Journal of Classification 25:195-208 (2008)
DOI: 10.1007/s00357-008-9024-6

Bounds of Resemblance Measures for Binary (Presence/Absence) Variables

Matthijs J. Warrens

Leiden University, The Netherlands

Abstract: Bounds of association coefficients for binary variables are derived using the arithmetic-geometric-harmonic mean inequality. More precisely, it is shown which presence/absence coefficients are bounds with respect to each other. Using the new bounds it is investigated whether a coefficient is in general closer to either its upper or its lower bound.

Keywords: Association coefficients; Similarity coefficients; 2×2 table; Minimum value; Harmonic mean; Geometric mean; Arithmetic mean; Maximum value.

1. Introduction

In data analysis an important role is played by association coefficients. A coefficient is a measure of similarity or resemblance of two entities or variables. An example is Pearson's product-moment correlation for two continuous variables. Coefficients for other types of variables can be found in, for example, Goodman and Kruskal (1954), Hubálek (1982), and Gower and Legendre (1986). In this paper we focus on measures for binary variables. These presence/absence coefficients are usually defined using the four dependent quantities a, b, c, and d presented in Table 1. Quantities a, b, c, and d may be probabilities as well as counts. Probabilities are used here for notational convenience.

The author would like to thank two anonymous reviewers for their helpful comments and valuable suggestions on earlier versions of this article.

Author's Address: Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands, e-mail: warrens@fsw.leidenuniv.nl.

Published online 19 December 2008

Table 1. Bivariate proportions table for binary variables.

	Variable two		
Variable one	Value 1	Value 2	Total
Value 1	a	b	p_1
Value 2	c	d	q_1
Total	p_2	q_2	1

In choosing a coefficient, each measure has to be considered in the context of the data-analytic study of which it is a part (Gower and Legendre 1986, sec. 5). Because there are so many resemblance measures for binary variables to choose from, it is important that the different coefficients and their properties are better understood. For example, Gower (1986), Fichet (1986), Gower and Legendre (1986), and Bren and Batagelj (2006) studied metric and Euclidean properties; Batagelj and Bren (1995) discussed results on (ordinal) equivalence relations over coefficients; Baulieu (1989, 1997) presented classifications of presence/absence coefficients using certain desirable properties in different axiomatic frameworks; Janson and Vegelius (1981) and Gower and Legendre (1986) investigated Gramian properties and positive semidefiniteness of coefficient matrices; finally, Boyce and Ellison (2001) studied presence/absence coefficients in the context of fuzzy set ordination.

In this paper we study bounds of measures for two binary variables. It is well-known that many presence/absence coefficients are bounded by 0 and 1 or -1 and 1. More importantly, coefficients can be bounds with respect to each other. A variety of insights can be obtained from deriving which coefficient is a lower or an upper bound of another coefficient. For example, a relatively large number of coefficients defined on the same quantities can be bounds with respect to each other; in this case it is likely that these coefficients, apart from perhaps the smallest and largest coefficient, reflect the association of two variables in a similar way, but to a different extent: some have lower/higher values than others. For example, it holds that

$$\begin{aligned} 0 \quad &\leq \quad \frac{a^2}{(a+b)(a+c)} \quad \leq \quad \frac{a}{a+b+c} \quad \leq \quad \frac{a}{a+\max(b,c)} \\ &\leq \quad \frac{2a}{2a+b+c} \quad \leq \quad \frac{a}{\sqrt{(a+b)(a+c)}} \\ &\leq \quad \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right) \quad \leq \quad \frac{a}{a+\min(b,c)} \quad \leq 1 \end{aligned}$$

(Proposition 1, Section 3). Coefficients with the same quantities in the numerator and denominator, that are bounded, and are close to each other in

the ordering, are (likely to be) more similar. Thus, results on bounds provide means of classifying various measures. Also, knowing which coefficients are similar (in terms of the actual values) provides insight into the stability of a given algorithm: for which coefficients will a data analysis provide the same or similar results?

The paper is organized as follows. A variety of resemblance measures for binary variables are functions of two real variables. These functions are the minimum, the harmonic, geometric and arithmetic means (Pythagorean means), and the maximum. Some properties of the Pythagorean means are the main tools for studying bounds in this paper. The tools are presented in the next section. In Sections 3 and 4 we present some applications of the theorems from Section 2. In Section 3 we focus on measures that do not include the probability d (representing negative matches). Coefficients that use the covariance $(ad - bc)$ of two binary variables in the numerator are investigated in Section 4. Coefficients that have been proposed as chance-corrected measures are studied in Section 5. Section 6 contains the discussion. Some additional inequalities for association coefficients for 2×2 tables are presented (without proof) in the appendix.

Many presence/absence coefficients are fractions and can be defined using probabilities a, b, c, and d only. It may occur that for some combinations of a, b, c, and d, the value of the coefficient is indeterminate (Batagelj and Bren 1995; Warrens, 2008). For simplicity, it is assumed throughout the paper that the value of a coefficient is defined. Furthermore, the expression "if and only if" is sometimes abbreviated as "iff".

2. Pythagorean Means

Let x_1 and x_2 be positive real numbers. The harmonic, geometric and arithmetic mean of x_1 and x_2, denoted by $H(x_1, x_2)$, $G(x_1, x_2)$, $A(x_1, x_2)$, respectively, are defined as

$$H(x_1, x_2) = \frac{2x_1x_2}{x_1 + x_2}, \quad G(x_1, x_2) = \sqrt{x_1x_2}, \quad A(x_1, x_2) = \frac{x_1 + x_2}{2}.$$

A variety of presence/absence coefficients can be expressed as the minimum, harmonic mean, geometric mean, arithmetic mean, or maximum of two positive quantities. We consider two results for these functions. First it is shown how the five functions are related. Theorem 1 is a special case of the generalized mean inequality (Bullen 2003, chap. 3; Abramowitz and Stegun 1972, p. 10).

Theorem 1. $\min(x_1, x_2) \leq H(x_1, x_2) \leq G(x_1, x_2) \leq A(x_1, x_2) \leq \max(x_1, x_2)$ *with equality iff* $x_1 = x_2$.

By Theorem 1, the five functions can be ordered and each Pythagorean mean has two boundaries: $\min(x_1, x_2)$ and $G(x_1, x_2)$ for $H(x_1, x_2)$, $H(x_1, x_2)$ and $A(x_1, x_2)$ for $G(x_1, x_2)$, and $G(x_1, x_2)$ and $\max(x_1, x_2)$ for $A(x_1, x_2)$. We may inspect whether the value of a mean is in general closer to its upper or its lower bound. For each pair of two adjacent functions we have the differences

$$\begin{aligned} H(x_1, x_2) - \min(x_1, x_2) &= \frac{\min(x_1, x_2)|x_1 - x_2|}{x_1 + x_2} \\ G(x_1, x_2) - H(x_1, x_2) &= \frac{\sqrt{x_1 x_2}(\sqrt{x_1} - \sqrt{x_2})^2}{x_1 + x_2} \\ A(x_1, x_2) - G(x_1, x_2) &= \frac{(\sqrt{x_1} - \sqrt{x_2})^2}{2} \\ \max(x_1, x_2) - A(x_1, x_2) &= \frac{|x_1 - x_2|}{2}. \end{aligned}$$

Some of these differences are ordered in the following way

Theorem 2.

$$\begin{aligned} G(x_1, x_2) - H(x_1, x_2) &\overset{(i)}{\leq} A(x_1, x_2) - G(x_1, x_2) \\ &\overset{(ii)}{\leq} \max(x_1, x_2) - A(x_1, x_2) \end{aligned}$$

and

$$A(x_1, x_2) - H(x_1, x_2) \overset{(iii)}{\leq} \max(x_1, x_2) - A(x_1, x_2)$$

with equality iff $x_1 = x_2$.

Proof (i): By assumption $x_1 \neq x_2$

$$\begin{aligned} \sqrt{x_1} - \sqrt{x_2} &\neq 0 \\ (\sqrt{x_1} - \sqrt{x_2})^4 &> 0 \\ (x_1 + x_2)^2 + 4x_1 x_2 &> 4\sqrt{x_1 x_2}(x_1 + x_2) \\ \frac{(x_1 + x_2)^2 + 4x_1 x_2}{2(x_1 + x_2)} &> 2\sqrt{x_1 x_2} \\ \frac{x_1 + x_2}{2} + \frac{2x_1 x_2}{x_1 + x_2} &> 2\sqrt{x_1 x_2} \\ \frac{x_1 + x_2}{2} - \sqrt{x_1 x_2} &> \sqrt{x_1 x_2} - \frac{2x_1 x_2}{x_1 + x_2}. \end{aligned}$$

Proof 1 (ii): Assume $x_1 > x_2$. Then $x_1 - x_2 > (\sqrt{x_1} - \sqrt{x_2})^2$ iff $2\sqrt{x_1 x_2} > 2x_2$.

Proof 2 (ii) *and proof* (iii): Both inequalities may be deduced from equality

$$\max(x_1, x_2) - A(x_1, x_2) = A(x_1, x_2) - \min(x_1, x_2) = \frac{|x_1 - x_2|}{2}.$$

■

Applications of Theorems 1 and 2 are presented in Sections 3 and 4.

3. Coefficients That Exclude Negative Matches

Sokal and Sneath (1963) (among others) make a distinction between coefficients that do or do not include the quantity d. If a binary variable is a coding of the presence or absence of a list of attributes or features, then d (usually) reflects the number of negative matches. In the field of numerical taxonomy quantity d is generally felt not to contribute to similarity. In other words, presence/absence is viewed as an ordinal variable. In this case presence is 'more' in a sense than absence. If the variables are nominal, coefficients for which the quantities a and d are equally weighted are appropriate.

In this section we consider seven coefficients that do not include the negative matches. Following Sokal and Sneath (1963), the convention is adopted of calling a coefficient by its originator or the first we know to propose it. The exception to this rule is the Phi coefficient in Section 4. The coefficients are

$$S_{\text{Sorg}} = \frac{a^2}{p_1 p_2} \qquad \text{(Sorgenfrei 1958)}$$

$$S_{\text{Jac}} = \frac{a}{p_1 + p_2 - a} \qquad \text{(Jaccard 1912)}$$

$$S_{\text{BB}} = \frac{a}{\max(p_1, p_2)} \qquad \text{(Braun-Blanquet 1932)}$$

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2} \qquad \text{(Gleason 1920; Dice 1945)}$$

$$S_{\text{Och}} = \frac{a}{\sqrt{p_1 p_2}} \qquad \text{(Ochiai 1957)}$$

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) \qquad \text{(Kulczyński 1927)}$$

$$S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)} \qquad \text{(Simpson 1943).}$$

The coefficients are related by

Proposition 1. $0 \leq S_{\text{Sorg}} \overset{(i)}{\leq} S_{\text{Jac}} \overset{(ii)}{\leq} S_{\text{BB}} \leq S_{\text{Gleas}} \leq S_{\text{Och}} \leq S_{\text{Kul}} \leq S_{\text{Sim}} \leq 1.$

Proof (i): $S_{\text{Sorg}} \leq S_{\text{Jac}}$ iff $p_1p_2 - a(p_1+p_2) + a^2 \geq 0$ iff $(p_1-a)(p_2-a) \geq 0$.
Proof (ii): $S_{\text{Jac}} \leq S_{\text{BB}}$ iff $p_1 + p_2 \geq \max(p_1, p_2) + a$ iff $\min(p_1, p_2) \geq a$.
Since $S_{\text{BB}} = \min(x_1, x_2)$, $S_{\text{Gleas}} = H(x_1, x_2)$, $S_{\text{Och}} = G(x_1, x_2)$, $S_{\text{Kul}} = A(x_1, x_2)$, and $S_{\text{Sim}} = \max(x_1, x_2)$, where

$$x_1 = \frac{a}{p_1} \quad \text{and} \quad x_2 = \frac{a}{p_2}$$

the remaining inequalities follow from application of Theorem 1.
■

The ordering of the seven coefficients for ordinal variables is established in Proposition 1. Note that this is the inequality used for illustrative purposes in Section 1.

Next we may inspect whether the value of a certain coefficient is in general closer to its upper or its lower bound. We have the differences

$$\begin{aligned}
S_{\text{Och}} - S_{\text{Gleas}} &= \frac{a\sqrt{p_1p_2}(\sqrt{p_1} - \sqrt{p_2})^2}{p_1 + p_2} \\
S_{\text{Kul}} - S_{\text{Gleas}} &= \frac{a(p_1 - p_2)^2}{2p_1p_2(p_1 + p_2)} \\
S_{\text{Kul}} - S_{\text{Och}} &= \frac{a(\sqrt{p_1} - \sqrt{p_2})^2}{2p_1p_2} \\
S_{\text{Sim}} - S_{\text{Kul}} = S_{\text{Kul}} - S_{\text{BB}} &= \frac{a|p_1 - p_2|}{2p_1p_2}.
\end{aligned}$$

The value of coefficient S_{Och} is closer to the value of measure S_{Gleas} than to the value of index S_{Kul}. The value of coefficient S_{Kul} is closer to measure S_{Och} and S_{Gleas} than to the value of coefficient S_{Sim}.

Proposition 2. $S_{\text{Och}} - S_{\text{Gleas}} \leq S_{\text{Kul}} - S_{\text{Och}} \leq S_{\text{Sim}} - S_{\text{Kul}}$ *and* $S_{\text{Kul}} - S_{\text{Gleas}} \leq S_{\text{Sim}} - S_{\text{Kul}}$.

The claim follows from using the definitions of these coefficients in the proof of Proposition 1 together with Theorem 2.

4. Coefficients with the Covariance in the Numerator

It may be required that the value of a similarity coefficient is zero in the absence of association between two variables. The covariance between two binary variables is given by $(ad - bc)$. Coefficients with quantity $(ad - bc)$ in the numerator have zero value if the two variables are statistically independent. We first consider coefficients

$$S_{\text{Cohen}} = \frac{2(ad-bc)}{p_1q_2 + p_2q_1} \qquad \text{(Kappa; Cohen 1960)}$$

$$S_{\text{Phi}} = \frac{ad-bc}{\sqrt{p_1p_2q_1q_2}} \qquad \text{(Phi coefficient; Yule 1912)}$$

$$S_{\text{MP}} = \frac{2(ad-bc)}{p_1q_1 + p_2q_2} \qquad \text{(Maxwell and Pilliner 1968)}$$

$$S_{\text{Fleiss}} = \frac{(ad-bc)(p_1q_1 + p_2q_2)}{2p_1q_2p_2q_1} \qquad \text{(Fleiss 1975, p. 656)}$$

$$S_{\text{Loe}} = \frac{ad-bc}{\min(p_1q_2, p_2q_1)} \qquad \text{(Loevinger 1948).}$$

Propositions 3 and 4 are applications of Theorem 1. Coefficients S_{Cohen}, S_{Phi}, and S_{Loe} are related by

Proposition 3. $0 \leq |S_{\text{Cohen}}| \leq |S_{\text{Phi}}| \leq |S_{\text{Loe}}| \leq 1$.

Proof: $S_{\text{Cohen}} = H(x_1, x_2)$, $S_{\text{Phi}} = G(x_1, x_2)$, and $S_{\text{Loe}} = \max(x_1, x_2)$, where

$$x_1 = \frac{ad-bc}{p_1q_2} \quad \text{and} \quad x_2 = \frac{ad-bc}{p_2q_1}.$$

■

Coefficients S_{MP}, S_{Phi}, and S_{Fleiss} are related by

Proposition 4. $0 \leq |S_{\text{MP}}| \leq |S_{\text{Phi}}| \leq |S_{\text{Fleiss}}| \leq 1$.

Proof: As noted by Fleiss (1975, p. 656), we have $S_{\text{MP}} = H(x_1, x_2)$, $S_{\text{Phi}} = G(x_1, x_2)$, and $S_{\text{Fleiss}} = A(x_1, x_2)$, where

$$x_1 = \frac{ad-bc}{p_1q_1} \quad \text{and} \quad x_2 = \frac{ad-bc}{p_2q_2}.$$

■

The absolute value of S_{Phi} is in general closer to the absolute value of coefficient S_{Cohen} than to value of coefficient S_{Loe}. Furthermore, the absolute value of S_{Phi} is in general closer to the absolute value of coefficient S_{MP} than to value of coefficient S_{Fleiss}.

Proposition 5. $|S_{\text{Phi}}| - |S_{\text{Cohen}}| \leq |S_{\text{Loe}}| - |S_{\text{Phi}}|$ *and* $|S_{\text{Phi}}| - |S_{\text{MP}}| \leq |S_{\text{Fleiss}}| - |S_{\text{Phi}}|$.

The claim follows from using the definitions of the coefficients in the proofs of Propositions 3 and 4, together with Theorem 2.

In addition to coefficients S_{Cohen}, S_{Phi}, S_{MP}, S_{Fleiss}, and S_{Loe} some other coefficients are considered in this section as well. The absolute values of coefficients

$$S_{\text{Bau}} = \frac{4(ad-bc)}{(a+b+c+d)^2} \qquad \text{(Baulieu 1989, p. 244)}$$

$$\text{and} \quad S_{\text{Yule1}} = \frac{ad-bc}{ad+bc} \qquad \text{(Yule 1900)}$$

are, respectively, lower and upper bounds for the absolute values of coefficients

$$S_{\text{Mich}} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2} \qquad \text{(Michael 1920)}$$

$$S_{\text{Yule2}} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \qquad \text{(Yule 1912)}$$

and S_{Cohen}, S_{Phi}, S_{Loe}, and S_{Fleiss}.

Proposition 6. *$|S_{\text{Bau}}|$ is a lower bound of $|S_{\text{Mich}}|$, $|S_{\text{Phi}}|$, $|S_{\text{Loe}}|$, $|S_{\text{Fleiss}}|$, and $|S_{\text{Yule1}}|$.*

Proof: We have $|S_{\text{Bau}}| \leq |S_{\text{Mich}}|$ iff

$$1 = (a+d+b+c)^2 = (a+d)^2+(b+c)^2+2(a+d)(b+c) \geq (a+d)^2+(b+c)^2.$$

Inequality $|S_{\text{Bau}}| \leq |S_{\text{Phi}}|$ holds iff $p_1p_2q_1q_2 \leq 1/16$. Since $p_1 + q_1 = p_2 + q_2 = 1$, we have $\max(p_1q_1) = 1/4$ and $\max(p_2q_2) = 1/4$, from which it follows that $\max(p_1p_2q_1q_2) = 1/16$.

Inequalities $|S_{\text{Bau}}| \leq |S_{\text{Loe}}|$ and $|S_{\text{Bau}}| \leq |S_{\text{Fleiss}}|$ follow from inequality $|S_{\text{Bau}}| \leq |S_{\text{Phi}}|$ and Propositions 3 and 4. Inequality $|S_{\text{Bau}}| \leq |S_{\text{Yule1}}|$ follows from inequality $|S_{\text{Bau}}| \leq |S_{\text{Phi}}|$ and Proposition 7.
■

Proposition 7. *$|S_{\text{Yule1}}|$ is an upper bound of $|S_{\text{Yule2}}|$, $|S_{\text{Mich}}|$, $|S_{\text{Phi}}|$, $|S_{\text{Cohen}}|$, and $|S_{\text{MP}}|$.*

Proof: Inequality $|S_{\text{Yule2}}| \leq |S_{\text{Yule1}}|$ follows from

$$\frac{ad-bc}{ad+bc} \geq \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \quad \text{for} \quad ad \geq bc$$

$$\text{and} \qquad \frac{ad-bc}{ad+bc} \leq \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \quad \text{for} \quad ad \leq bc.$$

Inequality $|S_{\text{Mich}}| \leq |S_{\text{Yule1}}|$ holds iff

$$(a+d)^2 + (b+c)^2 \geq 4(ad+bc)$$
$$a^2 + d^2 - 2ad + b^2 + c^2 - 2bc \geq 0$$
$$(a-d)^2 + (b-c)^2 \geq 0.$$

Inequality $|S_{\text{Phi}}| \leq |S_{\text{Yule1}}|$ holds iff $p_1p_2q_1q_2 \geq (ad+bc)^2$. The latter inequality is true since

$$p_1q_1 = (a+b)(c+d) \geq ad+bc$$
$$\text{and} \qquad p_2q_2 = (a+c)(b+d) \geq ad+bc.$$

Inequalities $|S_{\text{Cohen}}| \leq |S_{\text{Yule1}}|$ and $|S_{\text{MP}}| \leq |S_{\text{Yule1}}|$ follow from inequality $|S_{\text{Phi}}| \leq |S_{\text{Yule1}}|$ and Propositions 3 and 4.
■

5. Chance-corrected Coefficients

When comparing two variables some degree of agreement may be expected due to chance alone. A coefficient may be corrected for association due to chance if it does not have zero value in the case that the variables are statistically independent. Coefficient

$$S_{\text{Cohen}} = \frac{2(ad-bc)}{p_1q_2 + p_2q_1}$$

is an example of a coefficient that is corrected for chance. The chance-corrected coefficients considered in this section have a form

$$\frac{a+d-E(a+d)}{1-E(a+d)}$$

where $E(a+d)$ is unique for each coefficient. The other chance-corrected coefficients are

$$S_{\text{GK}} = \frac{2\min(a,d)-b-c}{2\min(a,d)+b+c} \quad \text{(Goodman and Kruskal 1954, p. 758)}$$

$$\text{and} \qquad S_{\text{Scott}} = \frac{4ad-(b+c)^2}{(p_1+p_2)(q_1+q_2)} \quad \text{(Scott 1955)}.$$

Coefficients S_{GK}, S_{Scott}, and S_{Cohen} are related by

Proposition 8. $-1 \leq S_{\text{GK}} \overset{(i)}{\leq} S_{\text{Scott}} \overset{(ii)}{\leq} S_{\text{Cohen}} \leq 1$.
Inequality (ii) is also proved in Blackman and Koval (1993, p. 216).

Proof (i): We have $S_{\text{GK}} \leq S_{\text{Scott}}$ if and only if

$$E(a+d)_{\text{GK}} \geq E(a+d)_{\text{Scott}}$$
$$\frac{\max(p_1+p_2, q_1+q_2)}{2} \geq \left(\frac{p_1+p_2}{2}\right)^2 + \left(\frac{q_1+q_2}{2}\right)^2.$$

Assume $(p_1+p_2) \geq (q_1+q_2)$. Then

$$\frac{p_1+p_2}{2}\left(1-\frac{p_1+p_2}{2}\right) \geq \left(\frac{q_1+q_2}{2}\right)^2$$
$$\frac{p_1+p_2}{2}\left(\frac{q_1+q_2}{2}\right) \geq \left(\frac{q_1+q_2}{2}\right)^2$$
$$\frac{p_1+p_2}{2} \geq \frac{q_1+q_2}{2}.$$

Proof (ii): We have $S_{\text{Scott}} \leq S_{\text{Cohen}}$ iff

$$E(a+d)_{\text{Scott}} \geq E(a+d)_{\text{Cohen}}$$
$$\left(\frac{p_1+p_2}{2}\right)^2 + \left(\frac{q_1+q_2}{2}\right)^2 \geq p_1p_2 + q_1q_2.$$

Since

$$A(p_1,p_2) = \frac{p_1+p_2}{2} \geq \sqrt{p_1p_2} = G(p_1,p_2)$$
$$\text{and} \qquad A(q_1,q_2) = \frac{q_1+q_2}{2} \geq \sqrt{q_1q_2} = G(q_1,q_2)$$

the desired inequality follows from application of Theorem 1.
■

6. Discussion

Bounds of resemblance measures for binary variables were derived in this paper using the arithmetic-geometric-harmonic mean inequality. It was shown that some coefficients are bounds of each other, and that the values of some coefficients are in general more similar compared to the values of other presence/absence coefficients. The arithmetic-geometric-harmonic mean inequality may also be used to obtain bounds of parameter families instead of individual coefficients. For instance, let

$$u_1(x,\theta) = \frac{x}{x+\theta b} \quad \text{and} \quad u_2(x,\theta) = \frac{x}{x+\theta c}$$

where $\theta > 0$ to avoid negative values, and where x can for instance be the quantities a, $a + d$, or $a + \sqrt{ad}$. Gower and Legendre (1986, p. 13) defined the parameter families

$$S_{\text{GL1}}(\theta) = \frac{a}{a + \theta(b + c)} \quad \text{and} \quad S_{\text{GL2}}(\theta) = \frac{a + d}{a + \theta(b + c) + d}.$$

Family $S_{\text{GL1}}(\theta)$ is equivalent to the harmonic mean of $u_1(a, 2\theta)$ and $u_2(a, 2\theta)$, whereas family $S_{\text{GL2}}(\theta)$ is equivalent to the harmonic mean of $u_1(a + d, 2\theta)$ and $u_2(a + d, 2\theta)$. Members of the two families are

$$S_{\text{Jac}} = S_{\text{GL1}}(\theta = 1) = \frac{a}{a + b + c}$$
$$S_{\text{Gleas}} = S_{\text{GL1}}(\theta = 1/2) = \frac{2a}{2a + b + c}$$

and

$$S_{\text{SM}} = S_{\text{GL2}}(\theta = 1) = \frac{a + d}{a + b + c + d} \quad \text{(Sokal and Michener 1958).}$$

A straightforward application of Theorem 1 is

Theorem 3. $\min(u_1, u_2) \leq H(u_1, u_2) \leq G(u_1, u_2) \leq A(u_1, u_2) \leq \max(u_1, u_2)$.

For example, from Theorem 3 it follows that

$$0 \leq \frac{a}{a + \theta\max(b, c)} \leq \frac{2a}{2a + \theta(b + c)} \leq \frac{a}{\sqrt{(a + \theta b)(a + \theta c)}}$$
$$\leq \frac{1}{2}\left(\frac{a}{a + \theta b} + \frac{a}{a + \theta c}\right) \leq \frac{a}{a + \theta\min(b, c)} \leq 1.$$

The inequality is a (partial) parametrized version of the inequality in Section 1. From a mathematical point of view, Theorem 3 may be used to obtain more general results using parameter families compared to the results for individual coefficients in Sections 3 to 5. Practitioners are perhaps more interested in the bounds for individual coefficients derived in this paper. Some additional bounds can be found in the appendix. The inequalities are presented without proof. Some bounds are not difficult to derive, others may be obtained using some of the tools discussed in this paper.

7. Appendix

In this appendix we note the following inequalities without proof.
$0 \leq S_{\text{RR}} \leq S_{\text{Jac}} \leq S_{\text{SM}} \leq 1$, where

$$S_{\text{RR}} = \frac{a}{a+b+c+d} \qquad \text{(Russel and Rao 1940)}$$
$$S_{\text{Jac}} = \frac{a}{a+b+c} \qquad \text{(Jaccard 1912)}$$
$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \qquad \text{(Sokal and Michener 1958).}$$

$0 \leq S_{\text{SS1}} \leq S_{\text{SS2}} \leq 1$, where

$$S_{\text{SS1}} = \frac{ad}{\sqrt{p_1 p_2 q_1 q_2}} \qquad \text{(Sokal and Sneath 1963)}$$
$$S_{\text{SS2}} = \frac{1}{4}\left(\frac{a}{p_1} + \frac{a}{p_2} + \frac{d}{q_1} + \frac{d}{q_2}\right) \qquad \text{(Sokal and Sneath 1963).}$$

$0 \leq S_{\text{Jac}} \leq S_{\text{BUB}} \leq S_{\text{SS3}} \leq 1$, where

$$S_{\text{Jac}} = \frac{a}{a+b+c} \qquad \text{(Jaccard 1912)}$$
$$S_{\text{BUB}} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}} \qquad \text{(Baroni-Urbani and Buser 1976, p. 258)}$$
$$S_{\text{SS3}} = \frac{2(a+d)}{2a+b+c+2d} \qquad \text{(Sokal and Sneath 1963).}$$

$-1 \leq S_{\text{NS}} \leq S_{\text{McC}} \leq 1$, where

$$S_{\text{NS}} = \frac{2a-b-c}{2a+b+c} \qquad \text{(No source)}$$
$$S_{\text{McC}} = \frac{a^2-bc}{p_1 p_2} \qquad \text{(McConnaughey 1964).}$$

$S_{\text{Mich}} \leq S_{\text{SM}} \leq 1$, where

$$S_{\text{Mich}} = \frac{4(ad-bc)}{(a+b+c+d)^2} \qquad \text{(Michael 1920)}$$
$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \qquad \text{(Sokal and Michener 1958).}$$

References

ABRAMOWITZ, M. and STEGUN, I.A. (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (9th print.), New York: Dover.

BARONI-URBANI, C. and BUSER, M.W. (1976), "Similarity of Binary Data," *Systematic Zoology, 25,* 251-259.

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.

BAULIEU, F.B. (1997), "Two Variant Axiom Systems for Presence/absence Based Dissimilarity Coefficients," *Journal of Classification, 14,* 159-170.

BLACKMAN, N. J.-M. and KOVAL, J.J. (1993), "Estimating Rater Agreement in 2×2 Tables: Correction for Chance and Intraclass Correlation," *Applied Psychological Measurement, 17,* 211-223.

BOYCE, R.L. and ELLISON, P.C. (2001), "Choosing the Best Similarity Index when Performing Fuzzy Set Ordination on Binary Data," *Journal of Vegetational Science, 12,* 711-720.

BRAUN-BLANQUET, J. (1932), *Plant Sociology: The Study of Plant Communities* (Authorized English translation of Pflanzensoziologie), New York: McGraw-Hill.

BREN, M. and BATAGELJ, V. (2006), "The Metric Index," *Croatica Chemica Acta, 79,* 399-410.

BULLEN, P.S. (2003), *Handbook of Means and Their Inequalities*, Dordrecht, The Netherlands: Kluwer.

COHEN, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement, 14,* 37-46.

DICE, L.R. (1945), "Measures of the Amount of Ecologic Association Between Species", *Ecology, 26,* 297-302.

FICHET, B. (1986), "Distances and Euclidean Distances for Presence-Absence Characters and Their Application to Factor Analysis," in *Multidimensional Data Analysis*, eds., J. de Leeuw, W.J. Heiser, J.J. Meulman, and F. Critchley, Leiden: DSWO Press, pp. 23-46.

FLEISS, J.L. (1975), "Measuring Agreement Between Two Judges on the Presence or Absence of a Trait," *Biometrics, 31,* 651-659.

GLEASON, H.A. (1920), "Some Applications of the Quadrat Method," *Bulletin of the Torrey Botanical Club, 47,* 21-33.

GOODMAN, L.A. and KRUSKAL, W. H. (1954), "Measures of Association for Cross Classifications," *Journal of the American Statistical Association, 49,* 732-764.

GOWER, J.C. (1986), "Euclidean Distance Matrices," in *Multidimensional Data Analysis*, eds., J. de Leeuw, W.J. Heiser, J.J. Meulman, and F. Critchley, Leiden: DSWO Press, pp. 11-22.

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

HUBÁLEK, Z. (1982), "Coefficients of Association and Similarity Based on Binary (Presence-Absence) Data: An Evaluation," *Biological Reviews, 57*, 669-689.

JACCARD, P. (1912), "The Distribution of the Flora in the Alpine Zone," *The New Phytologist, 11*, 37-50.

JANSON, S. and VEGELIUS, J. (1981), "Measures of Ecological Association," *Oecologia, 49*, 371-376.

KULCZYŃSKI, S. (1927), "Die Pflanzenassociationen der Pienenen," *Bulletin International de L'Académie Polonaise des Sciences et des Letters, classe des sciences mathematiques et naturelles, Serie B, Supplément II, 2*, 57-203.

LOEVINGER, J.A. (1948), "The Technique of Homogeneous Tests Compared with Some Aspects of Scale Analysis and Factor Analysis," *Psychological Bulletin, 45,* 507-530.

MAXWELL, A.E. and PILLINER, A.E.G. (1968), "Deriving Coefficients of Reliability and Agreement for Ratings," *British Journal of Mathematical and Statistical Psychology, 21,* 105-116.

MCCONNAUGHEY, B.H. (1964), "The Determination and Analysis of Plankton Communities," *Marine Research, Special No, Indonesia*, 1-40.

MICHAEL, E.L. (1920), "Marine Ecology and the Coefficient of Association: A Plea in Behalf of Quantitative Biology," *Journal of Animal Ecology, 8,* 54-59.

OCHIAI, A. (1957), "Zoogeographic Studies on the Soleoid Fishes Found in Japan and Its Neighboring Regions," *Bulletin of the Japanese Society for Fish Science, 22*, 526-530.

RUSSEL, P.F. and RAO, T.R. (1940), "On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras," *Journal of Malaria Institute India, 3*, 153-178.

SCOTT, W.A. (1955), "Reliability of Content Analysis: The Case of Nominal Scale Coding," *Public Opinion Quarterly, 19,* 321-325.

SIMPSON, G.G. (1943), "Mammals and the Nature of Continents," *American Journal of Science, 241,* 1-31.

SOKAL, R.R. and MICHENER, C. D. (1958), "A Statistical Method for Evaluating Systematic Relationships", *University of Kansas Science Bulletin, 38*, 1409-1438.

SOKAL, R.R. and SNEATH, R. H. (1963), *Principles of Numerical Taxonomy*, San Francisco: W. H. Freeman and Company.

SORGENFREI, T. (1958), *Molluscan Assemblages from the Marine Middle Miocene of South Jutland and Their Environments*, Copenhagen: Reitzel.

YULE, G.U. (1900), "On the Association of Attributes in Statistics," *Philosophical Transactions, Series A, 194,* 257-319.

YULE, G.U. (1912), "On the Methods of Measuring the Association between Two Attributes," *Journal of the Royal Statistical Society, 75,* 579-652.

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