University of Ljubljana
Institute of Mathematics, Physics and Mechanics
Department of Mathematics
Jadranska 19, 1 000 Ljubljana, Slovenia

Preprint series, Vol. 35 (1997), 561

The Metric Index

Matevž Bren, Vladimir Batagelj

ISSN 1318-4865

May 1997

Math.Subj.Class.(1991): 54 E, 62 H 30.

Presented at the OSDA-95 International Conference on Ordinal and Symbolic Data Analysis, June 20-23, 1995, Paris, France and the IFCS-96 fifth conference of International Federation of Classification Societies *Data Science, Classification and Related Methods*, March 27-30, 1996, Kobe, Japan.

This work was **supported** by the Ministry of Science and Technology of Slovenia, Project J2–6191–101–94.

Shortened version of this paper **submitted to** the *Journal of Classification.*

Ljubljana, May 1997

The Metric Index*

Matevž Bren
University of Maribor, Faculty of Organizational Sciences
Prešernova 11, 4 000 Kranj, Slovenia
matevz.bren@fov.uni-mb.si

Vladimir Batagelj
University of Ljubljana, Faculty of Mathematics and Physics
Jadranska 19, 1 000 Ljubljana, Slovenia
vladimir.batagelj@uni-lj.si

Abstract

We discuss the power transformation that provides an arbitrary even dissimilarity to be a semi-distance or definite dissimilarity to be a distance. We present the method for the metric index computation and concrete results on determining the metric index for 16 standard dissimilarity measures on binary vectors.

Keywords: dissimilarity spaces, metric spaces, association coefficients, metric index.

AMS Subj. Class. (1991): 54 E, 62 H 30.

1 Introduction

In Batagelj, Bren (1996) and Batagelj, Bren (1997) we presented

- an analytical method of determining the metric index for two families of dissimilarities that contain some well-known similarity measures (Kendall, Sokal-Michener; Rogers and Tanimoto; Jaccard; Dice, Czekanowski; Sokal and Sneath) on binary vectors;

*We would like to thank dr. Aleš Založnik for numerous remarks and suggestions that significantly improved the presentation of the material.

- a computational approach to determine (the upper bound for) the metric index with a local optimization procedure, and
- an evaluation of (the upper bounds of) metric indexes for a list of 16 standard dissimilarities on binary vectors and for some usual nonlinear transformations of these dissimilarities.

In this paper we give a method for determining the metric index for any even nonnegative dissimilarity measure on binary vectors and we summarize all these results.

2 Preliminaries

2.1 Dissimilarities

A mapping $d: \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ is a *dissimilarity measure* on the set of *units* $\mathcal{E}$ iff it is

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

P2. straight: $d(X,X) \leq d(X,Y)$ 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 with $\mathcal{D}_+$ the set of all dissimilarities on $\mathcal{E}$.

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

D3. *definite* iff $d(X,Y) = 0 \Rightarrow X = Y$;

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

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

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

A dissimilarity d on $\mathcal{E}$ is said to be a *semi-distance* iff it is metric. Naturally, a dissimilarity space $(\mathcal{E}, d)$ is a semi-metric space iff d is a semi-distance. $(\mathcal{E}, d)$ is a metric space iff d is also definite; then, d is a *distance* (a *metric*). We denote the subset of $\mathcal{D}_+$ comprising all semi-distances by $\mathcal{D}_\infty$ – the notation established in Van Cutsem (1994).

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

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

where d is a nonnegative dissimilarity measure (on $\mathcal{E}$) and for $\alpha = 0$ we define d^0 as a *discrete distance*

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

The power transformation preserves properties **D1** – **D4** and **D6**.

Joly and Le Calvé (1986) proved

- $d \in \mathcal{D}_\infty \;\Rightarrow\; d^\alpha \in \mathcal{D}_\infty$ *for all* α: $\quad 0 \leq \alpha \leq 1$;
- $d \in \mathcal{D}_+ \;\Rightarrow$ *there exists a unique nonnegative number* $p \in \overline{\mathbb{R}}$, *such that*
  $d^\alpha \in \mathcal{D}_\infty$ *for all* α: $\alpha \leq p$ *and*
  $d^\alpha \notin \mathcal{D}_\infty$ *for all* α: $\alpha > p$.

We call this threshold value

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

the *metric index* of the dissimilarity d. We can extend the definition of a metric index to all nonnegative dissimilarity measures and therefore we decided to use this terminology instead of the *distance index* we used in previous papers. If a dissimilarity d is not even then $p(d) = 0$. For an ultrametric dissimilarity d we have $p(d) = \infty$.

2.2 Association coefficients

In the case when all m properties measured on each unit are of presence/absence type, a description of a unit X is a binary vector $x = [x_1, x_2, \ldots, x_m] \in \mathbb{B}^m$, $\mathbb{B} := \{0,1\}$, where

$x_i = 1,\quad$ if unit X has the i-th property,
$x_i = 0,\quad$ if X lacks the i-th property, $\quad i = 1, 2, \ldots, m$.

For $x, y \in \mathbb{B}^m$ we denote

$x\,y := \sum_{i=1}^m x_i\,y_i\quad$ – the scalar product and

$\overline{x} := [1 - x_i]\quad$ – the complementary vector of x

and we define counters:

$a := x\,y$ – numbers of properties which units X and Y share,
$b := x\,\overline{y}$ – numbers of properties which X has and Y lacks,
$c := \overline{x}\,y$ – numbers of properties which Y has and X lacks,
$d := \overline{x}\,\overline{y}$ – numbers of properties which both X and Y lack

where $a + b + c + d = m$. [1] With these counters several association coefficients are defined (Hubalek (1982), Gower and Legendre (1986), Baulieu (1989), Lerman (1971))– see Table 1.

Since triangle inequality (**D5**) implies evenness we consider only even (**D4**) dissimilarity measures and for our analysis we select 16 such measures (see d_i in Table 1).

2.2.1 Cases of indeterminacy

In cases of indeterminacy (expressions of the form $\frac{0}{0}$) we use definitions proposed in Batagelj and Bren (1995). We eliminate the indeterminacies by appropriately defining values in critical cases. This solution substantially simplifies our study and also permits writing robust computer programs for calculation of association coefficients.

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

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

thus ensuring $s_6(X, X) = 1$.

To make this reading easier, the second column of Table 1 includes also the labels used there.

3 Determining the metric index

3.1 The properties of metric index

In this section we give another proof of Joly and Le Calvé theorem (extended to even nonnegative dissimilarity measures) which suggests also a method for determining the metric index.

[1] Using the letter d to denote the counter and/or the dissimilarity measure might be confusing, but its use is evident from the context.

	measure	s_i	definition	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.	Jaccard (1900)	s_6	$\frac{a}{a+b+c}$	$1-s$	Y	Y	Y	Y
4.	Kulczynski (1927), T^{-1}	s_7	$\frac{a}{b+c}$	s^{-1}	Y	Y	Y	N
5.	Kulczynski	s_{10}	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	$1-s$	Y	Y	Y	N
6.	Sokal & Sneath (1963), un_4	s_{11}	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$	$1-s$	Y	Y	Y	N
7.	Q_0	s_{12}	$\frac{bc}{ad}$	s	N	N	N	N
8.	Yule (1927), Q	s_{14}	$\frac{ad-bc}{ad+bc}$	$\frac{1}{2}(1-s)$	N	N	N	N
9.	– bc –	s_{15}	$\frac{4bc}{m^2}$	s	Y	N	N	N
10.	Driver & Kroeber (1932), Ochiai (1957)	s_{16}	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$1-s$	Y	Y	Y	N
11.	Sokal & Sneath (1963), un_5	s_{17}	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$1-s$	Y	Y	Y	N
12.	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
13.	Baroni-Urbani, Buser (1976), S^{**}	s_{19}	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$1-s$	Y	Y	Y	N
14.	Braun-Blanquet (1932)	s_{20}	$\frac{a}{\max(a+b,a+c)}$	$1-s$	Y	Y	Y	Y
15.	Simpson (1943)	s_{21}	$\frac{a}{\min(a+b,a+c)}$	$1-s$	Y	N	N	N
16.	Michael (1920)	s_{22}	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$\frac{1}{2}(1-s)$	N	Y	Y	N

Table 1: Association coefficients

THEOREM 1 *For any even nonnegative dissimilarity measure d on $\mathcal{E}$ for its metric index p we have*
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 evaluate the values $d(X,Y)$, $d(X,Z)$ and $d(Y,Z)$ and we rename the vectors so that

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

We can assume that $d(X,Y) > 0$: by the theorem assumption d is even and therefore $d(X,Y) = 0$ would imply $d(X,Z) = d(Y,Z)$. The triangle inequality holds for such a triple as an equality or as an inequality $0 \leq 2\,d(X,Z)$ – they remain valid also for all d^α, $\alpha > 0$.

An evident consequence of the inequality

$$0 \leq \alpha \leq 1: \qquad u^\alpha + v^\alpha \geq (u+v)^\alpha$$

is that for all triples $X, Y, Z \in \mathcal{E}$ for which the triangle inequality holds, 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 the triangle inequality holds on $\mathcal{E}$ then $p(d) \geq 1$. Now, if the triangle inequality holds for all d^α, $\alpha > 1$, d is an ultrametric (see also Résultat 2 in Joly, Le Calvé (1986)). To prove this, let for all $\alpha > 1$

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

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

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

When there will be no possible confusion we'll omit the indexes. We have $u \leq v \leq 1$ and for all $\alpha > 1$

$$1 \leq u^\alpha + v^\alpha.$$

This is possible only if at least v is equal to 1. Therefore d is ultrametric and $p(d) = \infty$.

In the case $p < \infty$ there exist 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 $\alpha > 1$ we define $\delta := d^\alpha$. We have

$$\delta(X,Y) + \delta(X,Z) \le \delta(Y,Z).$$

For δ its metric index $p(\delta) < 1$ and it holds $p(d) = \alpha\, p(\delta)$.

Therefore, without loss of generality, we can limit our 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 with $d(Y,Z)$ we get

$$0 \;<\; u + v \;<\; 1\,.$$

Let us observe the function $f\colon [0,1] \to \mathbb{R}$ defined by

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

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

$$f'(\alpha) = u^\alpha \,\ln u + v^\alpha \,\ln v < 0 \qquad \text{for} \quad 0 < u, v < 1\,.$$

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

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

Now, if for all triples $X, Y, Z \in \mathcal{E}$ that fail the triangle inequality, we compute the values α_0 then $p = \inf \alpha_0 \ge 0$. $\boxed{\text{Q.E.D.}}$

COROLLARY 2 *If the set of units $\mathcal{E}$ is finite $p(d) > 0$.*

Proof: See the proof of the theorem. Since $\alpha_0 > 0$, because there is a finite number of α_0s, $p = \inf \alpha_0 > 0$. $\boxed{\text{Q.E.D.}}$

REMARK 3 For a nonnegative dissimilarity measure d on $\mathbb{B}^m$ let us denote $p_m = p_m(d)$ its metric index and

$$p_\infty(d) \;:=\; \inf\{p_m(d);\; m \in \mathbb{N}\}$$

the metric index of d valid on all sets of units $\mathbb{B}^m$.

REMARK 4 Let d be metric on $\mathcal{E}$. If there exists a triple $X, Y, Z \in \mathcal{E}$ such that $d(X,Z),\ d(Y,Z) > 0$ and

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

then $p(d) = 1$.

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

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

we have

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

Therefore, for $\alpha > 1$, d^α is not metric. $\boxed{\text{Q.E.D.}}$

3.2 The method

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

$$u_{xyz}^\alpha + v_{xyz}^\alpha = 1$$

holds for some triple $x, y, z \in \mathbb{B}^m$.

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

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

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

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

It is continuous and from its derivatives on $\mathcal{R}$

$$\begin{aligned}
&F'_u = \alpha u^{\alpha-1} > 0,\\
&F'_v = \alpha v^{\alpha-1} > 0,\\
&F'_\alpha = u^\alpha \ln u + v^\alpha \ln v < 0,\\
&F''_{uu} = \alpha(\alpha-1)u^{\alpha-2} < 0,\\
&F''_{vv} = \alpha(\alpha-1)v^{\alpha-2} < 0,\\
&F''_{\alpha\alpha} = u^\alpha \ln^2 u + v^\alpha \ln^2 v > 0
\end{aligned}$$

we see that F is increasing and concave in the directions u and v and decreasing and convex in the direction α.

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

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

and compute

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

Now we restrict ourselves to

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

and compute

$$v_0 := \min\{v_{xyz};\ (x,y,z) \in \mathcal{B}_0\};$$

finally, α_0 is determined by

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

PROPOSITION 5 *If $v_0 = u_0$ then $p_m = \alpha_0$ is the metric index of d on $\mathbb{B}^m$.*

REMARK 6 For $u_0 = v_0$ the exponent α_0 is a solution of the equation

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

so we have a formula:

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

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

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

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

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

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

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

Since F is a decreasing function of α

$$\alpha_0 \leq \alpha$$

and $p_m = \alpha_0$ holds. $\boxed{\text{Q.E.D.}}$

Question: What if $v_0 > u_0$?

Let us denote $s := u + v$ an 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

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

It is continuous there and from its derivatives on $\mathcal{W}$

$$\begin{aligned}
&\Phi'_u = \alpha u^{\alpha-1} - \alpha(s-u)^{\alpha-1} > 0,\\
&\Phi'_s = \alpha(s-u)^{\alpha-1} > 0,\\
&\Phi'_\alpha = u^\alpha \ln u + (s-u)^\alpha \ln(s-u) < 0
\end{aligned}$$

we see that Φ is increasing in the directions u and s and decreasing in the α direction.

We denote $s_0 := u_0 + v_0$ and compute

$$u = \min\{u_{xyz};\ (x, y, z) \in \mathcal{B} \setminus \mathcal{B}_0\}.$$

Again we restrict to

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

and compute

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

PROPOSITION 7 *If $s \geq s_0$ than $\alpha > \alpha_0$ holds.*

Proof: Because $u_0 < u$, $s_0 \leq s = u + v$ and Φ is 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 exists a unique value α such that

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

Since Φ is decreasing function in α

$$\alpha_0 < \alpha$$

holds. Q.E.D.

REMARK 8 If $s \geq s_0$ for all $(x, y, z) \in \mathcal{B} \setminus \mathcal{B}_0$ than $p_m = \alpha_0$ holds.

Question: What if $s < s_0$?

PROPOSITION 9 *If* $u \geq 2^{-\frac{1}{\alpha_0}}$ *then* $p_m = \alpha_0$.

Proof: As the values $u_0,\ v_0\ (u_0 < v_0), \alpha_0$ and u are determined we evaluate

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

Since $d(x, y) \leq d(x, z)$ 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

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

Q.E.D.

EXAMPLE 10 For the values $u_0 = 0.1$ and $v_0 = 0.5$ we compute the value $\alpha_0 = 0.519$ and the boundary value $u_b = 0.263$ (see Table 2). If the next value of u

$$u = \min\{u_{xyz};\ (x, y, z) \in \mathcal{B} \setminus \mathcal{B}_0\}$$

that we compute is greater than 0.263 the value of p_m is determined: $p_m = 0.519$.

Table 2: Boundary value of u

u_0	v_0	α_0	u_b
0.1	0.5	0.519	0.263

Now we can summarize the whole procedure of determining p_m and p_∞:

1. STEP: We compute u_0 and v_0 . If $v_0 = u_0$ then $p_m = -\frac{\log 2}{\log u_0}$;

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

3. STEP: If $u < u_b$ then we determine $\mathcal{B}_1, v$ and s.
If $s \geq s_0$ we set $\mathcal{B}_0 = \mathcal{B}_1 \cup \mathcal{B}_0$ and we continue with step 2. Otherwise for
CASE A: $F(u, v, \alpha_0) = u^{\alpha_0} + v^{\alpha_0} \geq 1$ we again set $\mathcal{B}_0 = \mathcal{B}_1 \cup \mathcal{B}_0$ and we continue with step 2.
CASE B: $F(u, v, \alpha_0) = u^{\alpha_0} + v^{\alpha_0} < 1$ we set $u_0 = u$, $v_0 = v$, $\mathcal{B}_0 = \mathcal{B}_1 \cup \mathcal{B}_0$ and we 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

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

It may happen that p_m does not depend on m.

3.2.1 The Method Written in Integer Coordinates

We can express any definition of a dissimilarity measure on binary vectors in the form that involves counters on three binary vectors. For vectors $x, y, z \in \mathbb{B}^m$ we count frequencies of all possible triples componentwise. There are eight possible triples from 111 to 000 and we denote their frequencies with $A, B, \ldots, H$ (see Table 3). It

Table 3: 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

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

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

we can express

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

where δ_{xy}, δ_{xz} and δ_{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. We denote

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

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

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

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

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

3.3 Local optimization procedure

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

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

We use a local optimization procedure

initial (read, random) x, y, z;
$p := 1$;
while $\exists (u,v,w) \in N(x,y,z) : d^p(u,v) + d^p(v,w) < d^p(u,w)$ **do begin**
    $p :=$Solve($\alpha : d^\alpha(u,v) + d^\alpha(v,w) = d^\alpha(u,w)$);
    $x := u;\ y := v;\ z := w$
end

over a neighborhood of the initial triple

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

This means that we get a triple of binary vectors (x', y', z') from the neighborhood if, at the selected position, 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:

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

EXAMPLE 11 For the Sokal & Sneath dissimilarity $d_{11} = 1 - un_4$ we obtain for $m = 20$ the (local) optimum solution

$$\begin{aligned} x^* &= [\,00000000000011111111\,], \\ y^* &= [\,00000000000011111110\,], \\ z^* &= [\,00000000000000000001\,] \end{aligned}$$

with the corresponding dissimilarities

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

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

In the integer coordinate terminology: the upper bound of the metric index p_{20} equal to $\overline{p}_{20} = 0.500366330$ is attained when $B = 7$, $C = 1$ and $H = 12$.

4 The results

All results presented in following sections are obtained with the method described in the section 3.2.1 and supported also by the local optimization procedure.

4.1 Russel and Rao

If we write Russel and Rao dissimilarity measure

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

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

$$\begin{aligned} \delta_{xy}(A, B, \ldots, H) &= (C + D + E + F + G + H)/m\,, \\ \delta_{xz}(A, B, \ldots, H) &= (B + D + E + F + G + H)/m \quad \text{and} \\ \delta_{yz}(A, B, \ldots, H) &= (B + C + D + F + G + H)/m\,. \end{aligned}$$

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

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

The minimal positive value $u_0 = 1/m$ is attained when

$$\begin{aligned} &C+D+E+F+G+H = 1 \quad \text{and} \\ &B+C+D+F+G+H = m \end{aligned}$$

that is for $A = E = 0$ and $B = m-1$ and $C = 1$ or $S = 1$. We get v_0 when

$$B = m-1 \quad \text{and } C = 1: \qquad v_0 = \frac{m-1}{m}.$$

In this case

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

and so $\alpha_0 = 1$. Since

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

$p_m = \alpha_0 = 1$ and therefore $p_\infty = 1$ (see remark 4). This is in agreement with the fact that the triangle inequality holds for d_1 (see Table 1).

4.2 Families S_θ and T_θ

Gower and Legendre (1986) 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 4). Investigating metric and Euclidean properties of the dissimilarities $1 - S_\theta$, $1 - T_\theta$ and $\sqrt{1-S_\theta}$, $\sqrt{1-T_\theta}$ they found out that these properties depend on θ: there exists (critical) value θ_M such that for values of θ near 0 these dissimilarities are not metric – but for $\theta \ge \theta_M$ they become metric.

These results are summarized in Gower's and Legendre's theorems 9 and 10.

Theorem 9.
$1 - S_\theta$ is metric for $\theta \ge 1$ and $\sqrt{1-S_\theta}$ is metric for $\theta \ge 1/3$. If $\theta < 1$ then $1 - S_\theta$ may not be metric and if $\theta < 1/3$, $\sqrt{1-S_\theta}$ may be non-metric.

Theorem 10.
Says the same for $1 - T_\theta$ and $\sqrt{1-T_\theta}$.

Table 4: Some well-known similarities.

	definition	θ	p_∞
Kendall, Sokal-Michener (1958)	$\frac{a+d}{m}$	S_1	1
Rogers and Tanimoto (1960)	$\frac{a+d}{m+b+c}$	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 and Sneath	$\frac{a}{a+2(b+c)}$	T_2	$\frac{\log 2}{\log \frac{3}{2}} = 1.709$

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 answer our question.

In the integer coordinates we express

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

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

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

Investigating all possibilities to get the smallest positive value of $u_{xyz} = \delta_{xy}/\delta_{yz}$ gives us the (not unique) solution $B = m-1$ and $C = 1$. In this case

$$s_0 = u_0 + v_0 = \frac{1}{t\,(m-1)+1} + \frac{m-1}{t+m-1} = \frac{t\,(m^2-2m+2)+2m-2}{t^2\,(m-1)+t\,(m^2-2m+2)+m-1}$$

and $s_0 < 1$ holds for $t > 1$.

To increase u we pose more generally $B+C+H=m,\ B,C>0$ ($B=m-1,\ C=1$ is just a special case) and calculate the sum s

$$s=u+v=\frac{C}{t\,B+C}+\frac{B}{t\,C+B}=\frac{t\,(B^2+C^2)+2\,BC}{t^2\,BC+t\,(B^2+C^2)+BC}.$$

With the notation $\xi:=B/C$ we express this 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

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

we see that s is decreasing with t and that $s'_\xi=0$ for $t=0,\pm1$ or $\xi=\pm1$. Since $t>1$ and $\frac{1}{m-1}\leq\xi\leq m-1$ we get, when t is fixed, the minimal sum for $\xi=1$, that is for $B=C$. In this case

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

PROPOSITION 12 *If* $B>0,\ C>0$ *and* $\alpha=\frac{\log 2}{\log(t+1)},\quad t>1,$ *then*

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

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 $\mathcal{R}:=\{\,(t,\xi);\,t>1\wedge 0\leq\xi\leq 1\}$;
- $\phi(t,0)=\phi(t,1)=1$;
- $\phi(1,\xi)=1$ and $\lim_{t\to\infty}\phi(t,\xi)=1$.

Searching for the minimal value of the function ϕ 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 $\phi''_{\xi\xi}(t,1) > 0$ iff $\alpha\, t - 1 > 0$, that is iff

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

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'll show that there exists 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 $\xi,\ 0 < \xi < 1$, we rewrite the equation $\phi'_\xi(t,\xi) = 0$:

$$\begin{aligned} \xi^2 &= \left(\frac{t\,\xi^2+\xi}{t+\xi}\right)^{\alpha+1} \\ \log \xi^2 &= (1+\alpha)\ \log \frac{t\,\xi^2+\xi}{t+\xi} \end{aligned}$$

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

$$\begin{aligned} f'(\xi) &= \tfrac{2}{\xi}\,, &\qquad g'_t(\xi) &= (1+\alpha)\,\tfrac{t\left(1+2\,t\,\xi+\xi^2\right)}{\xi\,(t+\xi)\,(1+t\,\xi)}\,, \\ f''(\xi) &= -\tfrac{2}{\xi^2}\,, &\qquad g''_t(\xi) &= (1+\alpha)\,\tfrac{-t\left(t+2\,\xi+2\,t^2\,\xi+4\,t\,\xi^2+2\,t^3\,\xi^2+4\,t^2\,\xi^3+t\,\xi^4\right)}{\xi^2\,(t+\xi)^2\,(1+t\,\xi)^2} \end{aligned}$$

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

$$f'(\xi) = \tfrac{2}{\xi} \qquad \text{and}$$
$$g_t'(\xi) = (1+\alpha)\,\left(\tfrac{1}{\xi} + \tfrac{t^2-1}{t} + \tfrac{1-t^4}{t^2}\,\xi + \mathrm{O}(\xi^2)\right) \overset{\xi\to 0}{\longrightarrow} \tfrac{1+\alpha}{\xi}$$

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 $(\xi_t, f(\xi_t)) = (\xi_t, g_t(\xi_t))$, $\xi_t \in (0,1)$. Q.E.D.

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

$$p_\infty = p_m = \frac{\log 2}{\log(t+1)}\,.$$

Now we'll 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 for any triple $x,y,z \in \mathbb{B}^m$ in the integer coordinates, we get

$$\delta_{xy}(A,B,\ldots,H;t) = \frac{C+D+E+F}{t\,(A+B+G+H)+C+D+E+F}\,,$$
$$\delta_{xz}(A,B,\ldots,H;t) = \frac{B+D+E+G}{t\,(A+C+F+H)+B+D+E+G}\,,$$
$$\delta_{yz}(A,B,\ldots,H;t) = \frac{B+C+F+G}{t\,(A+E+D+H)+B+C+F+G}\,.$$

We get u_0 if

$$\delta_{xy} = \frac{1}{t\,(m-1)+1} \qquad \text{and} \qquad \delta_{yz} = 1.$$

That is when $A+B+G+H = m-1$,
$C+D+E+F = 1$,
$A+E+D+H = 0$ and
$B+C+F+G > 0$.

Therefore: $A = E = D = H = 0$, $B + G = m - 1$ and $C = 1$ or $F = 1$. There are two possibilities, but since

$$\delta_{..}(\ldots, C, \ldots, F, \ldots) = \delta_{..}(\ldots, F, \ldots, C, \ldots)$$

we need to consider only one

$$B + G = m - 1 \quad \text{and } C = 1: \qquad u_0 = \frac{1}{t\,(m-1)+1}\,, \qquad v_0 = \frac{m-1}{t+m-1}\,.$$

Since this is just the special case of $B + C + G = m$ we have more generally

$$u = \frac{C}{t\,(B+G)+C} \qquad \text{and} \qquad v = \frac{B+G}{t\,C+B+G}\,.$$

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 Proposition 12 we get the minimal exponent α for $\xi = 1$, that is, when $B + G = C$. So we have

$$B + C + G = m \qquad \text{and} \qquad B + G = C$$

that implies $m = 2\,C$; m is even $(m = 2n,\ n \in \mathbb{N})$. In this case

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

If m is odd $(m = 2n + 1,\ n \in \mathbb{N})$ ξ is never equal to 1, what means that $u^\alpha + v^\alpha$ is always greater then 1 and hence

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

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

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

This is the value of distance index p_∞ also for the second family.

4.3 Kulczinsky T

In Batagelj and Bren (1995) we defined the values of the dissimilarity $T = (b+c)/a$ in the cases of indeterminacy by

$$T(x,y) \;:=\; \begin{cases} 0 & \text{for } a_{xy} = 0,\; d_{xy} = m \\ \infty & \text{for } a_{xy} = 0,\; d_{xy} < m \end{cases}$$

Therefore, for any triple $x, y, z \in \mathbb{B}^m$ such that $a_{xy}, a_{xz} \neq 0$, but $a_{yz} = 0$ and $b_{yz} + c_{yz} > 0$ we have

$$T(x,y) = \frac{b_{xy} + c_{xy}}{a_{xy}}\,, \qquad T(x,z) = \frac{b_{xz} + c_{xz}}{a_{xz}} \qquad \text{and} \quad T(y,z) = \infty.$$

Hence, if we define $u^0 = 1$ for all $u \in \overline{\mathbb{R}^+}$ p_m is equal to 0.

EXAMPLE 13 For the triple

$\mathbf{1}$:= [0,. . . ,0,0,1]
$\mathbf{2}$:= [0,. . . ,0,1,0]
$\mathbf{3}$:= [0,. . . ,0,1,1]

we count

pair	a	b	c	d
1 3	1	0	1	$m-2$
2 3	1	0	1	$m-2$
1 2	0	1	1	$m-2$

and compute

$$T(\mathbf{1},\mathbf{3}) = T(\mathbf{2},\mathbf{3}) = 1\,, \qquad T(\mathbf{1},\mathbf{2}) = \infty.$$

4.4 Kulczinsky

The Kulczinsky's dissimilarities written in the integer coordinates for $x, y, z \in \mathbb{B}^m$ are

$$\begin{aligned}
\delta_{xy}(A,B,\ldots,H) &= \tfrac{1}{2}\left(\tfrac{C+D}{A+B+C+D} + \tfrac{E+F}{A+B+E+F}\right), \\
\delta_{xz}(A,B,\ldots,H) &= \tfrac{1}{2}\left(\tfrac{B+D}{A+C+B+D} + \tfrac{E+G}{A+C+E+G}\right), \\
\delta_{yz}(A,B,\ldots,H) &= \tfrac{1}{2}\left(\tfrac{B+F}{A+E+B+F} + \tfrac{C+G}{A+E+C+G}\right).
\end{aligned}$$

If we assume $0 < \delta_{xy} \le \delta_{xz} \le \delta_{yz}$ the minimal positive value $u_0 = 1/(2m)$ is attained when $\delta_{xy} = 1/(2m)$ and $\delta_{yz} = 1$. We have three possibilities

$$B = m-1 \ \wedge \ (C = 1 \ \veebar \ D = 1 \ \veebar \ F = 1)$$

but only the first one $B = m-1$ and $C = 1$ gives the minimal value of v:

$$v_0 = \frac{1}{2}\frac{m-1}{m} = \frac{1}{2} - \frac{1}{2m}.$$

The sum is $s_0 = 1/2$.

Now, to increase u, we put $B + C + H = m,\ B > 0,\ C > 0$. We get

$$u = \frac{1}{2}\frac{C}{B+C}, \qquad v = \frac{1}{2}\frac{B}{B+C} \qquad \text{and} \quad s = \frac{1}{2}.$$

As we proved in the section 3.1 there exist a unique value α_0 such that

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

and since ϕ is increasing function of u we have

$$1 = u_0^{\alpha_0} + v_0^{\alpha_0} = \phi(u_0, \frac{1}{2}, \alpha_0) \le \phi(u, \frac{1}{2}, \alpha_0)$$

for all u. Therefore $p_m = \alpha_0$ is the solution of the equation

$$(m-1)^{p_m} + 1 = (2m)^{p_m}\,. \tag{1}$$

For different dimensions we compute

m	2	3	4	5	$\cdots$	20	$\cdots$	50	$\cdots$	100
p_m	0.5	0.4895	0.4749	0.4620	$\cdots$	0.3798	$\cdots$	0.3357	$\cdots$	0.3085

The values of metric indexes written in this and following tables are exact to the fourth decimal place and according to the statement of Joly and Le Calvé in subsection 2.1 if we take less decimals the values have to be exact too.

From the equation (1) we see also that for large dimensions

$$p_m \stackrel{m\to\infty}{\longrightarrow} 0 = p_\infty\,.$$

4.5 Sokal & Sneath un_4

Sokal & Sneath dissimilarity $1 - un_4$ written in integer coordinates for $x, y, z \in \mathbb{B}^m$ gives

$$\delta_{xy}(A,B,\ldots,H) = \tfrac{1}{4}\left(\tfrac{C+D}{A+B+C+D} + \tfrac{E+F}{A+B+E+F} + \tfrac{C+D}{C+D+G+H} + \tfrac{E+F}{E+F+G+H}\right),$$
$$\delta_{xz}(A,B,\ldots,H) = \tfrac{1}{4}\left(\tfrac{B+D}{A+C+B+D} + \tfrac{E+G}{A+C+E+G} + \tfrac{B+D}{B+D+F+H} + \tfrac{E+G}{E+G+F+H}\right),$$
$$\delta_{yz}(A,B,\ldots,H) = \tfrac{1}{4}\left(\tfrac{B+F}{A+E+B+F} + \tfrac{C+G}{A+E+C+G}\tfrac{B+F}{B+F+D+H} + \tfrac{C+G}{C+G+D+H}\right).$$

If we assume $0 < \delta_{xy} \le \delta_{xz} \le \delta_{yz}$ and consider all possibilities to get the lowest value of u we get the (not unique) solution $B + C + H = m$. If we put $B = m - C - H$ we have

$$u(C,H) = \frac{C\,(m^2 - C^2)}{(m-H)\,(-4\,C^2 - 4\,C\,H - H^2 + 4\,C\,m + 3\,H\,m)}\,.$$

PROPOSITION 14 *The function $u(C,H)$ is increasing in C.*

Proof: We denote $\chi := H/m$, $\zeta := C/m$ and since $B + C + H = m$ we have $0 < \chi + \zeta < 1$. With this notation we rewrite

$$u(\chi,\zeta) = \frac{\zeta\,(1-\zeta^2)}{(1-\chi)\,(3\,\chi - \chi^2 + 4\,\zeta - 4\,\chi\,\zeta - 4\,\zeta^2)}\,.$$

The first derivative

$$u'_\zeta = \frac{3\,\chi - \chi^2 + 4\,\zeta^2 - 9\,\chi\,\zeta^2 + 3\,\chi^2\,\zeta^2 - 8\,\zeta^3 + 8\,\chi\,\zeta^3 + 4\,\zeta^4}{(1-\chi)\,(3\,\chi - \chi^2 + 4\,\zeta - 4\,\chi\,\zeta - 4\,\zeta^2)^2}$$

is indeed positive for $\chi, \zeta > 0$ and $\chi + \zeta < 1$:
the denominator is positive;
we rewrite the nominator in the form of a quadratic polynomial

$$P(\chi) := (3\zeta^2 - 1)\chi^2 + (8\zeta^3 - 9\zeta^2 + 3)\chi + 4\,\zeta^2\,(1-\zeta^2)\,.$$

When $\zeta = \frac{1}{\sqrt{3}}$ we have

$$P(\chi) = \frac{8}{3\sqrt{3}}\,\chi + \frac{16}{9} - \frac{8}{3\sqrt{3}} > 0$$

for $\chi > 1 - \frac{2}{\sqrt{3}} \doteq -0.15$.
When $\zeta \ne \frac{1}{\sqrt{3}}$ we have

$$D := 9 - 38\,\zeta^2 + 16\,\zeta^3 + 49\,\zeta^4 - 48\,\zeta^5 + 16\,\zeta^6 > 0$$

for $0 < \zeta < 1$ and the roots

$$\chi_{1,2}(\zeta) = \frac{-8\zeta^3 + 9\zeta^2 - 3 \pm \sqrt{D}}{2(3\,\zeta^2 - 1)}$$

are for

- $\zeta < \frac{1}{\sqrt{3}}$: $\chi_1 < 0$ and $\chi_2 > 1$
- $\zeta > \frac{1}{\sqrt{3}}$: $\chi_1 < 0$ and $\chi_2 < 0$.

Therefore $P(\chi)$ is also positive for $0 < \chi < 1$. $\boxed{\text{Q.E.D.}}$

PROPOSITION 15 *The sum $s = u + v$ is minimal if $C = 1$.*

Proof: We put $H = m - B - C$, we denote $\beta := B/m,\ \zeta := C/m$ and we rewrite

$$s(\beta, \zeta) = \frac{-\beta^2 + \beta\,\zeta - \zeta^2 + 1}{-\beta^2 + 2\,\beta\,\zeta - \zeta^2 - \beta - \zeta + 2}\,.$$

The first derivative

$$s'_\zeta = \frac{(1-\beta)\ \left(-\beta^2 + \zeta^2 + \beta - 2\,\zeta + 1\right)}{\left(\beta^2 - 2\,\beta\,\zeta + \zeta^2 + \beta + \zeta - 2\right)^2}$$

is positive for $\beta > 0$, $\zeta > 0$ and $\beta + \zeta < 1$:
the denominator is positive; the first factor of nominator is positive; the second factor we rewrite in the form of a quadratic polynomial

$$P(\zeta) := \zeta^2 - 2\,\zeta - \beta^2 + \beta + 1$$

that is positive too since $D = 4\,(\beta^2 - \beta) < 0$ for $\beta < 1$. $\boxed{\text{Q.E.D.}}$

COROLLARY 16 *If $B + C + H = m$ then the minimal exponent we get when $C = 1$.*

Proof: We fix the value B. We put $C = 1$, $H = m - B - 1$ and compute $u_0 := u(1/m, \chi)$ and $s_0 := s(\beta, 1/m)$ the minimal values of u and s when $B + C + H = m$. Since the function $\phi(u, s, \alpha)$ is increasing in u and in s we get for the unique value α_0 that solves the equation

$$u_0^{\alpha_0} + (s_0 - u_0)^{\alpha_0} = 1$$

the inequality

$$1 = \phi(u_0, s_0, \alpha_0) < \phi(u, s, \alpha_0).$$

Therefore any value α that solves the equation

$$u^\alpha + (s-u)^\alpha = 1$$

satisfies the inequality

$$1 = \phi(u, s, \alpha) < \phi(u, s, \alpha_0)$$

and since ϕ is decreasing function in α we get $\alpha_0 \leq \alpha$. Q.E.D.

Now, for $B + H = m - 1$ and $C = 1$ we have

$$u = \frac{(B+H)\ (2+B+H)}{(1+B)\ (4\,B + 3\,H + 3\,B\,H + 2\,H^2)}.$$

If we put $H = m - B - 1$ and rewrite

$$u = \frac{m^2 - 1}{(1+B)\ (-1 + 2\,B - B^2 - m - B\,m + 2\,m^2)}$$

the minimal value of u is attained when $B = (m - 1 + \sqrt{7m^2 - 8m + 4})/3$. That is for large m approximately $0.55\,m$. For $m = 20$ we have

B	1	2 $\cdots$	6	7	8	9	10	11	12 $\cdots$	19
H	18	17	13	12	11	10	9	8	7	0
u	0.26	0.18	0.089	0.082	0.078	0.074	0.0727	0.0723	0.0732	0.26
s	0.53	0.54	0.58	0.59	0.60	0.61	0.62	0.63	0.64	0.75
α	0.518	0.516	0.5004	0.5003	0.501	0.504	0.507	0.512	0.519	0.696

We see that $p_{20} = 0.5003$ we get when $B = 7$, $C = 1$ and $H = 12$. The local optimization procedure also confirms this result (see Example 11).

For different dimensions we calculate

m	3	4	5	$\cdots$	20	$\cdots$	50
B	1	1	1		7		19
C	1	1	1		1		1
H	1	2	3		12		30
p_m	0.6309	0.5959	0.5757		0.5003		0.4369

Therefore the extremal triple has the form

$$\begin{aligned} x^* &= [\overbrace{0,\ldots,0}^{3n},\overbrace{0,\ldots,0,1}^{2n}] \\ y^* &= [0,\ldots,0,1,\ldots,1,1] \\ z^* &= [0,\ldots,0,1,\ldots,1,0]. \end{aligned}$$

For $m = 5n,\ n \in \mathbb{N}$, we put $B = 2n-1,\ C = 1,\ H = 3n$ and we rewrite

$$u = \frac{25\,n^2-1}{8n\,(9\,n^2+2\,n-1)}\,, \qquad v = \frac{(2\,n-1)\,(3\,n+1)\,(7\,n-1)}{8n\,(9\,n^2+2\,n-1)}\,.$$

p_{5n} is the solution of the equation

$$u^{p_{5n}} + v^{p_{5n}} = 1 \tag{2}$$

that we solve numerically (see Table 5).

Table 5: The metric indexes for $1 - un_4$ dissimilarity

m	5	10	15	20	$\cdots$	50	$\cdots$	100
u	0.3	0.19	0.108	0.083		0.034		0.017
v	0.3	0.383	0.484	0.508		0.553		0.568
p_m	0.5757	0.5482	0.5209	0.5003		0.4369		0.3954

From the equation (2) we see also that for large dimensions

$$p_m \overset{m\to\infty}{\longrightarrow} 0 = p_\infty\,.$$

4.6 Driver & Kroeber, Ochiai

The Driver & Kroeber dissimilarity $1 - s_{16} = 1 - \frac{a}{\sqrt{(a+b)\,(a+c)}}$ written in integer coordinates for $x, y, z \in \mathbb{B}^m$ gives

$$\delta_{xy}(A, B, \ldots, H) = 1 - \frac{A+B}{\sqrt{(A+B+C+D)\,(A+B+E+F)}},$$
$$\delta_{xz}(A, B, \ldots, H) = 1 - \frac{A+C}{\sqrt{(A+C+B+D)\,(A+C+E+G)}},$$
$$\delta_{yz}(A, B, \ldots, H) = 1 - \frac{A+E}{\sqrt{(A+E+B+F)\,(A+E+C+G)}}.$$

If $0 < \delta_{xy} \le \delta_{xz} \le \delta_{yz}$ the smallest positive value of u is attained when

$$\delta_{yz} = 1: \qquad A + E = 0 \;\Rightarrow\; A = E = 0$$

and

$$\delta_{xy} \text{ is minimal:} \qquad A + B = m - 1 \;\wedge (C = 1 \;\veebar\; D = 1 \;\veebar\; E = 1 \;\veebar\; F = 1).$$

Only the first possibility $C = 1$ gives the value v_0 less then 1. Therefore we put $B + C = m$ and we get

$$u = 1 - \sqrt{\frac{B}{m}}, \qquad v = 1 - \sqrt{\frac{C}{m}} \qquad \text{and } s = 2 - \sqrt{\frac{B}{m}} - \sqrt{\frac{C}{m}}.$$

The minimal positive value of u is attained when $B = m - 1$:

$$u_0 = \min u = 1 - \sqrt{\frac{m-1}{m}},$$

but the minimal value of s when $B = C = m/2$:

$$\min s = 2 - \sqrt{2}.$$

From Table 6 we see that $p_{20} = 0.553$ we get when $B = 19$ and $C = 1$ (or $B = 1,\ C = 19$). Therefore the extremal triple has the form

Table 6: Values of u, s and α for different B when $m = 20$.

B	19	18	17	$\cdots$	11	10	9	$\cdots$	2	1
u	0.025	0.051	0.078		0.258	0.293	0.329		0.684	0.776
s	0.80	0.73	0.69		0.587	0.586	0.587		0.73	0.80
α	0.553	0.557	0.559		0.56440	0.56448	0.56440		0.557	0.553

$$
\begin{aligned}
x^* &= [1,\ldots,1,1] \\
y^* &= [1,\ldots,1,0] \\
z^* &= [0,\ldots,0,1]
\end{aligned}
$$

For different dimensions we calculate

m	3	4	5	$\cdots$	20	$\cdots$	50	$\cdots$	100
p_m	0.5636	0.5624	0.5613		0.5534		0.5484		0.5449

Since u_0 tends to 0 for large m but $\min s$ remains constant the extremal triple has the same form for all dimensions. It holds

$$p_m \stackrel{m\to\infty}{\longrightarrow} 0.5 = p_\infty$$

4.7 Sokal & Sneath, un_5

The Sokal & Sneath dissimilarity $1 - un_5 = 1 - \frac{ad}{\sqrt{(a+b)\,(a+c)\,(d+b)\,(d+c)}}$ written in integer coordinates for $x, y, z \in \mathbb{B}^m$ gives

$$
\begin{aligned}
\delta_{xy}(A,B,\ldots,H) &= 1 - \frac{(A+B)\,(G+H)}{\sqrt{(A+B+C+D)\,(A+B+E+F)\,(G+H+C+D)\,(G+H+E+F)}}\,, \\
\delta_{xz}(A,B,\ldots,H) &= 1 - \frac{(A+C)\,(F+H)}{\sqrt{(A+C+B+D)\,(A+C+E+G)\,(F+H+B+D)\,(F+H+E+G)}}\,, \\
\delta_{yz}(A,B,\ldots,H) &= 1 - \frac{(A+E)\,(D+H)}{\sqrt{(A+E+B+F)\,(A+E+C+G)\,(D+H+B+F)\,(D+H+C+G)}}\,.
\end{aligned}
$$

If $0 < \delta_{xy} \leq \delta_{xz} \leq \delta_{yz}$ the smallest positive value of u is attained when

$$
\begin{array}{ll}
\delta_{yz} = 1: & (A + E = 0 \ \wedge \ B + F \neq 0 \ \vee \ C + G \neq 0) \vee \\
& (D + H = 0 \ \wedge \ B + F \neq 0 \ \vee \ C + G \neq 0) \\
\text{and} & \\
\delta_{xy} \text{ minimal:} & \max(A+B)\,(G+H) \ \wedge (C = 1 \ \veebar \ D = 1 \ \veebar \ E = 1 \ \veebar \ F = 1).
\end{array}
$$

Let us consider the first possibility $A = E = 0$ (for $D = H = 0$ results are the same):

To get minimal δ_{xy} we put $B + G + H = m - 1$. In this case only the possibility $C = 1$ gives the value δ_{xz} less then 1 and we have

$$u = 1 - \sqrt{\frac{B\,(G+H)}{(B+1)\,(G+H+1)}}\,, \qquad v = 1 - \frac{H}{\sqrt{(B+1)\,(G+1)\,(B+H)\,(G+H)}}\,.$$

The minimal value of u is attained when $B = G + H = (m-1)/2$ and in this case the minimal value of v when $G = 0$. Therefore, if m is odd ($m = 2n+1$), we get when $B = H = n$ and $C = 1$ the minimal values

$$u_0 = 1 - \frac{n}{n+1} \qquad \text{and} \qquad v_0 = 1 - \frac{1}{\sqrt{2\,(n+1)}}\,. \tag{3}$$

Indeed, for $B + C + H = m$ the function

$$u(B, C, H) = 1 - \sqrt{\frac{B\,H}{(B+C)\,(H+C)}}$$

is increasing in C. For $C = 1$ we rewrite

$$u(B, 1, H) = 1 - \sqrt{\frac{B\,H}{(B+1)\,(H+1)}}$$

and this function attaints the minimal value when $B = H = (m-1)/2$. Now for $m = 2n+1$ we have (3).

To increase u we pose $B + H = m - 1$. We get

$$u(B, 1, H) = 1 - \sqrt{\frac{B\,H}{(B+1)\,(H+1)}} \qquad \text{and} \qquad v(B, 1, H) = 1 - \sqrt{\frac{H}{(B+1)\,(B+H)}}\,.$$

The minimal value of s is attained when $B = C = 1$ and $H = m - 2$:

$$\min s = s(1, 1, m-2) = 2 - \sqrt{2\,\frac{m-2}{m-1}}\,.$$

From Table 7 we see that $p_{20} = 0.59$ we get when $B = C = 1$ and $H = 18$.

Table 7: The values of u, v, s and α for different B when $m = 20$.

B	1	2	$\cdots$	8	9	10	11	$\cdots$	17	18
u	0.31	0.21		0.097	0.095	0.095	0.097		0.21	0.31
v	0.31	0.45		0.746	0.771	0.792	0.813		0.92	0.95
s	0.62	0.66		0.843	0.8660	0.887	0.910		1.13	1.26
α	0.59	0.61		0.71	0.74	0.76	0.80		1.41	1.96

Therefore the extremal triple has the form

$$\begin{aligned} i^* &= [0,\ldots,0,1,1] \\ y^* &= [0,\ldots,0,1,0] \\ z^* &= [0,\ldots,0,0,1] \end{aligned}$$

Hence we get the minimal value of α when

$$u(1,1,m-2) = v(1,1,m-2) = 1 - \sqrt{\frac{m-2}{2\,(m-1)}}$$

and so

$$p_m = -\frac{\log 2}{\log\left(1 - \sqrt{\frac{m-2}{2\,(m-1)}}\right)}\,.$$

For different dimensions we calculate

m	3	4	5	$\cdots$	20	$\cdots$	50	$\cdots$	100
p_m	1	0.8048	0.7313		0.5946		0.5759		0.5701

We see that u and v are decreasing in m and so is p_m . For large m we have

$$u = v \overset{m\to\infty}{\longrightarrow} \frac{2-\sqrt{2}}{2} \qquad \text{and} \qquad p_m \overset{m\to\infty}{\longrightarrow} p_\infty = -\frac{\log 2}{\log\frac{2-\sqrt{2}}{2}} \doteq 0.5645\,.$$

4.8 Pearson, ϕ

The Pearson dissimilarity

$$\frac{1}{2}(1-\phi) = \frac{1}{2}\left(1 - \frac{a\,d - b\,c}{\sqrt{(a+b)\,(a+c)\,(d+b)\,(d+c)}}\right)$$

written in integer coordinates for $x, y, z \in \mathbb{B}^m$ gives

$$\delta_{xy}(A,B,\ldots,H) = \tfrac{1}{2}\left(1 - \frac{(A{+}B)\,(G{+}H) - (C{+}D)\,(E{+}F)}{\sqrt{(A{+}B{+}C{+}D)\,(A{+}B{+}E{+}F)\,(G{+}H{+}C{+}D)\,(G{+}H{+}E{+}F)}}\right)\,,$$

$$\delta_{xz}(A,B,\ldots,H) = \tfrac{1}{2}\left(1 - \frac{(A{+}C)\,(F{+}H) - (B{+}D)\,(E{+}G)}{\sqrt{(A{+}C{+}B{+}D)\,(A{+}C{+}E{+}G)\,(F{+}H{+}B{+}D)\,(F{+}H{+}E{+}G)}}\right)\,,$$

$$\delta_{yz}(A,B,\ldots,H) = \tfrac{1}{2}\left(1 - \frac{(A{+}E)\,(D{+}H) - (B{+}F)\,(C{+}G)}{\sqrt{(A{+}E{+}B{+}F)\,(A{+}E{+}C{+}G)\,(D{+}H{+}B{+}F)\,(D{+}H{+}C{+}G)}}\right)\,.$$

Using the usual calculation the minimal positive value of u is attained when

$\delta_{yz} = 1$: $\quad A = E = D = H = 0 \ \wedge \ B + F \neq 0 \ \wedge \ C + G \neq 0$
and
δ_{xy} is minimal: $\max(A + B)\,(G + H) \ \wedge (C = 1 \ \vee \ F = 1)$.

For $A = E = D = H = 0$ we get:

$$u(B, C, F, G) = \frac{1}{2}\left(1 - \frac{B\,G - C\,F}{\sqrt{(B + C)\,(B + F)\,(G + C)\,(G + F)}}\right),$$

$$v(B, C, F, G) = \frac{1}{2}\left(1 - \frac{C\,F - B\,G}{\sqrt{(B + C)\,(B + F)\,(G + C)\,(G + F)}}\right)$$

and $s = u + v = 1$. We obtain the same sum for

$$(C = 0 \ \vee \ F = 0) \ \veebar \ (B = 0 \ \vee \ G = 0).$$

Now let us put only $A = E = 0$ (for $D = H = 0$ result is the same) and also $D = 0$ in order to get lover values of δ_{xy} and δ_{xz}. Consider the $\delta_{..}(B, C, F, G, H)$ for $H > 0$ we see that

- δ_{yz} is maximal when $\quad B + F = C + G = (m - H)/2$ and $H = 1$. In this case

$$\max \delta_{yz} = \frac{1}{2}\left(1 + \frac{m-1}{m+1}\right) \overset{m\to\infty}{\longrightarrow} 1.$$

- δ_{xy} is minimal when $C = 0$ or $F = 0$ and $B \neq 0$.

- δ_{xz} is minimal when $B = 0$ or $G = 0$ and $C \neq 0$.

Therefore we put $F = G = 0$ (for $B = C = 0$ result is the same) and for $B + C + H = m$ we have

$$u(B, C, H) = \frac{1 - \sqrt{\frac{B\,H}{(B+C)\,(C+H)}}}{1 + \sqrt{\frac{B\,C}{(B+H)\,(C+H)}}}.$$

If we rewrite this using $B = m - C - H$ we get

$$u(m - C - H, C, H) = \frac{1 - \sqrt{\frac{H\,(m-C-H)}{(C+H)\,(m-H)}}}{1 + \sqrt{\frac{C\,(m-C-H)}{(C+H)\,(m-C)}}}$$

defined for $0 < C+H < m$. The minimal positive value of u is attained when $C = 1$:

$$v(B,1,H) = \frac{1 - \sqrt{\frac{H}{B+B^2+H+B\,H}}}{1 + \sqrt{\frac{B}{B+H+B\,H+H^2}}},$$

$$s(B,1,H) = \frac{2 - \sqrt{\frac{B\,H}{1+B+H+B\,H}} - \sqrt{\frac{H}{B+B^2+H+B\,H}}}{1 + \sqrt{\frac{B}{B+H+B\,H+H^2}}}$$

and the minimal value of s when $B = C = 1$:

$$\min s = s(1,1,m-2) = 2\,\frac{m-1}{m}\left(1 - \sqrt{\frac{m-2}{2\,(m-1)}}\right) \stackrel{m\to\infty}{\longrightarrow} 2 - \sqrt{2} \doteq 0.59\,.$$

For $m = 20$ we have

B	1	2	$\cdots$	9	10	11	12	13	$\cdots$	17	18
u	0.30	0.19		0.079	0.078	0.078	0.079	0.082		0.13	0.18
v	0.30	0.42		0.638	0.645	0.648	0.649	0.647		0.60	0.56
s	0.60	0.61		0.717	0.723	0.726	0.728	0.729		0.73	0.74
α	0.5696	0.57		0.580	0.582	0.585	0.588	0.592		0.63	0.67

We see that $p_{20} = 0.5696$ we get when $B = C = 1$ and $H = 18$. Therefore the extremal triple has the form

$$\begin{aligned} i^* &= [0,\ldots,0,1,1] \\ y^* &= [0,\ldots,0,1,0] \\ z^* &= [0,\ldots,0,0,1] \end{aligned}$$

Hence the minimal value of α is attained when

$$u(1,1,m-2) = v(1,1,m-2) = \frac{m-1}{m}\left(1 - \sqrt{\frac{m-2}{2\,(m-1)}}\right)$$

and so

$$p_m = -\frac{\log 2}{\log \frac{m-1}{m}\left(1 - \sqrt{\frac{m-2}{2\,(m-1)}}\right)}\,.$$

Since u, v and p_m are decreasing in m, we have

$$u = v \stackrel{m\to\infty}{\longrightarrow} \frac{2-\sqrt{2}}{2} \qquad \text{and} \qquad p_m \stackrel{m\to\infty}{\longrightarrow} p_\infty = -\frac{\log 2}{\log \frac{2-\sqrt{2}}{2}} \doteq 0.5645\,.$$

Table 8: The values of metric index of ϕ for different dimensions.

m	3	4	5	$\cdots$	20	$\cdots$	50	$\cdots$	100
p_m	0.6309	0.6033	0.5920		0.5696		0.5664		0.5654

4.9 Baroni-Urbani, S^{**}

The Baroni-Urbani dissimilarity

$$1 - S^{**} = \frac{b+c}{a+b+c+\sqrt{ad}}$$

written in integer coordinates for $x, y, z \in \mathbb{B}^m$ gives

$$\delta_{xy}(A, B, \ldots, H) = \frac{C+D+E+F}{A+B+C+D+E+F+\sqrt{(A+B)\,(G+H)}}\,,$$
$$\delta_{xz}(A, B, \ldots, H) = \frac{B+D+E+G}{A+B+C+D+E+G+\sqrt{(A+C)\,(F+H)}}\,,$$
$$\delta_{yz}(A, B, \ldots, H) = \frac{B+C+F+G}{A+B+C+E+F+G+\sqrt{(A+E)\,(D+H)}}\,.$$

Using the usual analysis the minimal positive value of u is attained when

$\delta_{yz} = 1$: $\quad A = E = 0 \ \wedge\ B + C + F + G \neq 0$
and
δ_{xy} minimal: $\quad \max(A + B)\,(G + H) \ \wedge\ C + D + E + F = 1.$

We set $A = E = 0$ and to get the minimal value of v with the minimal value of u also $D = F = G = 0$:

$$u(B, C, H) = \frac{C}{B + C + \sqrt{B\,H}} \qquad \text{and} \qquad v(B, C, H) = \frac{B}{B + C + \sqrt{C\,H}}\,.$$

Since $H = m - C - B$ we rewrite

$$u(B, C, m - B - C) = \frac{C}{B + C + \sqrt{B\ (m - B - C)}}$$

defined for $0 < B + C < m$. From the first derivative

$$u(B, C, m{-}B{-}C)'_C = -\frac{C\left(1 - \frac{B}{2\sqrt{B\,(m-B-C)}}\right)}{\left(B + C + \sqrt{B\ (m - B - C)}\right)^2} + \frac{1}{B + C + \sqrt{B\ (m - B - C)}}$$

we see that u is increasing in C. For $C = 1$ we have

$$s(B,1,H) = \frac{1}{1+B+\sqrt{BH}} + \frac{B}{1+B+\sqrt{H}}$$

and in this case

$$\min s = s(1,1,m-2) = \frac{2}{2+\sqrt{m-2}} \overset{m\to\infty}{\longrightarrow} 0\,.$$

For $m = 20$ and $C = 1$ we calculate

B	1	2	3	4	$\cdots$	16	17	18
u	0.160	0.113	0.091	0.078		0.042	0.042	0.043
s	0.320	0.394	0.466	0.529		0.896	0.918	0.943
α	0.378	0.412	0.439	0.462		0.708	0.745	0.799

We see that $p_{20} = 0.378$ we get when $B = C = 1$ and $H = 18$. Therefore the extremal triple has the form

$$\begin{aligned} i^* &= [0,\ldots,0,1,1] \\ y^* &= [0,\ldots,0,1,0] \\ z^* &= [0,\ldots,0,0,1] \end{aligned}$$

Hence the minimal value of α is attained when

$$u(1,1,m-2) = v(1,1,m-2) = \frac{1}{2+\sqrt{m-2}}$$

and so

$$p_m = \frac{\log 2}{\log(2+\sqrt{m-2})}\,.$$

We see that u and v are decreasing in m and so is p_m – see Table 9. For large m values of $u = v$ tends to 0 and we have

$$p_m \overset{m\to\infty}{\longrightarrow} 0 = p_\infty\,.$$

Table 9: The values of metric index of S^{**} for different dimensions

m	2	3	4	5	$\cdots$	20	$\cdots$	50	$\cdots$	100
p_m	1	0.6309	0.5644	0.5263		0.3784		0.3166		0.2798

4.10 Braun - Blanquet

As a consequence of remark 4 we get for Braun-Blanquet $1 - s_{20}$ distance the metric index equal to 1 on all $\mathbb{B}^m$. Indeed, if we pose $\mathbf{1} := [0,\ldots,0,0,1]$, $\mathbf{2} := [0,\ldots,0,1,0]$, $\mathbf{3} := [0,\ldots,0,1,1] \in \mathbb{B}^m$ for any $m \geq 2$ lie on a straight line:

$$d_{20}(\mathbf{1},\mathbf{2}) = 1, \qquad d_{20}(\mathbf{1},\mathbf{3}) = d_{20}(\mathbf{2},\mathbf{3}) = \frac{1}{2}.$$

4.11 Michael

The Michael dissimilarity measure

$$\frac{1}{2}\,(1 - s_{22}) = \frac{1}{2}\left(1 - \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}\right)$$

written in integer coordinates for $x, y, z \in \mathbb{B}^m$ gives

$$\begin{aligned}
\delta_{xy}(A,B,\ldots,H) &= \tfrac{1}{2}\left(1 - \tfrac{4\,((A+B)\,(G+H)-(C+D)\,(E+F))}{(A+B+G+H)^2+(C+D+E+F)^2}\right),\\
\delta_{xz}(A,B,\ldots,H) &= \tfrac{1}{2}\left(1 - \tfrac{4\,((A+C)\,(F+H)-(B+D)\,(E+G))}{(A+C+F+H)^2+(B+D+E+G)^2}\right),\\
\delta_{yz}(A,B,\ldots,H) &= \tfrac{1}{2}\left(1 - \tfrac{4\,((A+E)\,(D+H)-(B+F)\,(C+G))}{(A+E+D+H)^2+(B+F+C+G)^2}\right).
\end{aligned}$$

Using the usual analysis the minimal positive value of u is attained when

$\delta_{yz} = 1$: $\quad A = E = D = H = 0 \ \wedge\ B + F = C + G$
and
δ_{xy} minimal: $\quad \max(A+B)\,(G+H)\ \wedge (C + D = 0 \ \vee\ E + F = 0)$.

For $A = E = D = H = 0$ the result is the same as with Pearson dissimilarity: $s = u + v = 1$. If we put

$$(C = 0 \ \vee\ F = 0)\ \veebar\ (B = 0 \ \vee\ G = 0)$$

sum is the same.

Now let us put only $A = E = 0$ (for $D = H = 0$ result is the same) and to get the maximal value of δ_{yz} also $D + H = 1$ and $B + F = C + G = (m-1)/2$:

$$\max \delta_{yz} = \frac{1}{2}\left(1 + \frac{4\,\frac{m-1}{2}\,\frac{m-1}{2}}{1+(m-1)^2}\right) = \frac{1}{2}\left(1 + \frac{(m-1)^2}{1+(m-1)^2}\right) \overset{m\to\infty}{\longrightarrow} 1.$$

To get lower values of δ_{xy} and δ_{xz} we set $D = F = G = 0$ and in this case we have

$$u = v = \frac{1 - \frac{8\,(m-1)}{(m-1)^2+(m+1)^2}}{1 + \frac{(m-1)^2}{(m-1)^2+1}} \overset{m\to\infty}{\longrightarrow} \frac{1}{2}\,.$$

Since

$$\delta_{yz}(A,B,C,D,E,F,G,H) = \delta_{yz}(A+E, B+F, C+G, D+H, 0, 0, 0, 0) =$$

$$= \delta_{yz}(A+E, B+F, C+G, 0, 0, 0, 0, D+H)$$

and to get lower values of δ_{xy} and δ_{xz} $(F + H \neq 0 \;\;\wedge\;\; G + H \neq 0)$ we pose $D = E = F = G = 0$:

$$u(A,B,C,H) = \frac{1 - \frac{4\,(A+B)\,H}{C^2+(A+B+H)^2}}{1 - \frac{4\,(A\,H - B\,C)}{(B+C)^2+(A+H)^2}}\,, \qquad v(A,B,C,H) = \frac{1 - \frac{4\,(A+C)\,H}{B^2+(A+C+H)^2}}{1 - \frac{4\,(A\,H - B\,C)}{(B+C)^2+(A+H)^2}}\,.$$

From the local optimization procedure we see that

$$\min u(A,B,C,H) = u(A,A,A,2\,A)\,.$$

Since for any $n \in \mathbb{N}$

$$u(n\,A, n\,B, n\,C, n\,H) = u(A,B,C,H)$$

the minimal value of u obtained with the local optimization procedure is

$$\min u(A,B,C,H) = u(1,1,1,2) = 0.08497$$

and since

$$v(A,B,C,H) = u(A,C,B,H)$$

the minimal value for v is the same. So we get the upper bound $\overline{p}_m$ of the metric index equal to

$$\overline{p}_m = \overline{p}_\infty = -\frac{\log 2}{\log 0.08497} = 0.2811\,.$$

If we put $a := A/m$ and consider the function $u(a, a, a, 1-3\,a)$ on the interval $[0, 1/3]$, we get

$$u(a,a,a,1-3\,a) = \frac{\left(1-4\,a+8\,a^2\right)\,\left(1-10\,a+26\,a^2\right)}{1-10\,a+42\,a^2-64\,a^3+48\,a^4}$$

and

$$u'_a(a,a,a,1-3\,a) = \frac{4\,\left(-1+16\,a-128\,a^2+632\,a^3-1804\,a^4+2592\,a^5-1120\,a^6\right)}{\left(1-10\,a+42\,a^2-64\,a^3+48\,a^4\right)^2}\,.$$

If we solve the equation $u'_a = 0$ numerically, we get

$$a = 0.195109\,.$$

This gives the minimum of u and v equal to 0.08375 and the new upper bound of metric index equal to

$$\overline{p}_m = \overline{p}_\infty = -\frac{\log 2}{\log 0.08375} = 0.2795\,.$$

5 Discussion

We presented an approach to determination of the metric index for a given dissimilarity between binary vectors and applied it to some well known similarity measures. The results obtained offer new information which can be used when selecting similarity measures for applications.

For Ochiai $1 - s_{16}$, Sokal & Sneath $1 - un_5$, Pearson $(1-\phi)/2$ and Dice $1 - T_{1/2}$ dissimilarities we also confirmed the well known fact, that a square roots of these dissimilarities are distances.

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

6 References

Baulieu, F.B. (1989): A classification of presence/absence based dissimilarity coefficients. *Journal of Classification,* 6, p.233-246.

Batagelj, V., Bren, M. (1995): Comparing Resemblance Measures. *Journal of Classification,* Vol. 12, No. 1, p.73-90.

Batagelj, V., Bren, M. (1996): Determining the Distance Index. *Ordinal and Symbolic Data Analysis: Proceedings of the International Conference on Ordinal and Symbolic Data Analysis – OSDA'95* ed. E. Diday, Y. Lechevallier, O. Opitz. Springer - Verlag Berlin Heidelberg New York 1996, p. 238-251.

Batagelj, V., Bren, M. (1997): Determining the Distance Index, II. *Data Science, Classification and Related Methods: Proceedings of the International Federation of Classification Societies – IFCS96.* In print, Springer Verlag, Tokyo.

Gower, J.C., and Legendre, P. (1986): Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification,* 3, p.5-48.

Hubálek, Z. (1982): Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Review,* 57, p.669-689.

Joly, S., Le Calvé, G. (1986): Etude des puissances d'une distance. *Statistique et Analyse des Données,* 11/3, p.30-50.

Joly, S., Le Calvé, G. (1994): Similarity functions. In *Classification and Dissimilarity Analysis,* p.67-87, editor: Van Cutsem, B., Lecture Notes in Statistics, Springer Verlag: New York.

Lerman, I.C. (1971): Indice de similarité et préordonnance associée. *Ordres totaux finis. Travaux du séminaire sur les ordres totaux finis,* Aix-en-Provence, 1967; Paris: Mouton.

Van Cutsem, B., ed.(1994): *Classification and Dissimilarity Analysis,* Lecture Notes in Statistics, Springer Verlag: New York.