CROATICA CHEMICA ACTA
CCACAA **79** (3) 399–410 (2006)

ISSN-0011-1643
CCA-3106
Original Scientific Paper

The Metric Index

Matevž Bren[a,c,*] and Vladimir Batagelj[b,c]

[a]*Faculty of Organizational Sciences, University of Maribor, Kidričeva 55a, 4000 Kranj, Slovenia*

[b]*Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia*

[c]*Institute of Mathematics, Physics and Mechanics, University of Ljubljana, 1000 Ljubljana, Slovenia*

RECEIVED NOVEMBER 28, 2005; REVISED MARCH 14, 2006; ACCEPTED APRIL 7, 2006

Keywords
dissimilarity
similarity and association coefficients
dissimilarity spaces
metric index
metric spaces

The power transformation that turns an arbitrary even dissimilarity into a semidistance or a definite dissimilarity into a distance is discussed. A method for the metric index computation is deducted and applied to determine the metric indices of 19 standard dissimilarity measures on dichotomous data.

INTRODUCTION

In many contexts, such as drug screening, reaction/no reaction testing or presence/absence sampling, educational and psychological tests... collected data are of dichotomous type and therefore units are presented with binary vectors. If we aim to present such units graphically or classify them into classes, the first step is to calculate 'distances' between units or classes of units. A group of dissimilarity measures designed for binary vectors is popularly known as matching coefficients.

In his article,[1] Hubálek gives a list of 43 association (affinity) or similarity (resemblance) coefficients and stresses:

"The purpose of this survey is to compare both theoretically and empirically a substantial majority of association and similarity coefficients (used, or just proposed), and to select a group of those which will be found most useful and theoretically least objectionable."

Continuing his work from this and other lists, we have chosen and compared 22 association coefficients.[2] We found classes of order equivalent coefficients and we calculated distances between these classes.

In the paper of Verbanac *at al.*,[3] the authors comment on this approach:

"Use of Daylight binary fingerprints as structural descriptors and Tanimoto coefficient as a similarity measure is a common practice, in spite of their limitations."

In this paper, we discuss a further step to a better knowledge and adequate application of association and similarity coefficients. Namely, some data analysis methods demand that dissimilarities used should be Euclidean or metric or..., but dissimilarity measures used in applications often do not have all these nice properties (see Table I). For example, graphical representations of multivariate data, widely used in research and applications of many disciplines, are based on techniques of representing a set of observations by a set of points in a low-dimensional real (usually) Euclidean vector space, so that observations that are similar to one another are represented by points that are close together. Multidimensional scaling (MDS)

* Author to whom correspondence should be addressed. (E-mail: matevz.bren@fov.uni-mb.si)

techniques belong to this family of graphical representations. Here, different approximations give rise to the different techniques of MDS: Metric MDS and Nonmetric MDS – see Cox and Cox.[4] Assumption of the metric MDS techniques (called Classical Scaling or Principal Coordinates Analysis) is that the dissimilarities between observations are distances within a set of points in some Euclidean space.

We discuss power transformation as a method that transforms nonmetric dissimilarities into metric distances, or nonmetric dissimilarity measures into metric dissimilarity measures. There is a simple geometric interpretation of this passage from nonmetric to metric: any power between 0 and 1 of dissimilarity d increases the short distances – usually the range of dissimilarity is [0, 1], relatively more than the long distances, retaining the metric inequality to hold for 'metric triples' and forcing it to hold for 'nonmetric triples'. Also, if the range of a dissimilarity is $[0, \infty]$, the power transformation d^α for α between 0 and 1 is a concave function and therefore it has the sub-additivity property that retains the metric inequality to hold for the 'metric triples' and forces it to hold for 'nonmetric triples'.

As a brief example, if the values of dissimilarity d between units X, Y and Z are $d(X, Y) = 0.49$, $d(X, Z) = 0.25$ and $d(Y, Z) = 0.25$, the triangle inequality holds

$$d(X, Y) = 0.49 \leq 0.25 + 0.25 = d(X, Z) + d(Z, Y)$$

and if we calculate the square root of dissimilarity d,

$$d^{\frac{1}{2}}(X, Y) = 0.7 \leq 0.5 + 0.5 = d^{\frac{1}{2}}(X, Z) + d^{\frac{1}{2}}(Z, Y)$$

the triangle inequality remains. But if we begin with 'nonmetric triple' U, V and W with the dissimilarities $d(U, V) = 0.81$, $d(U, W) = 0.25$ and $d(V, W) = 0.25$, the triangle inequality does not hold

$$d(U, V) = 0.81 \not\leq 0.25 + 0.25 = d(U,W) + d(V,W)$$

and calculating the square root of dissimilarity d,

$$d^{\frac{1}{2}}(U, V) = 0.9 \leq 0.5 + 0.5 = d^{\frac{1}{2}}(U, W) + d^{\frac{1}{2}}(V, W)$$

we gain the triangle inequality. If we continue this calculation with even lower values of the exponent, for example 1/4

$$d^{\frac{1}{4}}(U, V) = 0.948 \leq 0.707 + 0.707 = d^{\frac{1}{4}}(U, W) + d^{\frac{1}{4}}(V, W)$$

the triangle inequality remains and if we continue, at the end, we get the discrete distance for the 0 value of the exponent.

Any exponent higher than 1 has the opposite effect, making it more difficult for triangles to close in the dissimilarity space.

Since our idea is to transform a given nonmetric dissimilarity measure into a metric distance or semi-distance, the natural task is to find the highest value of the exponent that forces all triangles to close in a dissimilarity space. With this exponent, the changes caused by the power transformation of the given dissimilarities between units are the least possible.

We call this highest value of the exponent the metric index of the dissimilarity measure.

In this paper, we

- develop a method for determining the metric index for even nonnegative dissimilarity measures on binary vectors, and
- evaluate the exact values of metric indices for 19 dissimilarity coefficients.

PRELIMINARIES

Dissimilarities

Consider a set of units $\mathcal{E}$. Function d is a dissimilarity measure on the set of units $\mathcal{E}$ if and only if it is

P1. symmetric: $d(X, Y) = d(Y, X)$ for all $X, Y \in \mathcal{E}$, and

P2. straight: $d(X, X) \leq d(Y, X)$ for all $X, Y \in \mathcal{E}$.

A dissimilarity measure d that is

D1. nonnegative: $d(X, Y) \geq 0$ for all $X, Y \in \mathcal{E}$, and

D2. vanishes on the diagonal: $d(X, X) = 0$ for all $X \in \mathcal{E}$

is a dissimilarity on $\mathcal{E}$. The ordered pair $(\mathcal{E}, d)$ is a dissimilarity space. We denote by $\triangle_+$ the set of all dissimilarities on $\mathcal{E}$.

Moreover, a dissimilarity measure d on $\mathcal{E}$ is said to be

D3. definite if and only if $d(X, Y) = 0 \Rightarrow X = Y$;

D4. even (semi-definite) if and only if $d(X, Y) = 0 \Rightarrow$ for all $Z \in \mathcal{E}$: $d(X, Z) = d(Y, Z)$;

D5. metric if and only if the triangle inequality for all $X, Y, Z \in \mathcal{E}$: $d(X, Y) \leq d(X, Z) + d(Y, Z)$ holds.

D6. ultrametric if and only if the ultrametric inequality for all $X, Y, Z \in \mathcal{E}$: $d(X, Y) \leq \max\{d(X, Z), d(Y, Z)\}$ holds.

D7. additive if and only if Buneman's inequality or four-point condition $d(X, Y) + d(U, V) \leq \max(d(X, U) + d(Y, V), d(X, V) + d(Y, U))$ holds for all $X, Y, U, V \in \mathcal{E}$.

These properties are related as follows:

$$\text{D6} \Rightarrow \text{D5} \Rightarrow \text{D4} \Leftarrow \text{D3 and D7} \Rightarrow \text{D5}.$$

See also the list of dissimilarity measures on binary vectors and their properties in Table I.

Metric dissimilarity d on $\mathcal{E}$ is called a semi-distance. Naturally, a dissimilarity space $(\mathcal{E}, d)$ is a semi-metric space if and only if d is a semi-distance. $(\mathcal{E}, d)$ is a metric space if and only if d is also definite; then, d is a distance or a metric. We denote the subset of $\mathfrak{D}_+$ comprising all semi-distances by $\mathfrak{D}_\infty$ – the notation established in Classification and Dissimilarity Analysis.[5]

For all $\alpha > 0$, we introduce the power transformation by

$$d^\alpha(X, Y) := (d(X, Y))^\alpha \text{ for all } X, Y \in \mathcal{E}$$

where d is a nonnegative dissimilarity measure on $\mathcal{E}$; and for $\alpha = 0$, we define d^0 to be the discrete distance

$$d^0(X, Y) := \begin{cases} 0 & \text{for } X = Y, \\ 1 & \text{otherwise.} \end{cases}$$

We explicitly use the sign := for definitions and new notations.

The power transformation preserves properties D1 – D4 and D6. Note that we defined the power transformation for exponent $\alpha \geq 0$. If we define also

$$d^\infty(X, Y) := \lim_{\alpha \to \infty} d^\alpha(X, Y) \quad \text{for all} \quad X, Y \in \mathcal{E}$$

with d^∞, we can lose D4, which is the evenness property.

For example, let's $d(X, Y) < 1$, $d(X, Z) < 1$ and $d(Y, Z) > 1$. Then, the values of transformed d^∞ dissimilarities are

$$d^\infty(X, Y) = d^\infty(X, Z) = 0 \quad \text{and} \quad d^\infty(Y, Z) = \infty.$$

We see that, regardless of d, d^∞ is not even.

Joly and Le Calvé proved[5]

- $d \in \mathfrak{D}_\infty \Rightarrow d^\alpha \in \mathfrak{D}_\infty$ for all α: $0 \leq \alpha \leq 1$;
- $d \in \mathfrak{D}_+ \Rightarrow$ there is a unique nonnegative number $p \in \overline{\mathbb{R}}$, such that

$$d^\alpha \in \mathfrak{D}_\infty \quad \text{for all } \alpha: \quad \alpha \leq p, \text{ and}$$
$$d^\alpha \notin \mathfrak{D}_\infty \quad \text{for all } \alpha: \quad \alpha > p.$$

We call this threshold value

$$p = p(d) := \sup_\alpha \{ \text{for all } X, Y, Z \in \mathcal{E}: d^\alpha(X, Y) \leq d^\alpha(X, Z) + d^\alpha(Y, Z) \}$$

the metric index of dissimilarity d. We can extend the definition of a metric index to all nonnegative dissimilarity measures and therefore we decided to use this term instead of the distance index we used in previous papers.[6,7,8,9]

If a dissimilarity d is not even, then $p(d) = 0$.

For an ultrametric dissimilarity d, we have d^α ultrametric for all $\alpha \geq 0$ and therefore for an ultrametric d, we define $p(d) = \infty$.

Association Coefficients

In the case when all the properties measured on each unit are of the presence/absence type, the description of a unit is a binary vector with the i-th component equal to 1 if the unit has the i-th property, and equal to 0 if it lacks the i-th property. Therefore, if m is the number of measured properties, the description of a unit X is a binary vector $\mathbf{x} = [x_1, ..., x_m] \in \mathbb{B}^m$, $\mathbb{B} := \{0, 1\}$, where

$x_i = 1$, if unit X has the i-th property,

$x_i = 0$, if X lacks the i-th property, $i = 1, 2, ..., m$.

For $\mathbf{x}, \mathbf{y} \in \mathbb{B}^m$ we denote $\mathbf{x}\,\mathbf{y} := \sum_{i=1}^m x_i y_i$, $\overline{\mathbf{x}} := [1 - x_i]$ and we define the counters

$$a := \mathbf{x}\,\mathbf{y}, \quad b := \mathbf{x}\,\overline{\mathbf{y}}, \quad c := \overline{\mathbf{x}}\,\mathbf{y}, \quad d := \overline{\mathbf{x}}\,\overline{\mathbf{y}},$$

where $a + b + c + d = m$. Using the letter d to denote the counter and/or the dissimilarity measure might be confusing, but its use is always evident from the context. Several association coefficients are defined with these counters.[1,2,10] For example, Hubálek[1] in his article gives a list of 43 coefficients. From this and other lists, we have chosen[2] and compared 22 association coefficients and for our analysis here we have selected 19 nonequivalent association coefficients – see Table I. With adequate transformations – see column Diss. in Table I, we transformed them to the dissimilarity measures having the range $[0, 1]$ or $[0, \infty]$. Note that association coefficients Q_0 and – bc – are dissimilarities denoted d_{12} and d_{15} and therefore no transformation is needed.

Since the triangle inequality D5 implies evenness, we further consider only even D4 dissimilarity measures. Therefore, as defined at the end of previous section, the noneven dissimilarities Q_0, $\frac{1}{2}(1 - Q)$, – bc – and Simpson's $1 - s_{21}$ have a metric index equal to 0.

In cases of indeterminacy – expressions of the form $\frac{0}{0}$, we eliminate indeterminacies by appropriately defining values in critical cases. We use the definitions proposed in Ref. 2 – the second column of Table I contains the labels used there. This solution substantially simplifies our study and also permits to write robust computer programs for calculation of association coefficients.

For example, we defined Jaccard's coefficient by the expression:

$$s_6 := \begin{cases} 1 & \text{for } d = m, \\ \dfrac{a}{a+b+c} & \text{otherwise,} \end{cases}$$

thus ensuring $s_6(X, X) = 1$ also for the vector $\mathbf{0} := [0, 0 ... 0]$.

Table I. Association coefficients

Measure	Sim. s_i	Definition	Dissim. d_i	D2	D3	D4	D5
1. Russel and Rao (1940)	s_1	$\frac{a}{m}$	$1-s$	N	Y	Y	Y
2. Kendall, Sokal-Michener (1958)	s_2	$\frac{a+d}{m}$	$1-s$	Y	Y	Y	Y
3. Rogers & Tanimoto (1960)	s_3	$\frac{a+d}{a+d+2(b+c)}$	$1-s$	Y	Y	Y	Y
4. Jaccard (1900)	s_6	$\frac{a}{a+b+c}$	$1-s$	Y	Y	Y	Y
5. Kulczynski (1927), T^{-1}	s_7	$\frac{a}{b+c}$	s^{-1}	Y	Y	Y	N
6. Dice (1945), Czekanowski (1913)	s_8	$\frac{a}{a+\frac{1}{2}(b+c)}$	$1-s$	Y	Y	Y	N
7. Sokal & Sneath (1963), un_2	s_9	$\frac{a}{a+2(b+c)}$	$1-s$	Y	Y	Y	Y
8. Kulczynski	s_{10}	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	$1-s$	Y	Y	Y	N
9. Sokal & Sneath (1963), un_4	s_{11}	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$	$1-s$	Y	Y	Y	N
10. Q_0	d_{12}	$\frac{bc}{ad}$	d	Y	N	N	N
11. Yule (1927), Q	s_{14}	$\frac{ad-bc}{ad+bc}$	$\frac{1}{2}(1-s)$	N	N	N	N
12. – bc –	d_{15}	$\frac{4bc}{m^2}$	d	Y	N	N	N
13. Driver & Kroeber (1932), Ochiai (1957)	s_{16}	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$1-s$	Y	Y	Y	N
14. Sokal & Sneath (1963), un_5	s_{17}	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$1-s$	Y	Y	Y	N
15. Pearson, ϕ	s_{18}	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$\frac{1}{2}(1-s)$	Y	Y	Y	N
16. Baroni-Urbani, Buser (1976), S^{**}	s_{19}	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$1-s$	Y	Y	Y	N
17. Braun-Blanquet (1932)	s_{20}	$\frac{a}{\max(a+b,a+c)}$	$1-s$	Y	Y	Y	Y
18. Simpson (1943)	s_{21}	$\frac{a}{\min(a+b,a+c)}$	$1-s$	Y	N	N	N
19. Michael (1920)	s_{22}	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$\frac{1}{2}(1-s)$	N	Y	Y	N

DETERMINING THE METRIC INDEX

Properties of the Metric Index

In this section, we give an alternative proof of the Joly and Le Calvé theorem[5] extended to even nonnegative dissimilarity measures. This proof also suggests the method for determining the metric index explained in the next subsection (The Method).

Theorem 1. – For any even nonnegative dissimilarity measure d on $\mathcal{E}$ there is a unique nonnegative number p, its metric index, such that

d^α is metric for all $\alpha \leq p$, and

d^α is not metric for all $\alpha > p$.

Proof: For any three units $X, Y, Z \in \mathcal{E}$ we can always assume that

$$d(X, Y) \leq d(X, Z) \leq d(Y, Z) . \qquad (1)$$

We only consider cases such that $d(X, Y) > 0$, since by the theorem assumption d is even and therefore $d(X, Y) = 0$ implies $d(X,Z) = d(Y,Z)$. For such a triple, the triangle inequality holds as an equality

$$d(X, Y) + d(X, Z) = d(Y, Z) \quad \text{or}$$
$$d(X, Y) + d(Y, Z) = d(X, Z)$$

or as an inequality $d(X, Z) + d(Y, Z) = 2\, d(X, Z) \geq d(X, Y) = 0$, all remaining valid also for all d^α, $\alpha > 0$.

An evident consequence of the inequality

$$u^\alpha + v^\alpha \geq (u + v)^\alpha \quad \text{for all } \alpha, \quad 0 \leq \alpha \leq 1$$

is that for all triples $X, Y, Z \in \mathcal{E}$ that satisfy the triangle inequality, also

$$d^\alpha(X, Y) + d^\alpha(X, Z) \geq d^\alpha(Y,Z)$$

holds for all α, $0 \leq \alpha \leq 1$. Hence, if d is metric on $\mathcal{E}$, its metric index, if it exists, is more or equal to 1.

Now, if the triangle inequality holds for all d^α, $\alpha > 1$, d is an ultrametric – see also Résultat 2 in Joly and Le Calvé.[5] To prove this, let for all $\alpha > 1$

$$d^\alpha(X, Y) + d^\alpha (X, Z) \geq d^\alpha (Y, Z)\,. \qquad (2)$$

Since $d(Y, Z) > 0$, we can define

$$u_{XYZ} := \frac{d(X,Y)}{d(Y,Z)} \quad \text{and} \quad v_{XYZ} := \frac{d(X,Z)}{d(Y,Z)}\,.$$

We will omit the indexes, where there will be no possible confusion: from (1) we have that $u \leq v \leq 1$ and dividing (2) by $d^\alpha(Y, Z)$, we get

$$u^\alpha + v^\alpha \geq 1$$

for all $\alpha > 1$, which is possible only if at least v is equal to 1. Therefore, $d(X,Z) = d(Y,Z)$ and d is ultrametric. We pose $p(d) = \infty$.

When d is not an ultrametric, there is an $\alpha \in \mathbb{R}^+$ and a triple $X, Y, Z \in \mathcal{E}$, for which the dissimilarity measure d^α is not metric

$$d^\alpha(X,Y) + d^\alpha(X,Z) \leq d^\alpha(Y,Z).$$

If we denote $\delta := d^\alpha$, we have

$$\delta(X,Y) + \delta(X,Z) \leq \delta(Y,Z)$$

indicating that δ is not metric. Its metric index, if it exists, is less than 1 and the relation between metric indexes $p(d)$ and $p(\delta)$, if they exist, is $p(d) = \alpha\, p(\delta)$.

Therefore, without loss of generality, we can limit our further discussion to triples $X, Y, Z \in \mathcal{E}$ that do not obey the triangle inequality

$$d(X,Y) + d(X,Z) < d(Y,Z).$$

Dividing the inequality by $d(Y,Z)$, we get

$$0 < u + v < 1.$$

Consider the function f: $[0, 1] \to \mathbb{R}$ defined by

$$f(\alpha) := u^\alpha + v^\alpha - 1.$$

It is continuous, $f(0) = 1$, $f(1) = u + v - 1 < 0$ and for $0 < u, v < 1$

$$f'(\alpha) = u^\alpha \ln u + v^\alpha \ln v < 0.$$

We see that $f(\alpha)$ is strictly decreasing and of opposite signs at the endpoints of the interval $[0, 1]$. Hence, there is a unique value $\alpha_0 = \alpha_0(u, v) \in (0, 1)$ such that $f(\alpha_0) = 0$, that is

$$u^{\alpha_0} + v^{\alpha_0} = 1.$$

Therefore,

$$d^{\alpha_0}(X, Y) + d^{\alpha_0}(X, Z) = d^{\alpha_0}(Y, Z)$$

the metric inequality holds as an equality. For all α less than α_0, the function $f(\alpha)$ is positive

$$0 < f(\alpha) = u^\alpha + v^\alpha - 1$$

$$1 < u^\alpha + v^\alpha$$

$$d^\alpha(Y, Z) < d^\alpha(X, Y) + d^\alpha(X, Z)$$

the triangle inequality holds for the triple X, Y, Z. For all α more than α_0 the function $f(\alpha)$ is negative and the triangle inequality fails.

Now, if for all triples $X, Y, Z \in \mathcal{E}$ that fail the triangle inequality, we compute the values α_0, then there is unique

$$p := \inf \alpha_0 \geq 0$$

that we call the metric index of the dissimilarity d, $p = p(d)$. □

Corollary 2. – Under the conditions of Theorem 1, if the set of units $\mathcal{E}$ is finite, then $p(d) > 0$.

Proof: See the end of the preceding proof. Because there is a finite number of units, also the number of α_0s is finite and since $\alpha_0 > 0$, it follows that also p is positive, $p = \min \alpha_0 > 0$. □

Proposition 3. – Let d be metric on $\mathcal{E}$. If there is a triple $X, Y, Z \in \mathcal{E}$ such that $d(X, Z) > 0$, $d(Y, Z) > 0$ and

$$d(X, Y) = d(X, Z) + d(Y, Z)$$

then $p(d) = 1$.

Proof: Evidently $p(d) \geq 1$. Since for $\alpha > 1$ and u, v > 0, it follows that

$$(u + v)^\alpha > u^\alpha + v^\alpha$$

and we have

$$d^\alpha(X, Y) = (d(X, Z) + d(X, Z))^\alpha > d^\alpha(X, Z) + d^\alpha(Y, Z) .$$

The triangle inequality fails. Therefore, for $\alpha > 1$, d^α is not metric. □

When the set of units $\mathcal{E} = \mathbb{B}^m$ – the case of dissimilarity measures on binary vectors, in general, the metric index depends on their dimension. Therefore, for a nonnegative dissimilarity measure d, we denote its metric index on $\mathbb{B}^m$ by $p_m(d)$ or simply p_m and consequently the metric index $p(d)$ of d valid on all sets of units $\mathbb{B}^m$, $m \in \mathbb{N}$ is

$$p(d) = \inf\{p_m(d);\ m \in \mathbb{N}\}.$$

The Method

This method for determining the metric index of any even nonnegative dissimilarity measure on $\mathbb{B}^m$ is based on the steps of the proof of Theorem 1.

Let d be an even nonnegative dissimilarity measure on $\mathbb{B}^m$. Our task is to find $p_m(d)$ for an arbitrary $m \in \mathbb{N}$, that is the minimal exponent α such that

$$u^\alpha_{\mathbf{x}_0\mathbf{y}_0\mathbf{z}_0} + v^\alpha_{\mathbf{x}_0\mathbf{y}_0\mathbf{z}_0} = 1$$

holds for some triple $\mathbf{x}_0, \mathbf{y}_0, \mathbf{z}_0 \in \mathbb{B}^m$, and for all other triples $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{B}^m$

$$u^\alpha_{\mathbf{xyz}} + v^\alpha_{\mathbf{xyz}} > 1$$

holds.

On the region $\mathcal{R} \subset \mathbb{R}^3$ determined by inequalities

$$0 < u \leq v < 1, \quad u + v < 1 \quad \text{and} \quad 0 < \alpha < 1$$

we observe the function F: $\mathcal{R} \to \mathbb{R}$ defined by

$$F(u, v, \alpha) := u^\alpha + v^\alpha .$$

It is continuous and from its derivatives on $\mathcal{R}$ we see that F is increasing and concave in directions u and v and decreasing and convex in direction α. Therefore we can not apply methods of Linear or Mathematical programming where convexity is essential for the existence of the necessary and sufficient condition for global optima. Hence, we constructed a step by step method to find the $(\mathbf{x}_0, \mathbf{y}_0, \mathbf{z}_0)$ triple and corresponding exponent α.

For a fixed $m \in \mathbb{N}$, we denote

$$\mathcal{B} := \{(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathbb{B}^{3m};\ 0 < d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) < d(\mathbf{y}, \mathbf{z}) \wedge d(\mathbf{x}, \mathbf{y}) + d(\mathbf{x}, \mathbf{z}) < d(\mathbf{y}, \mathbf{z})\}$$

and determine $u_0 := \min\{u_{\mathbf{xyz}};\ (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B}\}$. Now we restrict ourselves to

$$\mathcal{B}_0 := \{(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B};\ u_{\mathbf{xyz}} = u_0\}$$

and determine $v_0 := \min\{v_{\mathbf{xyz}};\ (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B}_0\}$; finally, α_0 is determined by

$$u_0^{\alpha_0} + v_0^{\alpha_0} = 1 .$$

Proposition 4. – If $v_0 = u_0$ then $p_m = \alpha_0$ is the metric index of d on $\mathbb{B}^m$.

Proof: For any triple $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B}$ we have

$$u_0 = v_0 \leq u_{\mathbf{xyz}} \leq v_{\mathbf{xyz}} .$$

Since F is an increasing function of u as well as of v, this gives:

$$1 = F(u_0, u_0, \alpha_0) \leq F(u_{\mathbf{xyz}}, v_{\mathbf{xyz}}, \alpha_0) .$$

As we proved in the previous section, there is a unique value α such that

$$1 = u^\alpha_{\mathbf{xyz}} + v^\alpha_{\mathbf{xyz}} = F(u_{\mathbf{xyz}}, v_{\mathbf{xyz}}, \alpha) \leq F(u_{\mathbf{xyz}}, v_{\mathbf{xyz}}, \alpha_0) .$$

Since F is a decreasing function of α: $\alpha_0 \leq \alpha$ and $p_m = \alpha_0$ holds. □

Remark 5. – For $u_0 = v_0$ the exponent α_0 is obtained as the solution

$$\alpha_0 = -\frac{\log 2}{\log u_0}$$

of the equation $2\, u_0^{\alpha_0} = 1$.

And this is the end only if we are lucky that $u_0 = v_0$. If not?

Question: What if $v_0 > u_0$?

Let us denote $s := u + v$ and rewrite F in the form

$$F(u, v, \alpha) = u^\alpha + (s - u)^\alpha =: \Phi(u, s, \alpha) .$$

We study the function Φ on the region $\mathcal{W} \subset \mathbb{R}^3$ determined by inequalities

$$0 < u \leq \frac{s}{2}, \quad 0 < s < 1 \quad \text{and} \quad 0 < \alpha < 1.$$

Φ is continuous on $\mathcal{W}$ and from its derivatives on $\mathcal{W}$ we see that Φ is increasing in directions u and s and decreasing in the α direction. We denote $s_0 := u_0 + v_0$ and

compute $u = \min\{u_{\mathbf{xyz}};\ (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B} \setminus \mathcal{B}_0\}$. Again we restrict to

$$\mathcal{B}_1 := \{(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B};\ u_{\mathbf{xyz}} = u\}$$

and compute $v = \min\{v_{\mathbf{xyz}};\ (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B}_1\}$.

Proposition 6. – If $s \geq s_0$, then $\alpha > \alpha_0$ holds.

Proof: Because $u_0 < u$, $s_0 \leq s = u + v$ and Φ is an increasing function in u and s, we have

$$1 = u_0^{\alpha_0} + v_0^{\alpha_0} = \Phi(u_0, s_0, \alpha_0) < \Phi(u, s, \alpha_0)\,.$$

As we proved in the previous section, there is a unique value α such that

$$1 = u^{\alpha} + v^{\alpha} = \Phi(u, s, \alpha) < \Phi(u, s, \alpha_0)\,.$$

Since Φ is a decreasing function in α, $\alpha_0 < \alpha$ holds.

Remark 7. – If $s \geq s_0$ for all $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B} \setminus \mathcal{B}_0$, then $p_m = \alpha_0$ holds.
And again, this is the end only if we are lucky that $s \geq s_0$. If not?

Question: What if $s < s_0$ for some $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{B} \setminus \mathcal{B}_0$?

Proposition 8. – If $u \geq 2^{-\frac{1}{\alpha_0}}$, then $p_m = \alpha_0$.

Proof: As the values u_0, v_0 $(u_0 < v_0)$, α_0 and u are known, we determine

$$v = \min\{v_{\mathbf{xyz}};\ u_{\mathbf{xyz}} = u\}.$$

Since $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z})$, the lowest value we can get for v is $v = u$ and in this case the value of α is also minimal:

$$\min \alpha = -\frac{\log 2}{\log u}\,.$$

Therefore, we can get $\alpha < \alpha_0$ only if $\min \alpha < \alpha_0$, that is, if

$$u < 2^{-\frac{1}{\alpha_0}} =: u_b\,.$$

□

Now we can summarize the complete procedure of determining p_m and p:

STEP 1: Compute u_0 and v_0. If $v_0 = u_0$, then $p_m = -\dfrac{\log 2}{\log u_0}$;

STEP 2: If $v_0 > u_0$, then compute also α_0 and u. If $u \geq u_b$, then $p_m = \alpha_0$;

STEP 3: If $u < u_b$, then determine $\mathcal{B}_1$, v and s.

If $s \geq s_0$, set $\mathcal{B}_0 = \mathcal{B}_1 \cup \mathcal{B}_0$ and continue with step 2. Otherwise for

CASE A: $F(u, v, \alpha_0) = u^{\alpha_0} + v^{\alpha_0} \geq 1$: set $\mathcal{B}_0 = \mathcal{B}_1 \cup \mathcal{B}_0$ and continue with step 2.

CASE B: $F(u, v, \alpha_0) = u^{\alpha_0} + v^{\alpha_0} < 1$: set $u_0 = u$, $v_0 = v$, $\mathcal{B}_0 = \mathcal{B}_1 \cup \mathcal{B}_0$ and go back to step 1.

Since the sets $\mathcal{B}$, $\mathcal{B}_0$ and $\mathcal{B}_1$ are finite, we get in a finite number of steps the global minimum of all exponents and so we determine the value of p_m. Finally,

$$p = \inf\{p_m;\ m \in \mathbb{N}\}.$$

It may happen that p_m does not depend on m.

The Method Written in Integer Coordinates

We can rewrite the dissimilarity measures among binary vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{B}^m$ in the form that involves counters that count frequencies of all possible triples component wise. There are eight possible triples from 111 to 000 and we denote their frequencies by $A,B,...,H$ – see Table II, where

$$A + B + C + D + E + F + G + H = m.$$

Table II. Frequencies of all possible triples

x	1	1	1	1	0	0	0	0
y	1	1	0	0	1	1	0	0
z	1	0	1	0	1	0	1	0
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*

Since

$$\begin{aligned} a_{\mathbf{xy}} &= A + B & a_{\mathbf{xz}} &= A + C & a_{\mathbf{yz}} &= A + E \\ b_{\mathbf{xy}} &= C + D & b_{\mathbf{xz}} &= B + D & b_{\mathbf{yz}} &= B + F \\ c_{\mathbf{xy}} &= E + F & c_{\mathbf{xz}} &= E + G & c_{\mathbf{yz}} &= C + G \\ d_{\mathbf{xy}} &= G + H & d_{\mathbf{xz}} &= F + H & d_{\mathbf{yz}} &= D + H \end{aligned}$$

we can express

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &=: \delta_{\mathbf{xy}}(A,B,C,D,E,F,G,H), \\ d(\mathbf{x}, \mathbf{z}) &=: \delta_{\mathbf{xz}}(A,B,C,D,E,F,G,H), \\ d(\mathbf{y}, \mathbf{z}) &=: \delta_{\mathbf{yz}}(A,B,C,D,E,F,G,H) \end{aligned}$$

where $\delta_{\mathbf{xy}}$, $\delta_{\mathbf{xz}}$ and $\delta_{\mathbf{yz}}$ map the lattice

$$\mathcal{S}_m := \{(A,B,C,D,E,F,G,H) \in \mathbb{N}^8 :\ A + B + C + D + E + F + G + H = m\}$$

into the reals. Let

$$\mathcal{S} := \{(A,B,C,D,E,F,G,H) \in \mathcal{S}_m :\ 0 < \delta_{\mathbf{xy}} \leq \delta_{\mathbf{xz}} < \delta_{\mathbf{yz}} \wedge \delta_{\mathbf{xy}} + \delta_{\mathbf{xz}} < \delta_{\mathbf{yz}}\},$$

Table III. The results of metric index computation

	Dissimilarity measure	Dimension m							
		2	3	4	5	20	50	100	p
1.	$1 - s_1$								1
2.	$1 - s_2$								1
3.	$1 - s_4$								$\frac{\log 2}{\log \frac{3}{2}} = 1.709$
4.	$1 - s_6$								1
5.	s_7^{-1}								0
6.	$1 - s_8$								$\frac{\log 2}{\log 3} = 0.630$
7.	$1 - s_9$								$\frac{\log 2}{\log \frac{3}{2}} = 1.709$
8.	$1 - s_{10}$	0.5	0.489	0.474	0.462	0.379	0.335	0.308	0
9.	$1 - s_{11}$	0.706	0.630	0.595	0.575	0.500	0.436	0.395	0
10.	d_{12}								0
11.	$\frac{1}{2}(1 - s_{14})$								0
12.	d_{15}								0
13.	$1 - s_{16}$	0.564	0.563	0.562	0.561	0.553	0.548	0.544	0.5
14.	$1 - s_{17}$		1	0.804	0.731	0.594	0.575	0.570	0.564
15.	$\frac{1}{2}(1 - s_{18})$	1	0.630	0.603	0.592	0.569	0.566	0.565	0.564
16.	$1 - s_{19}$	1	0.630	0.564	0.526	0.378	0.316	0.279	0
17.	$1 - s_{20}$								1
18.	$1 - s_{21}$								0
19.	$\frac{1}{2}(1 - s_{22})$								0.279

$$u := \frac{\delta_{\mathbf{xy}}}{\delta_{\mathbf{yz}}} \quad \text{and} \quad v := \frac{\delta_{\mathbf{xz}}}{\delta_{\mathbf{yz}}}.$$

Of course, $0 < u + v < 1$ on $\mathcal{S}$. Now, we determine

$$u_0 := \min\{u : (A,B,C,D,E,F,G,H) \in \mathcal{S}\},$$
$$\mathcal{S}_0 := \{(A,B,C,D,E,F,G,H) \in \mathcal{S} : u = u_0\} \quad \text{and}$$
$$v_0 := \min\{v : (A,B,C,D,E,F,G,H) \in \mathcal{S}_0\}$$

and we follow steps 1 to 3 from the previous section just replacing $\mathcal{B}$, $\mathcal{B}_0$ and $\mathcal{B}_1$ by $\mathcal{S}$, $\mathcal{S}_0$ and $\mathcal{S}_1$, respectively.

Local Optimization Procedure

We can approach the problem of determining the metric index also by solving numerically the corresponding optimization problem

$$p = \operatorname{argmax}\{\alpha : \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{B}^m : d^\alpha(\mathbf{x}, \mathbf{y}) + d^\alpha(\mathbf{y}, \mathbf{z}) \geq d^\alpha(\mathbf{x}, \mathbf{z})\}.$$

We use a local optimization procedure

```
initial (read, random) x, y, z;
p := 1;
while ∃(u,v,w) ∈ N(x,y,z) : d^p(u,v) + d^p(v,w) < d^p(u,w)
do begin
    p := Solve(α : d^α(u,v) + d^α(v,w) = d^α(u,w) );
    x := u; y := v; z := w
end
```

over a neighbourhood of the current triple

$N(x, y, z) =$
$\{(x', y', z') : r \in 1..m \wedge (x'_r = \neg x_r \vee y'_r = \neg y_r \vee z'_r = \neg z_r)\}.$

This means that we get a triple of binary vectors (**x'**, **y'**, **z'**) from the neighbourhood if, at the selected position r, we simply change at least one digit in at least one of binary vectors **x**, **y** or **z**.

From the local minima (**x***, **y***, **z***) obtained by this procedure, we can usually guess a general pattern of 'extremal' triples, from which we compute an upper bound for p_m:

$$\overline{p}_m := p(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*), \quad \mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^* \in \mathbb{B}^m .$$

Example 9. – For the Sokal & Sneath dissimilarity $d_{11} = 1 - un_4$, we obtain for $m = 20$ the local optimum solution

$\mathbf{x}^* = [00000000000011111111]$,
$\mathbf{y}^* = [00000000000011111110]$,
$\mathbf{z}^* = [00000000000000000001]$

with the corresponding dissimilarities

$$d_{11}(\mathbf{x}^*, \mathbf{y}^*) = 0.050480769,$$
$$d_{11}(\mathbf{x}^*, \mathbf{z}^*) = 0.310855263,$$
$$d_{11}(\mathbf{y}^*, \mathbf{z}^*) = 0.611336032.$$

This gives the upper bound for the metric index $\overline{p}_{20} = 0.500366330$.

In terms of integer coordinates: the upper bound $\overline{p}_{20} = 0.500366330$ of the metric index p_{20} is attained when $B = 7$, $C = 1$ and $H = 12$. Solution obtained with the method confirms this results as the exact value of p_{20} – see Table III.

RESULTS

We summarize all our results on metric index computation in Table III. The values of metric indexes given in this table are exact to the third decimal place and according to the statement of Joly and Le Calvé in subsections *Dissimilarities*, if we take fewer decimals the values have to be exact too. We calculated these values of metric indices using the method explained in sections *The Method*, steps 1 to 3, and *The Method Written in Integer Coordinates*. If only the p value is given, then p_m does not depend on dimension m.

For dissimilarities: Russel and Rao, two families S_θ – contain Kendall, Rogers & Tanimoto and T_θ – contains Jaccard, Dice and Sokal & Sneath un_2, calculations of the metric index are in Appendix, other complete calculations with all steps and proofs are available as a preprint in the Preprint Series of the Institute of Mathematics, Physics and Mechanics, University of Ljubljana. All steps and proofs are also available at the www address http://vlado.fmf.uni-lj.si/vlado/vladounp.htm.

CONCLUSIONS

In the paper we discussed power transformation as a method that transforms nonmetric dissimilarities into metric distances, or nonmetric dissimilarity measures into metric dissimilarity measures. We presented a method for determining the metric index of a given dissimilarity between binary vectors and we applied it to some well known dissimilarity coefficients. The results obtained offer new information that can be used when selecting dissimilarity coefficients for applications.

We stress also that for Ochiai $1 - s_{16}$, Sokal & Sneath $1 - s_{17}$, Pearson $(1 - s_{18})/2$ and Dice $1 - s_8$ dissimilarities, we confirmed the well known results, that their square roots are distances.

We expect that the proposed method can be successfully applied to dissimilarities between other types of units.

Acknowledgments. – This work was supported in part by the Ministry of Science of Slovenia. We thank dr. Aleš Založnik for numerous remarks and suggestions that significantly improved the presentation of the material.

APPENDIX

The results presented in the following sections were suggested by the local optimization procedure and verified with the method described in the section The Method.

Russel and Rao

If we write Russel and Rao dissimilarity measure

$$d_1(\mathbf{x}, \mathbf{y}) := 1 - s_1(\mathbf{x}, \mathbf{y}) = 1 - \frac{a}{m} = \frac{b+c+d}{m}$$

in integer coordinates for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{B}^m$, we get

$\delta_{\mathbf{xy}}(A,B,...,H) = (C + D + E + F + G + H)/m,$
$\delta_{\mathbf{xz}}(A,B,...,H) = (B + D + E + F + G + H)/m$ and
$\delta_{\mathbf{yz}}(A,B,...,H) = (B + C + D + F + G + H)/m.$

If we assume $0 < \delta_{\mathbf{xy}} \le \delta_{\mathbf{xz}} \le \delta_{\mathbf{yz}}$ and write $S := D + F + G + H$, we have

$$u_{\mathbf{xyz}} = \delta_{\mathbf{xy}} / \delta_{\mathbf{yz}} = \frac{C+E+S}{B+C+S},$$

$$v_{\mathbf{xyz}} = \delta_{\mathbf{xz}} / \delta_{\mathbf{yz}} = \frac{B+E+S}{B+C+S}.$$

The minimal value $u_0 = 1/m$ is attained when $C + D + E + F + G + H = 1$ and $B + C + D + F + G + H = m$, that is, for $A = E = 0$, $B = m - 1$ and $C = 1$ or $S = 1$. We get v_0 when $B = m - 1$ and $C = 1$: $v_0 = (m - 1) / m$. In this case

Table IV. Some well-known similarities

	Definition	θ	p
Kendall, Sokal-Michener (1958)	$\frac{a+d}{m}$	S_1	1
Rogers & Tanimoto (1960)	$\frac{a+d}{a+d+2(b+c)}$	S_2	$\frac{\log 2}{\log \frac{3}{2}} = 1.709$
Jaccard (1900)	$\frac{a}{a+b+c}$	T_1	1
Dice (1945), Czekanowski (1913)	$\frac{a}{a+\frac{1}{2}(b+c)}$	$T_{1/2}$	$\frac{\log 2}{\log 3} = 0.630$
Sokal & Sneath	$\frac{a}{a+2(b+c)}$	T_2	$\frac{\log 2}{\log \frac{3}{2}} = 1.709$

$$s_0 = u_0 + v_0 = \frac{1}{m} + \frac{m-1}{m} = 1$$

and so $\alpha_0 = 1$. Since

$$s = u_{\mathbf{xyz}} + v_{\mathrm{xyz}} = \frac{B+C+2(D+E+F+G+H)}{B+C+D+F+G+H} \geq 1 ,$$

$p_m = \alpha_0 = 1$ and therefore $p = 1$ – see Proposition 3. This is in agreement with the fact that the triangle inequality holds for d_1 as written in Table I.

Families S_θ *and* T_θ

Gower and Legendre[10] introduced two families of functions

$$S_\theta = \frac{a+d}{a+d+\theta(b+c)} \quad \text{and} \quad T_\theta = \frac{a}{a+\theta(b+c)}$$

(where $\theta > 0$ to avoid negative values) that contain some well-known similarity measures – see Table IV.

Investigating metric and Euclidean properties of the dissimilarities $1 - S_\theta$, $1 - T_\theta$ and $(1 - S_\theta)^{1/2}$, $(1 - T_\theta)^{1/2}$, they found out that these properties depend on θ: there is a critical value θ_M such that for values of θ near 0 these dissimilarities are not metric – but for $\theta \geq \theta_M$ they become metric. These results are summarized in Gower's and Legendre's[10] theorems 9 and 10.

THEOREM 9. $1 - S_\theta$ is metric for $\theta \geq 1$ and $(1 - S_\theta)^{1/2}$ is metric for $\theta \geq 1/3$. If $\theta < 1$, then $1 - S_\theta$ may not be metric and if $\theta < 1/3$, $(1 - S_\theta)^{1/2}$ may be non-metric.

THEOREM 10. Says the same for $1 - T_\theta$ and $(1 - T_\theta)^{1/2}$.

If we write t := $1/\theta$, we see that for $t > 1$ the dissimilarity

$$d_t = 1 - T_{1/t} = \frac{b+c}{t\,a+b+c}$$

may not be metric. Therefore, we can ask:

Find the metric index p for the dissimilarity d_t.

First, we shall compute the metric index p_m for an arbitrary $m \in \mathbb{N}$. We shall see that it is independent of m and thus answers our question.

In the integer coordinates we express

$$d_t(\mathbf{x}, \mathbf{y}) = \delta_{\mathbf{xy}}(A,B,...,H;\, t) = \frac{C+D+E+F}{t(A+B)+C+D+E+F} ,$$

$$d_t(\mathbf{x}, \mathbf{z}) = \delta_{\mathbf{xz}}(A,B,...,H;\, t) = \frac{B+D+E+G}{t(A+C)+B+D+E+G} ,$$

$$d_t(\mathbf{y}, \mathbf{z}) = \delta_{\mathbf{yz}}(A,B,...,H;\, t) = \frac{B+C+F+G}{t(A+E)+B+C+F+G} .$$

Investigating all possibilities to get the smallest positive value of $u_{\mathbf{xyz}} = \delta_{\mathbf{xy}} / \delta_{\mathbf{yz}}$ gives us a possible solution $B = m - 1$ and $C = 1$. To increase u, we pose more generally $B + C + H = m$, $B,C > 0$ – now $B = m - 1$, $C = 1$ is just a special case and calculate the sum s

$$s = u + v = \frac{C}{tB+C} + \frac{B}{tC+B} =$$

$$\frac{t(B^2+C^2)+2BC}{t^2BC+t(B^2+C^2)+BC}$$

Denoting $\xi := B / C$, we re-express it in the form

$$s(t, \xi) = \frac{t\xi^2+2\xi+t}{t\xi^2+\xi(t^2+1)+t} .$$

From the first derivatives

$$s'_t = -\xi\frac{(t+\xi)^2+(t\xi+1)^2}{(t+\xi)^2\,(t\xi+1)^2} \quad \text{and} \quad s'_\xi = \frac{t(t^2-1)(\xi^2-1)}{(t+\xi)^2(1+t\xi)^2}$$

we see that s is decreasing with t and that $s'_\xi = 0$ for $t = 0, \pm 1$ or $\xi = \pm 1$. Since $t > 1$ and $1 / (m-1) \leq \xi \leq m-1$, we get, when t is fixed, the minimal sum for $\xi = 1$, that is for $B = C$. In this case

$$u = v = \frac{1}{t+1} \quad \text{and} \quad \alpha = \frac{\log 2}{\log(t+1)} .$$

Proposition 10. – If $B > 0$, $C > 0$ and $\alpha = \frac{\log 2}{\log(t+1)}$, $t > 1$, then

$$u^\alpha + v^\alpha = \left(\frac{C}{tB+C}\right)^\alpha + \left(\frac{B}{tC+B}\right)^\alpha \geq 1 .$$

Proof: For $\xi = B / C > 0$, we define

$$\phi(t, \xi) := u^{\alpha} + v^{\alpha} = \left(\frac{1}{t\xi+1}\right)^{\frac{\log 2}{\log(t+1)}} + \left(\frac{1}{\frac{t}{\xi}+1}\right)^{\frac{\log 2}{\log(t+1)}}$$

The function $\phi(t, \xi)$ has the following properties

- $\phi(t, \frac{1}{\xi}) = \phi(t, \xi)$ and so we can consider the function ϕ only on the rectangle $\mathscr{R} := \{(t, \xi); t > 1 \wedge 0 \le \xi \le 1\}$;
- $\phi(t, 0) = \phi(t, 1) = 1$;
- $\phi(1, \xi) = 1$ and $\lim_{t\to\infty} \phi(t, \xi) = 1$.

Searching for the minimal value of the function we compute the first derivative

$$\phi'_{\xi}(t, \xi) = \alpha\left(\frac{1}{t\xi+1}\right)^{\alpha-1} \frac{-t}{(t\xi+1)^2} + \alpha\left(\frac{\xi}{t+\xi}\right)^{\alpha-1} \frac{t}{(t+\xi)^2}$$

and for $\xi = 1$, we have

$$\phi'_{\xi}(t, 1) = -\alpha t\left(\frac{1}{1+t}\right)^{\alpha+1} + \alpha t\left(\frac{1}{1+t}\right)^{\alpha+1} = 0$$

From the second derivative

$$\phi''_{\xi\xi}(t, \xi) = -\frac{((1-\alpha)t+2\xi)\alpha t\left(\frac{\xi}{t+\xi}\right)^{\alpha}}{\xi^2(t+\xi)^2} + \frac{(1+\alpha)\alpha t^2\left(\frac{1}{t\xi+1}\right)^{\alpha}}{(t\xi+1)^2}$$

we get

$$\phi''_{\xi\xi}(t, 1) = \frac{2\alpha t(\alpha t-1)}{(1+t)^{\alpha}(1+t)^2}$$

and $\phi''_{\xi\xi}(t, 1) > 0$ if and only if $\alpha t - 1 > 0$, that is, if and only if

$$t > \frac{\log(t+1)}{\log 2} = \log_2(t+1),$$

which is true for $t > 1$. Since $\phi(t, 1) = 1$ for all $t > 1$, we see that for $\xi = 1$ the function ϕ has a local minimum equal to 1.

Now for fixed t, $t > 1$, we show that there is only one point $\xi_t \in (0, 1)$ such that $\phi'_{\xi}(t, \xi_t) = 0$. Therefore, ξ_t is a local maximum of the curve and $\phi(t, \xi) > 1$ for $t > 1$ and $0 < \xi < 1$.

We fix $t > 1$. For ξ, $0 < \xi < 1$, we rewrite the equation $\phi'_{\xi}(t, \xi) = 0$:

$$\xi^2 = \left(\frac{t\xi^2+\xi}{t+\xi}\right)^{\alpha+1}$$

$$\log \xi^2 = (1+\alpha)\log\frac{t\xi^2+\xi}{t+\xi}$$

and we denote the left and the right side of the last equation by $f(\xi)$ and $g_t(\xi)$, respectively. We have $f(1) = g_t(1) = 0$ and

$$\lim_{\xi\to+0} f(\xi) = \lim_{\xi\to+0} g_t(\xi) = -\infty.$$

From derivatives

$$f'(\xi) = \frac{2}{\xi},$$

$$g'_t(\xi) = (1+\alpha)\frac{t(1+2t\xi+\xi^2)}{\xi(t+\xi)(1+t\xi)},$$

$$f''(\xi) = -\frac{2}{\xi^2},$$

$$g''_t(\xi) = (1+\alpha)\frac{-t(t+2\xi+2t^2\xi+4t\xi^2+2t^3\xi^2+4t^2\xi^3+t\xi^4)}{\xi^2(t+\xi)^2(1+t\xi)^2}$$

we see that both functions are increasing and concave on the interval (0, 1). Since for all $t > 1$

$$g'_t(1) = (1+\alpha)\frac{2t}{t+1} > 2 = f'(1)$$

the curve g_t is steeper than f when $\xi = 1$. But comparing the power series

$$f'(\xi) = \frac{2}{\xi} \quad \text{and}$$

$$g'_t(\xi) = (1+\alpha)\left(\frac{1}{\xi} + \frac{t^2-1}{t} + \frac{1-t^4}{t^2}\xi + O(\xi^2)\right) \xrightarrow{\xi\to 0} \frac{1+\alpha}{\xi}$$

taking into account that: $2 > 1 + \alpha$, we see that for $\xi \to +0$ the curve f becomes steeper and therefore they intersect at exactly one point $(\xi_t, f(\xi_t)) = (\xi_t, g_t(\xi_t))$, $\xi_t \in (0, 1)$. □

An immediate consequence of this proposition is that $p_m = \alpha$. Since p_m does not depend on m, the metric index p of the dissimilarity d_t is also equal to

$$p = p_m = \frac{\log 2}{\log(t+1)}.$$

Now we will compute the metric index for the second family of dissimilarities

$$1 - S_{\theta} = \frac{b+c}{t(a+d)+b+c} =: d_t,$$

where $t := 1/\theta$. If we express d_t in integer coordinates, we get expressions similar to those for the family $1 - T_{\theta}$. Also, the results are similar: for B + C + G = m, we have

$$u = \frac{C}{t(B+G)+C} \quad \text{and} \quad v = \frac{B+G}{tC+B+G}$$

and for $\xi := (B + G) / C$, we get again

$$s(t, \xi) = u + v = \frac{t\xi^2 + 2\xi + t}{t\xi^2 + \xi(t^2+1) + t}.$$

According to Preposition 10, we get the minimal exponent α for $\zeta = 1$, that is, when $B + G = C$. This implies $m = 2C$; m is even. If we write $m = 2n$, $n \in \mathbb{N}$, we get in this case

$$u = v = \frac{1}{t+1} \quad \text{and} \quad \alpha = \frac{\log 2}{\log(t+1)} = p_{2n}.$$

If m is odd, $m = 2n + 1$, $n \in \mathbb{N}$, ξ is never equal to 1, therefore $u^\alpha + v^\alpha > 1$ and hence

$$p_{2n+1} > \frac{\log 2}{\log(t+1)} \quad \text{for all } n \in \mathbb{N}.$$

But for larger dimensions $\xi = n/(n + 1)$ tends to 1 and

$$p_{2n+1} \xrightarrow{n\to\infty} p_{2n} = \frac{\log 2}{\log(t+1)}.$$

This is the value of the metric index p also for the second family.

REFERENCES

1. Z. Hubálek, *Biol. Rev.* **57** (1982) 669–689.
2. V. Batagelj and M. Bren, *Journal of Classification* **12** (1995) 73–90.
3. D. Verbanac, D. Jelić, V. Stepanić, I. Tatić, D. Žiher, and S. Koštrun, *Croat. Chem. Acta* **78** (2005) 133–139.
4. T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, Chapman and Hall, London, 1994, p. 235.
5. S. Joly and G. Le Calvé, *Similarity functions*, in: B. Van Cutsem (Ed.), *Classification and Dissimilarity Analysis*, Lecture Notes in Statistics, Springer-Verlag, New York, 1994, pp. 67–87.
6. V. Batagelj and M. Bren, *Determining the Distance Index*, in: E. Diday, Y. Lechevallier, and O. Opitz (Eds.), *Ordinal and Symbolic Data Analysis*, Proc. of the International Conference – OSDA'95 Springer-Verlag, Berlin, 1996, pp. 238–251.
7. V. Batagelj and M. Bren, *Determining the Distance Index II*, in: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, and Y. Baba (Ed.), *Data Science, Classification and Related Methods*, Proc. of the International Federation of Classification Societies – IFCS96, Springer-Verlag, Tokyo, 1998, pp. 460–467.
8. M. Bren and V. Batagelj, *Metric index – overview and examples*, in: *Data Science, Classification and Related Methods*, Proc. of the International Federation of Classification Societies – IFCS98, 1998, pp. 36–39.
9. M. Bren and V. Batagelj, *The distance index*, in: A. Graovac (Ed.), Book of Abstracts, The 11th Dubrovnik International Course and Conference on the Interfaces among Mathematics, Chemistry and Computer Sciences, Dubrovnik, Croatia, 1996, p. 43.
10. J. C. Gower and P. Legendre, *Journal of Classification* **3** (1986) 5–48.

SAŽETAK

Metrički indeks nesličnosti

Matevž Bren i Vladimir Batagelj

U radu je diskutirana transformacija koja proizvoljnu nesličnost prevodi u poluudaljenost a definitnu nesličnost u udaljenost. Uveden je postupak za računanje metričkog indeksa primjenom kojeg je zatim određena njegova vrijednost za 19 standardnih mjera nesličnosti na dihotomnim podacima.