symmetry of the considered similarity indices is then approached from their respective versions adapted to two real scalar values. A striking geometry is observed for the real-valued Jaccard that closely resembles the generalized Kronecker delta function [5, 4].

The reported concepts and methods are also employed to propose a double Pearson correlation coefficient in which the effects of the values with same or opposite signs can be separated and controlled as a linear combination depending on a parameter α, in a manner similar to that adopted for the coincidence index in [6].

2 Basic Concepts

Given two vectors $\vec{x}$ and $\vec{y}$, both in $\mathbb{R}^N$, their respective inner (or scalar, or dot) product can be expressed as:

$$\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{N} x_i \; y_i = |\vec{x}||\vec{y}| \cos(\theta) \qquad (1)$$

where θ is the smallest angle between the two vectors and:

$$|\vec{x}| = \sqrt{\sum_{i=1}^{N} x_i^2}; \quad |\vec{y}| = \sqrt{\sum_{i=1}^{N} y_i^2} \tag{2}$$

Observe that the inner product is neither upper nor lower bound and can take positive, null (orthogonal vectors), or negative values.

The often adopted *cosine similarity* can be expressed as:

$$\cos(\theta) = \frac{\langle \vec{x}, \vec{y} \rangle}{|\vec{x}||\vec{y}|} \tag{3}$$

from which we see that this similarity index does not take into account the magnitudes of the operands x and y, but only the smallest angle between the two vectors. We also have that $-1 \leq \cos(\theta) \leq 1$.

Similarly, given two real functions $f(x)$ and $g(x)$, their *inner product* is defined as:

$$\langle f(x), g(x) \rangle = \int_S f(x)g(x)dx \tag{4}$$

where S is the combined support of both functions.

When applied to real functions, the inner product allows to consider the length, distance, orthogonality, and angle between functions in a manner analogous to that of the inner product applied to vectors. It is also interesting to contemplate the situation in which the inner product is applied to discretized versions of functions. Because functions are also vectors, in the sense of vector spaces, we will henceforth refer generically to both vectors and functions.

It is possible to derive a distance between two operands f and g from the inner product. First, we express the *norm* of a vector f in terms of the inner product as:

$$|f| = \sqrt{\langle f(x), f(x) \rangle} \tag{5}$$

so that we can now define the *distance* between two vectors f and g as:

$$d(f, g) = |f - g| = \sqrt{\langle f(x) - g(x) \rangle} \tag{6}$$

which corresponds to the *Euclidean distance* between f and g.

A *complete* inner product space is called a *Hilbert space* in functional analysis (e.g. [7, 8]), which can be informally understood as being an extension from vectors of linear algebra to real function spaces. In addition, the concept of proximity expressed in the inner product also relates to topological concepts. For instance, by *complete* it is meant that all Cauchy sequences in the metric space has a limit contained in that same space, which can be informally understood as the space not having 'gaps' or 'holes'.

Let X and Y be any two random variables described by respective density probabilities $p(x)$ and $p(y)$.

Their *average* and *variance* can be defined as:

$$\mu_X = \int_S xp(x)dx \tag{7}$$

$$\mu_Y = \int_S yp(y)dy \tag{8}$$

$$\sigma_X^2 = \int_S (x - \mu_X)^2 p(x)dx \tag{9}$$

$$\sigma_Y^2 = \int_S (y - \mu_Y)^2 p(y)dx \tag{10}$$

The respective *standard deviations* are:

$$\sigma_X = +\sqrt{\sigma_X^2} \tag{11}$$

$$\sigma_Y = +\sqrt{\sigma_Y^2} \tag{12}$$

In case the joint density probability $p(x, y)$ is known, we can define the *covariance* between X and Y as:

Given a random variable X, it can be *standardized* as:

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X} \tag{13}$$

The standardization procedure is often employed, especially to make a set of random variables more commensurate, therefore avoiding those variable with larger magnitudes to dominate. However, the decision to standardize or not depends on each specific data and problem. After standardization, each of the random variables will have zero means and unit standard deviation. In addition, most of the values will result inside the interval $[-2, 2]$.

The unbiased *covariance* between two random variables X and Y represented in terms of respective samples $\vec{x} = x_1, x_2, \ldots, x_N$ and $\vec{y} = y_1, y_2, \ldots, y_N$ can be estimated as:

$$cov(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} [x_i - \mu_X] [y_i - \mu_Y] \tag{14}$$

which can be understood as a normalized inner product, in the sense that:

$$cov(X, Y) = \frac{1}{N-1} \langle X - \mu_X, Y - \mu_Y \rangle \tag{15}$$

Interestingly, when the cross-correlation is taken on standardized vectors $\vec{x}$ and $\vec{y}$, it becomes identical to the *Pearson correlation coefficient* $-1 \leq P(x, y) \leq 1$.

In summary, we have seen that the classic inner product, the L2 norm, the cosine similarity, the Euclidean distance, the cross-correlation, the covariance, and the Pearson correlation coefficient are all directly related to the inner product between two vector or function operands.

3 Conjoint Sign Functions

Given two signals $x(t)$ and $y(t)$, their respective sign functions can be expressed as:

$$s_x = s_x(t) = sign(x(t)) \tag{16}$$
$$s_y = s_y(t) = sign(y(t)) \tag{17}$$

We can now define the following *conjoint sign functions*:

$$s_p = |s_x + s_y| \tag{18}$$
$$s_m = |s_x - s_y| \tag{19}$$
$$s_{hp} = |s_x + s_y|/2 \tag{20}$$
$$s_{hm} = |s_x - s_y|/2 \tag{21}$$
$$s_{xy} = s_x s_y \tag{22}$$

Figure 1 illustrates the functions x_{h+}, x_{h-}, and x_{xy} respectively to $x = \sin(x)$ and $y = \cos(x)$.

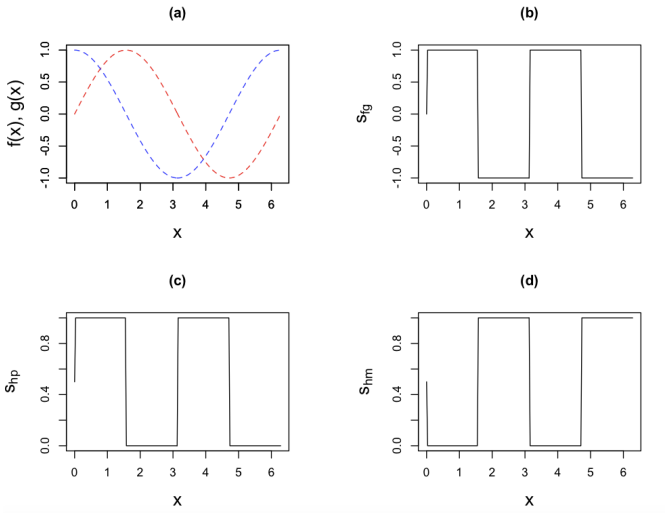


Figure 1: Functions $f(x)$ and $g(x)$ corresponding to a whole period of sine and cosine, and the respective conjoint sign functions s_{fg}, s_{hp} and s_{hm}. Observe that $s_{fg} = s_{hp} - s_{hm}$ and also that the $y-$axis in (b-d) have different limits.

The function in Equation 20 has been used in [9], and that in Equation 22 has been employed in [10, 11], both related to the L1 norm. The latter function has also appears in the *minmod* slope limiting function adopted in partial differential equations (e.g. [12]). The function in Equation 21 has been used in [6, 3].

We also have that:

$$s_{hp} = 1 - s_{hm} \tag{23}$$
$$s_{hm} = 1 - s_{hp} \tag{24}$$
$$s_{xy} = s_p - 1 \tag{25}$$
$$s_{xy} = 1 - s_m \tag{26}$$
$$s_{xy} = s_{hp} - s_{sm} \tag{27}$$
$$\tag{28}$$

The generalized Kronecker delta function has been suggested [3, 4] as a means to express not only same sign similarity as in the traditional Kronecker delta, but also opposite sign relationships. It can be expressed as:

$$\delta^{\pm}_{x,y} = \begin{cases} 1 & \iff x = y,\ x, y \neq 0 \\ 0 & \iff x = y = 0 \\ -1 & \iff x = -y,\ x, y \neq 0 \end{cases} \tag{29}$$

Figure 2 illustrates the generalized Kronecker delta function.

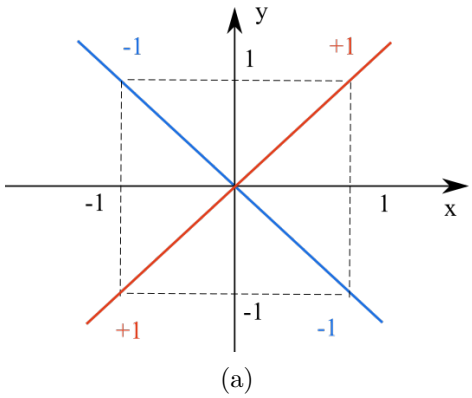


Figure 2: The generalized Kronecker delta function allows the strictest quantification of similarities between values with the same or opposite signs, providing a reference for other similarity indices.

As developed in [3], the generalized Kronecker delta function plays a critically important role in defining the most strict similarity test, from which other more tolerant similarity indices including the inner product and real-valued Jaccard index can be directly related.

4 Generalized Multisets

Multisets (e.g. [13, 14, 15, 16, 17, 18]) provide an intuitive and interesting extension of the classic set theory (e.g. [13, 19]) so as to allow the repetition of elements. Generalized versions of multisets [1, 2, 3, 4] have been developed to allow real, possibly negative multiplicities. Basically, the multiset subtraction is allowed to take negative values, so that the complement of a multiset A can be performed as $\Phi - A$, where the null multiset takes the

place of the universe set in set theory. The generalization of multiplicities to take real, possibly negative values [1, 2, 3, 4] allow new binary operations to be defined between two multifunctions $f(z)$ and $g(z)$, including but not being limited to:

$$f \cap g = \int_S \min\{f, g\}\, dz \tag{30}$$

$$f \cup g = \int_S \max\{f, g\}\, dz \tag{31}$$

$$f \sqcap g = \int_S s_{fg} \min\{s_f f, s_g g\}\, dz \tag{32}$$

$$f \sqcup g = \int_S s_{fg} \max\{s_f f, s_g g\}\, dz \tag{33}$$

$$f \sqcap_- g = \int_S |s_f - s_g|/2 \min\{s_f f, s_g g\}\, dz \tag{34}$$

$$f \sqcap_+ g = \int_S |s_f + s_g|/2 \min\{s_f f, s_g g\}\, dz \tag{35}$$

$$f \sqcup_- g = \int_S |s_f - s_g|/2 \max\{s_f f, s_g g\}\, dz \tag{36}$$

$$f \sqcup_+ g = \int_S |s_f + s_g|/2 \max\{s_f f, s_g g\}\, dz \tag{37}$$

$$f \tilde{\sqcap} g = \int_S \min\{s_f f, s_g g\}\, dz \tag{38}$$

$$f \tilde{\sqcup} g = \int_S \max\{s_f f, s_g g\}\, dz \tag{39}$$

Equations 30 and 31 correspond to the multiset counterparts of the set theory operations of intersection and union. However, we do not have, as could expected, that the intersection of a generic multiset x and the null multiset Φ corresponds to the null multiset. On the contrary, we typically have that:

$$x \cap \Phi \neq \Phi \tag{40}$$

The binary operations in Equations 32 and 33 can be understood as the intersection and union considering negative multiplicities. Now, we do have that:

$$x \sqcap \Phi = \Phi \tag{41}$$

Equations 35 and 34 can be understood as the intersection operation acting only when x and y have the same or opposite signs. Equations 37 and 36 are have the analogous effect regarding union.

Equations 38 and 39 can be understood as the intersection and union acting on the absolute values of multiplicities.

These several operations, which allow great flexibility for taking into account diverse combinations of operands signs, provide the basis for obtaining the similarity indices considered in the present work.

5 The Real-Valued Jaccard and Coincidence Indices

It has been described [1, 2, 3, 4] that, when generalized to real, possibly negative multiplicities, the Jaccard index becomes:

$$\mathcal{J}_R(f, g) = \frac{\int_S s_{fg} \min\{s_f f, s_g g\}\, dz}{\int_S \max\{s_f f, s_g g\}\, dz} = \frac{f \sqcap g}{f \tilde{\sqcup} g} \tag{42}$$

which has been called the *real-valued Jaccard similarity index*.

Interestingly, it can be verified that the following equation is identical to the previous one, therefore providing an alternative definition for the real-valued Jaccard index:

$$\mathcal{J}_R(f, g) = \frac{\int_S [f(z)\ g(z)]\, dz}{\left[\int_S \max\{s_f f, s_g g\}\, dz\right]^2} = \frac{\langle f, g\rangle}{[f \tilde{\sqcup} g]^2} \tag{43}$$

Thus, we have that:

$$\mathcal{J}_R(f, g) = \frac{f \sqcap g}{f \tilde{\sqcup} g} = \frac{\langle f, g\rangle}{[f \tilde{\sqcup} g]^2} \tag{44}$$

which then implies:

$$\boxed{\langle f, g\rangle = [f \sqcap g]\, [f \tilde{\sqcup} g]} \tag{45}$$

This result illustrates the flexibility of the generalized multiset operations described in Section 4. In addition, it establishes an important link between the real-valued Jaccard, as well as the coincidence indices, with the classic inner product. As a matter of fact, this result shows that the cross correlation, which consists in the successive sliding application of the inner product, can actually be used for calculation of the real-valued Jaccard index, and vice versa, provided the proper normalization is taken into account. This important link will be further considered in Section 6 in order to better understand the properties of the similarity indices considered in this work.

Given that the Jaccard index is not capable of taking into account the relative interiority between the two compared sets, vectors or functions [1], it has been complemented by considering the interiority index (also overlap [20]) which, when adapted to real, possibly negative values yields:

$$\mathcal{I}_R(f, g) = \frac{\int_S \min\{s_f f, s_g g\}\, dx}{\min\{S_f, S_g\}} = \frac{f \sqcap g}{\min\{S_f, S_g\}} \tag{46}$$

where:

$$S_f = \int_S s_f f(z) dz \tag{47}$$

$$S_g = \int_S s_g g(z) dz \tag{48}$$

The *coincidence index* can now be expressed [1, 2, 3, 4] as corresponding to the product between the real-valued Jaccard and interiority indices:

$$\mathcal{C}_R(f,g) = \frac{[f \sqcap g]\,[f \tilde{\sqcap} g]}{[f \tilde{\sqcup} g]\,\min\{S_f, S_g\}} \tag{49}$$

or, in expanded notation:

$$\mathcal{C}_R(f,g) = \\ = \frac{\left[\int_S s_{fg}\,\min\{sf_f, s_g g\}\,dz\right]\,\left[\int_S \min\{s_f f, s_g g\}\,dz\right]}{\left[\int_S \max\{sf_f, s_g g\}\,dz\right]\,[\min\{S_f, S_g\}]} \tag{50}$$

6 The Geometry of Similarity

In this section we study in some detail the geometry and symmetries of the above presented similarity indices in order to better understand the properties and effects of each index.

We start by considering the real-valued Jaccard similarity index rewritten for two real scalar values $f = x$ and $g = y$:

$$\mathcal{J}_R(x,y) = \frac{s_{xy}\,\min\{|x|,|y|\}}{\max\{s_x x, s_y y\}} \tag{51}$$

We can then separate the numerator and denominator as:

$$A_1(x,y) = s_{xy}\,\min\{s_x x, s_y y\} = x \sqcap y \tag{52}$$
$$A_2(x,y) = \max\{s_x x, s_y y\} = x \tilde{\sqcup} y \tag{53}$$

which, as two scalar fields on the (x,y) domain, can be visualized.

We will also consider the following additional fields, directly related to the inner product:

$$A_3(x,y) = x\ y \tag{54}$$
$$A_4(x,y) = [\max\{s_x x, s_y y\}]^2 = [x \tilde{\sqcup} y]^2 \tag{55}$$

as well as the multiset operation:

$$A_5(x,y) = \min\{s_x x, s_y y\} = x \tilde{\sqcap} y \tag{56}$$

When rewritten for scalar real values x and y, we have that $f \tilde{\sqcap} g = \min\{S_x, S_y\}$ and therefore the coincidence index becomes:

$$\mathcal{C}_R(x,y) = \frac{[x \sqcap y]\,[x \tilde{\sqcap} y]}{[x \tilde{\sqcup} y]\,[x \tilde{\sqcap} y]} = \frac{x \sqcap y}{x \tilde{\sqcup} y} = \mathcal{J}_R(x,y) \tag{57}$$

However, observe that this will not, in general, be the case with higher dimensional vector operands, in which case the real-valued Jaccard and coincidence index will be typically distinct.

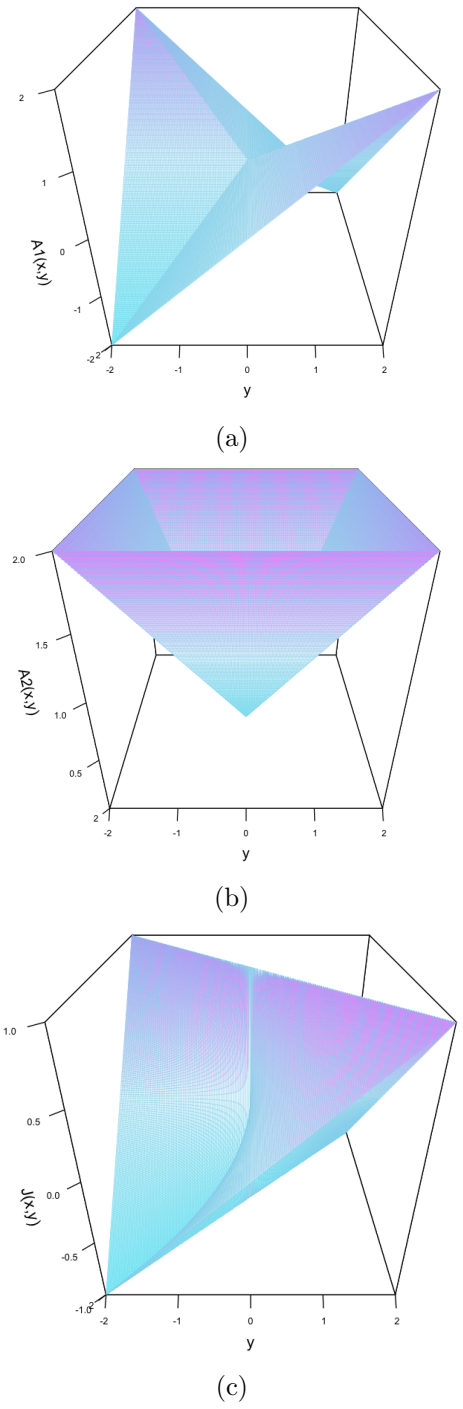


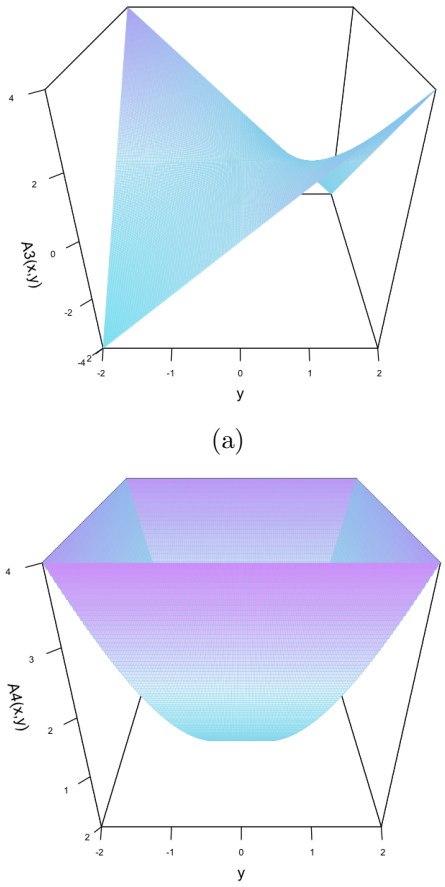
Figure 3: The scalar fields $A_1(x,y)$ in Equation 52 (a); $A_2(x,y)$ in Equation 53 (b); and the real-valued Jaccard index $\mathcal{J}_R(x,y) = A_1(x,y)/A_2(x,y$ (c) for $-2 \leq x \leq 2$ and $-2 \leq y \leq 2$. Observe the striking geometry of the real-valued Jaccard index shown in (c), which is closely related to the generalized Kronecker delta function.

Figure 3 illustrates the fields $A_1(x,y)$ (a) and $A_2(x,y)$ (b) for $-2 \leq x \leq 2$ and $-2 \leq y \leq 2$.

Observe the striking geometry of the surface in Figure 3(c). As a more detailed verification will reveal, this function resembles closely the generalized Kronecker delta

function [5]. Observe the gradual, linear rotation from the identity line $x = y$ to the anti-identity line $x = -y$. It is this specific geometry of the real-valued Jaccard index that contributes to enhanced performance for template matching observed in [21, 5].

Figure 10 presents the multiset operations in Equations 54 and 55, the former of which being directly related to the inner-product.



(a)

(b)

Figure 4: The scalar fields $A_3(x, y)$ in Equation 55 (a); $A_2(x, y)$ in Equation 53 (b); for $-2 \leq x \leq 2$ and $-2 \leq y \leq 2$. Recall that, for scalar values, $\mathcal{C}_R(x, y) = \mathcal{J}_R(x, y) = A_3(x, y)/A_4(x, y) = A_1(x, y)/A_2(x, y)$, which is shown in Fig. 3(c).

Figure 5 depicts the multiset absolute union operation $x\tilde{\sqcap}y$, which plays an important role in both the interiority and coincidence indices.

We are now in position to analyze more closely the geometry of the cross-correlation and real-valued Jaccard index. Given the quadrant symmetries of these indices, we henceforth focus our attention only on the first quadrant, $x \geq 0$ and $y \geq 0$.

Let's consider the real-valued Jaccard similarity applied

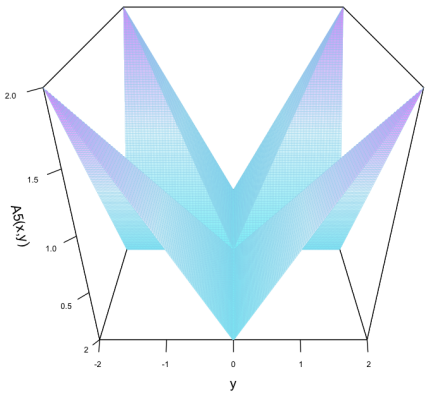


Figure 5: The scalar field obtained for the multiset operation $x\tilde{\sqcap}y$, where x and y are real scalar values, for $-2 \leq x \leq 2$ and $-2 \leq y \leq 2$.

to real scalar values. We have the following situations:

$$\begin{cases} x > y & \iff \mathcal{J}_R(x, y) = \frac{y}{x} \\ x = y = 0 & \iff \mathcal{J}_R(x, y) = 0 \\ x < y & \iff \mathcal{J}_R(x, y) = \frac{x}{y} \end{cases} \tag{58}$$

Given the involved symmetries, we can further restrict our attention to $x > y$.

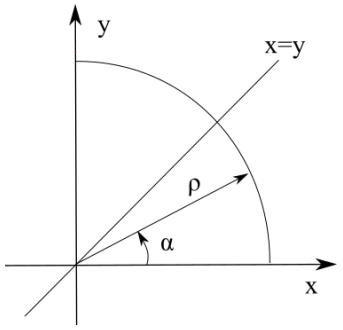


Figure 6: The geometrical construction adopted for studying the geometry of the scalar field defined by the real-valued Jaccard similarity index assuming scalar operands in the case of $x > y$.

If we fix our attention to the points (x, y), with $x > y$, so that $\rho = \sqrt{x^2 + y^2}$ is equal to a fixed constant $\tilde{\rho}$, we will find that:

$$\mathcal{J}_R(\alpha, \rho = \tilde{\rho}) = \frac{y}{x} = \tan(\alpha) \tag{59}$$

Therefore, we have that the real-valued Jaccard similarity along the semi-circle defined by α and $\tilde{\rho}$ increases with $\tan\alpha$ as we go from $\alpha = 0$ to $\alpha = \pi/4$. At $\alpha = \pi/4$, we have $\mathcal{J}_R(x, y) = \tan(\pi/4) = 1$, which corresponds to the classic Kronecker delta function.

The above geometry analysis reveals that the real-valued Jaccard similarity index corresponds to a version of the Kronecker delta function in which the similarity decreases with the tangent of α as one rotates from the

maximum crest corresponding to the classic Kronecker delta function. As such, the real-valued Jaccard indeed implements a more strict quantification of the similarities, as described in [1, 3, 4].

The rotation from the identity to the anti-identity line which, as seen above, follows the tangent of the angle, implies that the sectioning of the height of the real-valued Jaccard scalar field by level sets with the same tolerance will yield sections with similar area, as illustrated in Figure 7, therefore implying that the similarity quantification by the real-valued Jaccard scalar field is nearly uniform with the angle and image values.

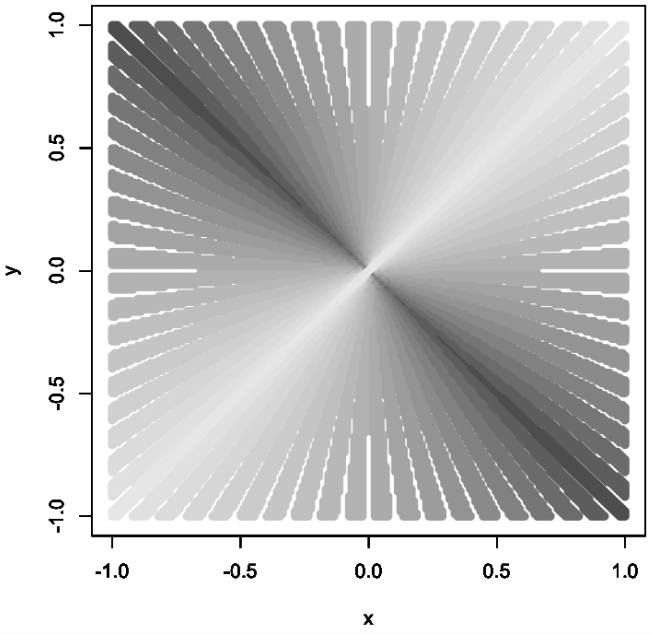


Figure 7: The partitioning of the real-valued Jaccard function by allowing fixed similarity tolerances yields, as a consequence of the tangent mapping identified in this work, a substantially uniform coverage of the similarities. Level sections shown for 20 uniform similarity intervals.

If necessary, the real-valued Jaccard function can be immediately modified to yield perfectly uniform sections, which can be obtained by making:

$$\mathcal{J}_U(x,y) = \arctan(\mathcal{J}_R(x,y)) \tag{60}$$

which is illustrated in Figure 8. This version of the real-valued Jaccard similarity index may be called the *isotropic Jaccard index*. The coincidence index can be modified in a similar manner.

It is also interesting to consider the possibility of having:

$$\mathcal{J}_R(f,g) = \frac{[f \sqcap g]^D}{f \tilde{\sqcup} g} \tag{61}$$

Figure 9 illustrates the scalar versions of the real-valued Jaccard index with the numerator taken to the power of

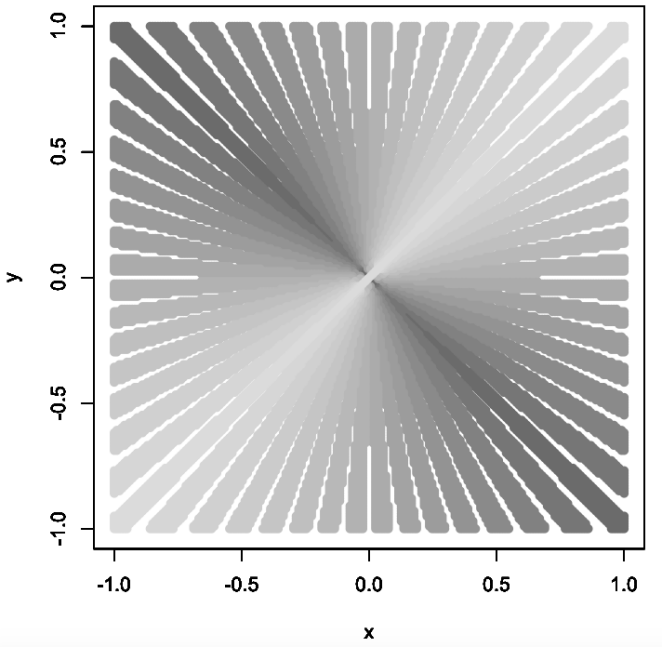


Figure 8: The partitioning of the real-valued Jaccard function into perfectly homogenous level sections, obtained by taking the arctangent of the real-valued Jaccard similarities. Level sections shown for 20 uniform similarity intervals.

D for $D = 2, 3, 5$ and 21. Two particularly interesting properties can be observed. First, we have that even values of D will imply the real-valued Jaccard index to become related to the absolute value of the generalized Kronecker delta function, which can be of interest for certain applications. Second, it is interesting to observe that the real-valued Jaccard index converges to the generalized Kronecker delta function as $D \to \infty$, for D odd.

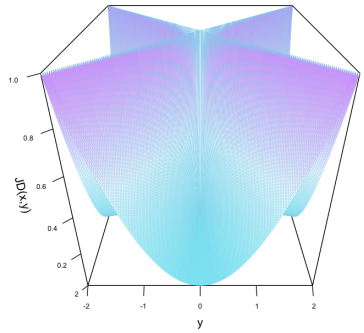
The above results indicate that the poser D controls how much the real-valued Jaccard index is strict regarding the quantification of similarity. More strict similarity quantifications will be characterized by steeper crests in the respectively obtained geometries. A similar result can be considered for the case of the coincidence index.

7 The Double Pearson Coefficient

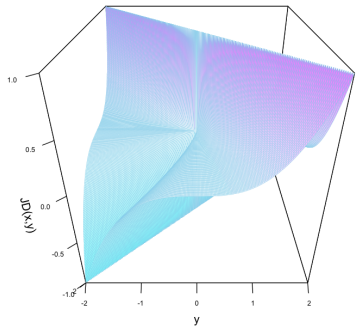
As described recently in [6], it is interesting to split Equation 32 as a combination of the indices proposed in [9], i.e.:

$$f \sqcap g = 2\,[\alpha]\,[f \sqcap_+ g] - 2\,[1-\alpha]\,[f \sqcap_- g] \tag{62}$$

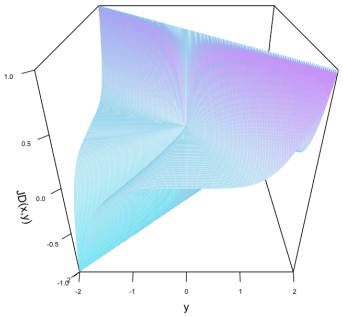
where $0 \leq \alpha \leq 1$ controls the contribution of the pairs of values x and y that have the same or opposite signs on the resulting integration. This resource has proven to allow an effective means for obtaining progressions of datasets represented as complex networks that are in-



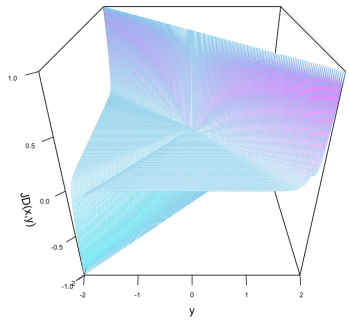
(a)



(b)



(c)



(d)

Figure 9: The scala version of the real-valued Jaccard index with the numerator taken to the power of D for $D = 2$ (a), 3, (b), 5 (c), and 21 (d). Observe the convergence of this index to the generalized Kronecker delta function as $D \to \infty$, D odd.

creasingly more connected for increasing values of α.

The concepts and methods reviewed and reported in this work paves the way to obtaining an analogous decomposition of the classic inner product as well as its normalized version known as the Pearson correlation coefficient.

Let's define the functionals:

$$\langle f, g \rangle_{-} = \int_S |s_x - s_y|/2 \; xy \; dz \tag{63}$$

$$\langle f, g \rangle_{+} = \int_S |s_x + s_y|/2 \; xy \; dz \tag{64}$$

$$\tag{65}$$

The double inner product can therefore be written as:

$$\langle f, g \rangle = [\alpha] \langle f, g \rangle_{-} - [1 - \alpha] \langle f, g \rangle_{+} \tag{66}$$

It may also be interesting to consider the two split terms separately in order to provide additional information about the joint variation of the two operands x and y.

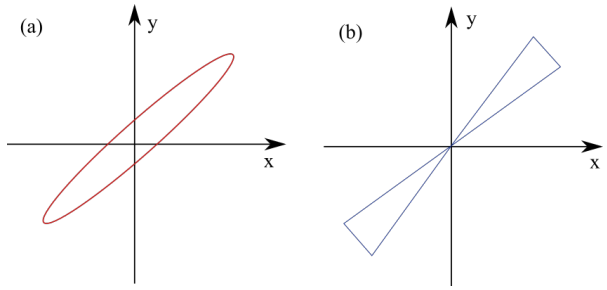


Figure 10: Two rather distinct point distributions which could have the same inner product or Pearson correlation coefficient. The consideration of the double Pearson correlation suggested in this work allows these situations to be effectively distinguished. For instance, in the case of this specific example, we would have that $\langle f, g \rangle_{-}$ would be non-zero at $(x = 0, y = 0)$ for (a) and zero for the point distribution in (b).

8 Concluding Remarks

The present work has addressed the properties of the inner product as well as the real-valued Jaccard and coincidence indices, with emphasis on the several interesting schemes that can be employed to take into account the sign of the operands. These developments are strongly based on the generalization of multisets to take into account real, possibly negative values. In particular, these generalizations allow several additional binary operators between multisets to be defined, many of which have been presented here.

Special attention has been give to the geometrical characterization of the surfaces arising by the several involved multiset operators which, for scalar values of the operands x and y, define respective scalar fields that can be conveniently visualized. This approach has allowed us to observe that the real-valued Jaccard and coincidence indices

present a geometry that closely resembles the generalized Kronecker delta function, involving a rotation from the identity line $x = y$ to the anti-identity line $x = -y$ following the tangent function. It has also been verified that taking the numerator of the real-valued Jaccard index to the power of D, with D odd, provides an effective manner to control de degree of how much strict the similarity quantification is performed. In particular, these functions converge to the generalized Kronecker delta as $D \to \infty$, D odd.

The reported approach also motivates the consideration of local and global properties of the geometry of the obtained surfaces, such as the respective gradients, as a means to formally specify criteria for the similarity quantification. For instance, one may aim at achieving minimum variation of the gradient magnitude as one moves from $x = y$ to $x = -y$.

The described concepts also paved the way to developing a double Pearson correlation coefficient, in which the contribution of the values x and y with the same or opposite signs can be separated and taken as a linear combination controlled by a respective parameter $0 \leq \alpha \leq 1$, in a similar manner to that described recently in [6, 22].

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

References

[1] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index, 2021.

[2] L. da F. Costa. Multisets. https://www.researchgate.net/publication/355437006_Multisets, 2021.

[3] L. da F. Costa. On similarity. https://www.researchgate.net/publication/355792673_On_Similarity, 2021.

[4] L. da F. Costa. Generalized multiset operations. https://www.researchgate.net/publication/356191988_Generalized_Multiset_Operations, 2021. [Online; accessed 10-Nov-2021].

[5] L. da F. Costa. Multiset neurons. https://www.researchgate.net/publication/356042155_Common_Product_Neurons, 2021.

[6] L. da F. Costa. Coincidence complex networks. https://www.researchgate.net/publication/355859189_Coincidence_Complex_Networks, 2021.

[7] E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley, 1989.

[8] W. Rudin. *Functional Analysis*. McGraw-Hill, 1991.

[9] B. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Publisher, Dordrecth, 1996.

[10] C. E. Akbas, A. Bozkurt, M. T. Arslan, H. Aslanoglu, and A. E. Cetin. L1 norm based multiplication-free cosine similiarity measures for big data analysis. In *IEEE Computational Intelligence for Multimedia Understanding (IWCIM)*, France, Nov. 2014.

[11] C. E. Akbas, A. Bozkurt, A. E. Cetin, R. Cetin-Atalay, and A. Uner. Multiplication-free neural networks. In *Signal Processing and Communications Applications Conference (SIU)*, Malatya, Turkey, May. 2015.

[12] B. Popov and O. Trifonov. Order of convergence of second order schemes based on the minmod limiter. *Math. Comp.*, 75:1735–1753, 2006.

[13] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[14] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[15] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[16] W. D. Blizard. The development of multiset theory. *Modern Logic*, 4:319–352, 1991.

[17] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[18] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.

[19] W. Rudin. *Elemets of Algebraic Topology*. Addison-Wesley, 1984.

[20] M. K. Vijaymeena and K. Kavitha. A survey on similarity measures in text mining. *Machine Learning and Applications*, 3(1):19–28, 2016.

[21] L. da F. Costa. Comparing cross correlation-based similarities. https://www.researchgate.net/publication/355546016_Comparing_Cross_Correlation-Based_Similarities, 2021.

[22] L. da F. Costa. A caleidoscope of datasets represented as networks
by the coincidence methodology. https://www.researchgate.net/publication/356392287_A_Caleidoscope_of_Datasets_Represented_as_Networks_by_the_Coincidence_Methodology, 2021.