Journal of Classification 33:141-148 (2016)
DOI: 10.1007/s00357-016-9200-z

# Inequalities Between Similarities for Numerical Data

Matthijs J. Warrens

University of Groningen, The Netherlands

**Abstract:** Similarity measures are entities that can be used to quantify the similarity between two vectors with real numbers. We present inequalities between seven well known similarities. The inequalities are valid if the vectors contain non-negative real numbers.

**Keywords**: Bray-Curtis similarity; Ruzicka similarity; Ellenberg similarity; Gleason similarity.

## 1. Introduction

Similarity measures and distances are important tools in pattern classification, clustering and information retrieval problems (Gower and Legendre 1986; Zuur, Ieno, and Smith 2007; Lesot, Rifqi, and Benhadda 2009). A similarity can be used to quantify the strength of the relationship between two vectors with numerical data. Popular choices are the Bray-Curtis similarity, the Ellenberg similarity and the Gleason similarity (Deza and Deza 2013). A similarity measure or a distance has to be considered in the context of the descriptive statistical study of which it is a part. The choice of a measure depends on the nature of the data and the type of analysis, for instance, cluster analysis or multidimensional scaling.

Corresponding Author's Address: Matthijs J. Warrens, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands, e-mail: m.j.warrens@rug.nl.

Published online: 13 May 2016

Since the choice of a proper similarity measure or a distance is often not an exact science, various authors have investigated which measure may be appropriate in a certain data-analytic context (Campbell 1978; Huhta 1979; Wolda 1981; Gower and Legendre 1986; Baulieu 1989; Batagelj and Bren 1995; Fechner and Schneider 2004; Albatineh, Niewiadomska-Bugaj, and Mihalko 2006; Cha 2007). Comparisons of similarities may not be conclusive, but they often provide some insight into the behavior of the similarities. A type of study that may enhance the understanding of similarities and how they are related is an analytic comparison (Warrens 2008b). Deza and Deza (2013) present a list of similarity measures that are used in practice. In this note we present inequalities between these similarities that hold for non-negative real data. Understanding how similarities are related may help a researcher decide which similarity to choose.

The note is organized as follows. Seven similarities of interest are introduced in the next section. It is also shown in Section 2 which similarities coincide if the vectors are restricted to the values 0 and 1. Inequalities between the seven similarities are presented in Section 3. The inequalities are valid if the vectors consist of non-negative real numbers. This is, for instance, the case if the vectors are species abundance distributions or probability density functions. The latter are popular functions for summarizing various types of pattern data (Cha 2007). Finally, Section 4 contains a conclusion.

## 2. Similarities

In this section, we briefly discuss seven well known similarities from the classification literature (Deza and Deza 2013, p. 292–294). The values of the similarities lie between 0 and 1, where 1 indicates perfect similarity and 0 indicates lack of similarity. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be non-zero vectors with non-negative real numbers. The notation $\sum x_i$ is short for the summation $\sum_{i=1}^{n} x_i$.

The intersection similarity (Cha 2007) is defined as

$$S_1 = \frac{\sum \min(x_i, y_i)}{\min(\sum x_i, \sum y_i)}.$$

Similarity $S_1$ is the complement of the intersection distance in Deza and Deza (2013, p. 294). The Kulczyński 2 similarity is given by

$$S_2 = \frac{1}{2}\left[\frac{\sum \min(x_i, y_i)}{\sum x_i} + \frac{\sum \min(x_i, y_i)}{\sum y_i}\right].$$

The Kulczyński 1 similarity (Deza and Deza 2013, p. 294) is not considered here because it has no upper bound. The Bray-Curtis similarity (Bray and Curtis 1957) is defined as

$$S_3 = \frac{2\sum \min(x_i, y_i)}{\sum (x_i + y_i)}.$$

The Motyka similarity (Deza and Deza 2013) is half $S_3$. The Motyka similarity is not considered here because its upper bound is $\frac{1}{2}$ instead of 1 if $x = y$. The Roberts similarity is given by

$$S_4 = \frac{\sum (x_i + y_i) \frac{\min(x_i, y_i)}{\max(x_i, y_i)}}{\sum (x_i + y_i)},$$

and the Ruzicka similarity is defined as

$$S_5 = \frac{\sum \min(x_i, y_i)}{\sum \max(x_i, y_i)}.$$

For similarities $S_1$ to $S_5$ we have $S = 1$ if $x = y$, and $S = 0$ if $x_i y_i = 0$ for all $i$.

Next, define

$$1_{x_i y_i \neq 0} = \begin{cases} 1, & \text{if } x_i y_i \neq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Using this indicator function the Ellenberg similarity is defined as

$$S_6 = \frac{\sum (x_i + y_i) 1_{x_i y_i \neq 0}}{\sum (x_i + y_i)(1 + 1_{x_i y_i = 0})},$$

while the Gleason similarity is given by

$$S_7 = \frac{\sum (x_i + y_i) 1_{x_i y_i \neq 0}}{\sum (x_i + y_i)}.$$

We have $S_6 = S_7 = 1$ if $x_i y_i \neq 0$ for all $i$, and $S_6 = S_7 = 0$ if $x_i y_i = 0$ for all $i$.

The seven similarities for numerical data $S_1$ to $S_7$ extend several well known similarities for two binary vectors (Albatineh, Niewiadomska-Bugaj, and Mihalko 2006; Warrens 2008a, 2009). The vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ might be restricted to the values 0 and 1, which may be regarded as formal scores for the states $-$ (absence) and $+$ (presence) of two binary objects. For example, the objects may be individuals that may or may not possess certain traits. Furthermore, the objects could be regions in which certain species do or do not occur.

The information in two binary vectors can be summarized by four dependent quantities: the number of attributes with $+$ on both objects ($A$),

the number of attributes with $+$ on one object and $-$ on the other object ($B$ and $C$), and the number of attributes with $-$ on both objects ($D$). It holds that $A + B + C + D = n$. Using the numbers $A$, $B$, $C$ and $D$ we may construct the following fourfold table of co-occurrence of two binary objects:

| | Object 1 | |
|---|---|---|
| Object 2 | $+$ | $-$ |
| $+$ | $A$ | $B$ |
| $-$ | $C$ | $D$ |

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are binary vectors similarity $S_1$ reduces to the Simpson similarity $A/(A+\min(B,C))$. Furthermore, similarities $S_3$, $S_4$ and $S_7$ coincide if $\boldsymbol{x}$ and $\boldsymbol{y}$ are restricted to the values 0 and 1. In this case the three similarities are equal to the Dice or Sørensen similarity $2A/(2A + B + C)$ (Deza and Deza 2013). Moreover, both $S_5$ and $S_6$ reduce to the Jaccard similarity $A/(A + B + C)$ if $x$ and $y$ are binary vectors (Warrens 2008a, 2009).

## 3. Inequalities

In this section, we present inequalities between the seven similarities from Section 2. We begin with inequalities between $S_1$ to $S_5$. We have the ordering $S_1 \geq S_2 \geq S_3 \geq S_4 \geq S_5$. The four inequalities are proved in Lemma 1, 2 and 3.

**Lemma 1.** $S_1 \geq S_2 \geq S_3$.

*Proof.* Let $a = \sum \min(x_i, y_i)/\sum x_i$ and $b = \sum \min(x_i, y_i)/\sum y_i$. Since $S_1 = \max(a, b)$ and $S_2 = (a + b)/2$, it follows that $S_1 \geq S_2$. Furthermore, since $S_2 = (a + b)/2$ and $S_3 = 2/(a^{-1} + b^{-1})$, the inequality $S_2 \geq S_3$ follows from the arithmetic-harmonic means inequality.
■

**Lemma 2.** $S_3 \geq S_4$.

*Proof.* Inequality $S_3 \geq S_4$ if and only if

$$2\sum \min(x_i, y_i) \geq \sum (x_i + y_i)\frac{\min(x_i, y_i)}{\max(x_i, y_i)}. \tag{1}$$

Let $a_i = \min(x_i, y_i)$ and $b_i = \max(x_i, y_i)$. We have $b_i \geq a_i$. Adding $b_i$ to both sides of this inequality we obtain $2b_i \geq a_i + b_i$. Multiplying both sides of the latter inequality by $a_i/b_i$ we obtain $2a_i \geq a_i(a_i + b_i)/b_i$. Summing the latter inequality over all $i$ we obtain (1), and thus $S_3 \geq S_4$.
■

**Lemma 3.** $S_4 \geq S_5$.

*Proof.* Let $a_i = \min(x_i, y_i)$ and $b_i = \max(x_i, y_i)$. We have $S_4 \geq S_5$ if and only if

$$\frac{\sum(a_i + b_i)\frac{a_i}{b_i}}{\sum(a_i + b_i)} \geq \frac{\sum a_i}{\sum b_i}. \tag{2}$$

Cross multiplying the fractions of (2) we obtain

$$\sum \frac{a_i^2}{b_i} \sum b_i + \sum a_i \sum b_i \geq \left(\sum a_i\right)^2 + \sum a_i \sum b_i,$$

or equivalently

$$\sum \frac{a_i^2}{b_i} \sum b_i \geq \left(\sum a_i\right)^2. \tag{3}$$

For real numbers $c_1, \ldots, c_n$ and $d_1, \ldots, d_n$ the Cauchy-Schwarz inequality is given by

$$\sum c_i^2 \sum d_i^2 \geq \left(\sum c_i d_i\right)^2. \tag{4}$$

Using $c_i = a_i/\sqrt{b_i}$ and $d_i = \sqrt{b_i}$ in (4) we obtain inequality (3), and thus $S_4 \geq S_5$.
■

Next, we consider the similarities $S_6$ and $S_7$. The inequality $S_6 \leq S_7$ follows from the inequality

$$\sum(x_i + y_i)(1 + 1_{x_i y_i = 0}) \geq \sum(x_i + y_i).$$

To show how $S_6$ and $S_7$ are related to the other similarities, we use the following lemma.

**Lemma 4.**

$$\sum(x_i + y_i)1_{x_i y_i \neq 0} \geq 2\sum \min(x_i, y_i). \tag{5}$$

*Proof.* If $x_i y_i \neq 0$, both $x_i$ and $y_i$ are positive, and we have $x_i + y_i \geq 2\min(x_i, y_i)$. Furthermore, if $x_i y_i = 0$, we have $\min(x_i, y_i) = 0$. Hence, inequality (5) is valid.
■

The inequality $S_7 \geq S_3$ follows from Lemma 4. Combining $S_7 \geq S_3$ with the previous results we have $S_7 \geq S_3 \geq S_4 \geq S_5$.

**Lemma 5.** $S_6 \geq S_5$.

*Proof.* Let

$$a = \sum (x_i + y_i) 1_{x_i y_i \neq 0},$$
$$b = \sum \min(x_i, y_i),$$
$$c = \sum \max(x_i, y_i).$$

Using the quantities $a$, $b$ and $c$, together with the identity

$$1 = 1_{x_i y_i = 0} + 1_{x_i y_i \neq 0},$$

the inequality $S_6 \geq S_5$ is equal to

$$\frac{a}{2(b+c) - a} \geq \frac{b}{c}. \tag{6}$$

Cross multiplying the fractions in (6) we obtain

$$a(b+c) \geq 2b(b+c). \tag{7}$$

Since $b + c > 0$, dividing (7) by $b + c$ yields $a \geq 2b$, which is inequality (5). The assertion then follows from Lemma 4.
■

## 4. Conclusion

In this note, we presented inequalities between seven well known similarities for two numerical vectors (Deza and Deza 2013). The inequalities are valid if the vectors contain non-negative real numbers. Examples of non-negative vectors are species abundance distributions and probability density functions. The latter are popular functions for representing various types of pattern data (Cha 2007).

The results are summarized in Theorem 6 below. The symbol $\geq$ indicates that the row similarity dominates the column similarity, while $\leq$ indicates that the row similarity never exceeds the column similarity. The remaining pairwise relations between the similarities are marked with the symbol $-$.

**Theorem 6.** *The following inequalities hold between the similarities:*

| | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|
| $S_1$ | $\geq$ | $\geq$ | $\geq$ | $\geq$ | $-$ | $-$ |
| $S_2$ | | $\geq$ | $\geq$ | $\geq$ | $-$ | $-$ |
| $S_3$ | | | $\geq$ | $\geq$ | $-$ | $\leq$ |
| $S_4$ | | | | $\geq$ | $-$ | $\leq$ |
| $S_5$ | | | | | $\leq$ | $\leq$ |
| $S_6$ | | | | | | $\leq$ |

For similarities $S_1$ to $S_5$ we always have the ordering $S_1 \geq S_2 \geq S_3 \geq S_4 \geq S_5$. It thus appears that similarities $S_1$ to $S_5$ are measuring the same concept of similarity but to a different extent.

For pairs of similarities marked with the symbol $-$ in Theorem 6, the ordering of the similarities depends on the data. For example, if $x = (4, 3, 2)$ and $y = (0, 2, 5)$ we have $S_1 = .571$, $S_6 = .600$ and $S_7 = .750$. Thus, for these data we have the ordering $S_7 > S_6 > S_1 > S_2 > S_3 > S_4 > S_5$. Furthermore, if $x = (9, 0, 1)$ and $y = (0, 6, 1)$ we have $S_1 = .143$, $S_2 = .121$, $S_3 = .118$, $S_4 = .118$, $S_5 = .063$, $S_6 = .063$ and $S_7 = .118$. Hence, for these data we have the ordering $S_1 > S_2 > S_7 = S_3 = S_4 > S_6 = S_5$.

## References


ALBATINEH, A.N., NIEWIADOMSKA-BUGAJ, M., and MIHALKO, D. (2006), "On Similarity Indices and Correction for Chance Agreement," *Journal of Classification, 23,* 301–313.

BATAGELJ, V., and BREN, M. (1995), "Comparing Resemblance Measures," *Journal of Classification, 12,* 73–90.

BAULIEU, F.B. (1989), "A Classification of Presence/Absence Based Dissimilarity Coefficients," *Journal of Classification, 6,* 233–246.

BRAY, J.R., and CURTIS, J.T. (1957), "An Ordination of the Upland Forest Communities of Southern Wisconsin," *Ecological Monographs, 27,* 325–349.

CAMPBELL, B.M. (1978), "Similarity Coefficients for Classifying Relevés," *Vegetatio, 37,* 101–109.

CHA, S. (2007), "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *International Journal of Mathematical Models and Methods in Applied Sciences, 1,* 300–307.

DEZA, M.M., and DEZA, E. (2013), *Encyclopedia of Distances,* Berlin: Springer-Verlag.

FECHNER, U., and SCHNEIDER, G. (2004), "Evaluation of Distance Metrics for Ligand-based Similarity Searching," *ChemBioChem, 5,* 538–540.

GOWER, J.C., and LEGENDRE, P. (1986), "Metric and Euclidean Properties of Dissimilarity Coefficients," *Journal of Classification, 3,* 5–48.

HUHTA, V. (1979), "Evaluation of Different Similarity Indices as Measures of Succession in Arthropod Communities of the Forest Floor after Clear-cutting," *Oecologia, 41,* 11–23.

LESOT, M.-J., RIFQI, M., and BENHADDA, H. (2009), "Similarity Measures for Binary and Numerical Data: A Survey," *International Journal of Knowledge Engineering and Soft Data Paradigms, 1,* 63–84.

WARRENS, M.J. (2008a), "On the Indeterminacy of Resemblance Measures for Binary (Presence/Absence) Data," *Journal of Classification, 25,* 125–136.

WARRENS, M.J. (2008b), "Bounds of Resemblance Measures for Binary (Presence/Absence) Variables," *Journal of Classification, 25,* 195–208.

WARRENS, M.J. (2009), "$k$-Adic Similarity Coefficients for Binary (Presence/Absence) Data," *Journal of Classification, 26,* 227–245.

WOLDA, H. (1981), "Similarity Indices, Sample Size and Diversity," *Oecologia, 50,* 296–302.

ZUUR, A.F., IENO, E.N., and SMITH, G.M. (2007), *Analysing Ecological Data,* New York: Springer.