Similarity coefficients for binary data : properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients

Warrens, M.J.

Part IV

Multivariate coefficients

CHAPTER 16

Coefficients that generalize basic characteristics

Fundamental entities in several domains of data analysis are resemblance measures or similarity coefficients. In most domains similarity measures are defined or studied for pairwise or bivariate (two-way) comparison. As an alternative to bivariate resemblance measures multivariate or multi-way coefficients may be considered. Multivariate coefficients can for example be used if one wants to determine the degree of agreement of three or more raters in psychological assessment, if one wants to know how similar the partitions obtained from three different cluster algorithms are, or if one is interested in the degree of similarity of three or more areas where certain types of species may or not may be encountered.

In this chapter multivariate formulations (for groups of objects of size k) of various of bivariate similarity coefficients (for pairs of objects) for binary data are presented. In this chapter the multivariate formulations are not functions of bivariate similarity coefficients, for example

$$\frac{S_{12} + S_{13} + S_{23}}{3} \qquad \text{(arithmetic mean)}.$$

Instead, an attempt is made in this chapter to present multi-way formulations that reflect certain basic characteristics of, and have a similar interpretation as, their two-way versions.

Chapter 16 is organized as follows. First, a class of two-way similarity coefficients for binary data is considered, that can be written as functions of two variables a and d, for example

$$S_{\text{Jac}} = \frac{a}{a+b+c} = \frac{a}{1-d}.$$

This class of coefficients is generalized by reformulating the two-way quantities a and d into multivariate variables $a^{(k)}$ and $d^{(k)}$. Similarity coefficients that can be defined using only the variables $a^{(k)}$ and $d^{(k)}$ are named after Bennani-Dosse (1993) and Heiser and Bennani (1997), who first presented these coefficients for the similarity of three variables.

For the second class of coefficients the quantity p_i (q_i), that is, the proportion of 1s (0s) in variable x_i, is involved in the definition. Throughout the chapter it is shown what properties from the two-way case are preserved with the multivariate formulations of various similarity coefficients presented here.

16.1 Bennani-Heiser coefficients

Many bivariate coefficients are written as functions of four dependent variables a, b, c and d. Although b and c are two separate variables, most coefficients are defined to be symmetric in b and c. As noted by Heiser and Bennani (1997, p. 195), a large number of two-way measures are characterized by the number of positive matches (a), negative matches (d), and mismatches (b, c). This is especially the case for similarity coefficients that are rational functions, linear in both numerator and denominator, for example

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \quad \text{or} \quad S_{\text{Jac}} = \frac{a}{a+b+c}.$$

Suppose $x_1, x_2, ..., x_k$ are k binary variables. Instead of variables a, b, c and d (as used and defined in Part I), we define for k binary variables and multivariate coefficients, the two variables

$$\begin{aligned} a^{(k)} &= \text{the proportions of 1s that} \quad x_1, x_2, ..., x_k \quad \text{share in the same positions} \\ d^{(k)} &= \text{the proportions of 0s that} \quad x_1, x_2, ..., x_k \quad \text{share in the same positions.} \end{aligned}$$

Similarity coefficients that can be defined using the variables $a^{(k)}$ and $d^{(k)}$ are named after Bennani-Dosse (1993) and Heiser and Bennani (1997), who first presented these coefficients for three variables. Although many Bennani-Heiser coefficients are linear in both numerator and denominator, it is not a necessary property. In the following, let $S^{(k)}$ denote a multivariate similarity coefficient for groups of size k.

Jaccard (1912) studied flora in several districts of the Alpine mountains. To measure the degree of similarity of two districts, Jaccard used the ratio

$$S_{\text{Jac}}^{(2)} = \frac{\text{Number of species common to the two districts}}{\text{Total number of species in the two districts}} = \frac{a^{(2)}}{1-d^{(2)}}.$$

A seemingly proper and straightforward 3-way formulation of Jaccard coefficient would be

$$S_{\text{Jac}}^{(3)} = \frac{\text{Number of species common to the three districts}}{\text{Total number of species in the three districts}} = \frac{a^{(3)}}{1 - d^{(3)}}.$$

The complement $1 - S_{\text{Jac}}^{(3)}$ was presented in Cox, Cox and Branco (1991, p. 200). The multivariate formulation of S_{Jac} is then given by

$$S_{\text{Jac}}^{(k)} = \frac{a^{(k)}}{1 - d^{(k)}}.$$

The two-way Jaccard coefficient S_{Jac} is a member of $S_{\text{GL1}}(\theta)$, given by

$$S_{\text{GL1}}(\theta) = \frac{a}{a + \theta(b + c)} = \frac{a}{(1 - \theta)a + \theta(1 - d)}$$

which is one of the parameter families studied for metric properties in Gower and Legendre (1986). A possible multivariate formulation of $S_{\text{GL1}}(\theta)$ is given by

$$S_{\text{GL1}}^{(k)}(\theta) = \frac{a^{(k)}}{(1 - \theta)a^{(k)} + \theta(1 - d^{(k)})}.$$

Members of $S_{\text{GL1}}^{(k)}(\theta)$ are (see Section 3.1)

$$\begin{aligned} S_{\text{GL1}}^{(k)}(\theta = 1) &= S_{\text{Jac}}^{(k)} = \frac{a^{(k)}}{1 - d^{(k)}} \\ S_{\text{GL1}}^{(k)}(\theta = 1/2) &= S_{\text{Gleas}}^{(k)} = \frac{2a^{(k)}}{1 + a^{(k)} - d^{(k)}} \\ S_{\text{GL1}}^{(k)}(\theta = 2) &= S_{\text{SS1}}^{(k)} = \frac{a^{(k)}}{2 - a^{(k)} - 2d^{(k)}}. \end{aligned}$$

The formulations of $S_{\text{GL1}}(\theta)$ and $S_{\text{GL2}}(\theta)$ (and their multivariate formulations presented in this chapter) are related to the concept of global order equivalence (Sibson, 1972; Batagelj and Bren, 1995). We first present a generalization of global order equivalence for multivariate coefficients that are Bennani-Heiser coefficients. Two Bennani-Heiser coefficients, $S^{(k)}$ and $S^{(k)*}$, are said to be globally order equivalent if

$$S(a_1^{(k)}, d_1^{(k)}) \;>\; S(a_2^{(k)}, d_2^{(k)})$$

$$\text{if and only if} \qquad S^*(a_1^{(k)}, d_1^{(k)}) \;>\; S^*(a_2^{(k)}, d_2^{(k)}).$$

If two coefficients are globally order equivalent, they are interchangeable with respect to an analysis method that is invariant under ordinal transformations. Proposition 16.1 is a straightforward generalization of Theorem 3.1.

Proposition 16.1. *Two members of* $S^{(k)}_{\mathrm{GL1}}(\theta)$ *are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S^{(k)}_{\mathrm{GL1}}(\theta)$, we have

$$\frac{a_1^{(k)}}{(1-\theta)a_1^{(k)}+\theta(1-d_1^{(k)})} > \frac{a_2^{(k)}}{(1-\theta)a_2^{(k)}+\theta(1-d_2^{(k)})}$$

$$\frac{a_1^{(k)}}{1-d_1^{(k)}} > \frac{a_2^{(k)}}{1-d_2^{(k)}}.$$

Since an arbitrary ordinal comparison with respect to $S^{(k)}_{\mathrm{GL1}}(\theta)$ does not depend on the value of θ, any two members of $S^{(k)}_{\mathrm{GL1}}(\theta)$ are globally order equivalent. □

Instead of positive matches only, one may also be interested in a similarity coefficient or resemblance measure that involves the negative matches. The simple matching coefficient is given by

$$\begin{aligned} S^{(2)}_{\mathrm{SM}} =& \frac{\text{Number of attributes present and absent in two objects}}{\text{Total number of attributes}} \\ =& a^{(2)} + d^{(2)}. \end{aligned}$$

The multivariate formulation of S_{SM} is then given by

$$S^{(k)}_{\mathrm{SM}} = a^{(k)} + d^{(k)}.$$

The simple matching coefficient (S_{SM}) belongs to another parameter family studied in Gower and Legendre (1986), which is given by

$$S_{\mathrm{GL2}}(\theta) = \frac{a+d}{\theta+(1-\theta)(a+d)}.$$

The multivariate extension of family $S_{\mathrm{GL2}}(\theta)$ is given by

$$S^{(k)}_{\mathrm{GL2}}(\theta) = \frac{a^{(k)}+d^{(k)}}{\theta+(1-\theta)(a^{(k)}+d^{(k)})}.$$

Members of $S^{(k)}_{\mathrm{GL2}}(\theta)$ are (see Section 3.1)

$$\begin{aligned} S^{(k)}_{\mathrm{GL2}}(\theta=1) &= S^{(k)}_{\mathrm{SM}} = a^{(k)} + d^{(k)} \\ S^{(k)}_{\mathrm{GL2}}(\theta=1/2) &= S^{(k)}_{\mathrm{SS2}} = \frac{2(a^{(k)}+d^{(k)}}{1+a^{(k)}+d^{(k)}} \\ S^{(k)}_{\mathrm{GL2}}(\theta=2) &= S^{(k)}_{\mathrm{RT}} = \frac{a^{(k)}+d^{(k)}}{2-a^{(k)}-d^{(k)}}. \end{aligned}$$

Proposition 16.2 demonstrates the global order equivalence property for $S^{(k)}_{\mathrm{GL2}}(\theta)$. The assertion is a straightforward generalization of Theorem 3.2.

Proposition 16.2. *Two members of* $S^{(k)}_{\mathrm{GL2}}(\theta)$ *are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S^{(k)}_{\mathrm{GL2}}(\theta)$, we have

$$\frac{a_1^{(k)} + d_1^{(k)}}{\theta + (1-\theta)(a_1^{(k)} + d_1^{(k)})} > \frac{a_2^{(k)} + d_2^{(k)}}{\theta + (1-\theta)(a_2^{(k)} + d_2^{(k)})}$$

$$a_1^{(k)} + d_1^{(k)} > a_2^{(k)} + d_2^{(k)}$$

which does not depend on the value of θ. □

Other Bennani-Heiser coefficients are generalizations of bivariate coefficients by Russel and Rao (1940) (S_{RR}) and Baroni-Urabani and Buser (1976, p. 258). Possible multivariate formulations of these coefficients are given by

$$S^{(k)}_{\mathrm{RR}} = a^{(k)}$$
$$S^{(k)}_{\mathrm{BUB}} = \frac{a^{(k)} + \sqrt{a^{(k)}d^{(k)}}}{1 - d^{(k)} + \sqrt{a^{(k)}d^{(k)}}}$$
$$\text{and} \quad S^{(k)}_{\mathrm{BUB2}} = \frac{2a^{(k)} + d^{(k)} - 1 + \sqrt{a^{(k)}d^{(k)}}}{1 - d^{(k)} + \sqrt{a^{(k)}d^{(k)}}}.$$

16.2 Dice's association indices

Let p_i and q_i denote the proportion of 1s, respectively 0s, in variable x_i. For the multivariate formulations presented in this section it is useful to work with a different generalization of the concept of globally order equivalent (Sibson, 1972). Let $x_{1,k} = \{x_1, x_2, ..., x_k\}$ and $y_{1,k} = \{y_1, y_2, ..., y_k\}$ denote two k-tuples. Two multivariate coefficients, S and S^*, are said to be globally order equivalent if

$$S(x_{1,k}) > S(y_{1,k}) \quad \text{if and only if} \quad S^*(x_{1,k}) > S^*(y_{1,k}).$$

Dice (1945, p. 298) proposed two-way association indices that consist of the amount of similarity between any two species x_1 and x_2, relative to the occurrence of either x_1 or x_2. Hence, for every pair of variables there are two measures, namely

$$S_{\mathrm{Dice1}} = \frac{a^{(2)}}{p_1} \quad \text{and} \quad S_{\mathrm{Dice2}} = \frac{a^{(2)}}{p_2}.$$

What became know as the Dice coefficient is Dice's coincidence index, which is the harmonic mean of the two association measures, given by

$$S^{(2)}_{\mathrm{Gleas}} = \frac{2a^{(2)}}{p_1 + p_2}.$$

Dice (1945, p. 300) already noted that the coefficients he proposed could be easily expanded to measure the amount of association between three or more species. Thus, for every triple of variables there are three coefficients, namely

$$\frac{a^{(3)}}{p_1},\ \frac{a^{(3)}}{p_2} \quad \text{and} \quad \frac{a^{(3)}}{p_3}.$$

The three-way extension of S_{Gleas} is then the harmonic mean of the three association indices, which is given by

$$S^{(3)*}_{\text{Gleas}} = \frac{3a^{(3)}}{p_1 + p_2 + p_3}$$

where the asterisk $(*)$ is used to denote that this formulation is different from the Bennani-Heiser multivariate generalization presented in the previous section. The corresponding multivariate formulation of S_{Gleas} is given by

$$S^{(k)*}_{\text{Gleas}} = \frac{k\, a^{(k)}}{\sum_{i=1}^{k} p_i}.$$

Instead of the harmonic mean, we may apply other special cases of the power mean (Section 3.2) to Dice's association indices, to obtain multivariate generalizations of various other two-way similarity coefficients. Hence, we obtain

$$\begin{aligned}
S^{(k)}_{\text{BB}} &= \frac{a^{(k)}}{\max(p_1, p_2, ..., p_k)} && \text{(minimum)} \\
S^{(k)}_{\text{Kul}} &= \frac{1}{k}\sum_{i=1}^{k} \frac{a^{(k)}}{p_i} && \text{(arithmetic mean)} \\
S^{(k)}_{\text{DK}} &= \frac{a^{(k)}}{\prod_{i=1}^{k} p_i^{1/k}} && \text{(geometric mean)} \\
S^{(k)}_{\text{Sim}} &= \frac{a^{(k)}}{\min(p_1, p_2, ..., p_k)} && \text{(maximum).}
\end{aligned}$$

In addition, the product of the two association indices defines a coefficient by Sorgenfrei (1958). Its multivariate extension is given by

$$S^{(k)}_{\text{Sorg}} = \frac{\left[a^{(k)}\right]^k}{\prod_{i=1}^{k} p_i}.$$

An alternative two-way formulation of S_{Kul} is given by

$$S^{(2)}_{\text{Kul}} = \frac{1}{2}\left[\frac{a^{(2)}}{p_1} + \frac{a^{(2)}}{p_2}\right] = \frac{a^{(2)}(p_1 + p_2)}{2p_1p_2}.$$

From this formulation we may present the alternative multivariate extension of $S^{(2)}_{\text{Kul}}$ given by

$$S^{(k)*}_{\text{Kul}} = \frac{\left[a^{(k)}\right]^{k-1} \sum_{i=1}^{k} p_i}{k \prod_{i=1}^{k} p_i}$$

where the asterisk $(*)$ is used to denote that this formulation is different from $S^{(k)}_{\text{Kul}}$.

A two-way coefficient by McConnaughey (1964) is given by

$$S^{(2)}_{\text{McC}} = \frac{a^{(2)}(p_1 + p_2) - p_1 p_2}{p_1 p_2}.$$

A possible multivariate generalization of $S^{(2)}_{\text{McC}}$ is given by

$$S^{(k)}_{\text{McC}} = \frac{\frac{2}{k}\left[a^{(k)}\right]^{k-1} \sum_{i=1}^{k} p_i - \prod_{i=1}^{k} p_i}{\prod_{i=1}^{k} p_i}.$$

As it turns out, multivariate formulation $S^{(k)*}_{\text{Kul}}$ preserves an order equivalence property with respect to $S^{(k)}_{\text{McC}}$, which is not preserved by power mean multivariate formulation $S^{(k)}_{\text{Kul}}$. Some additional notation is required: let $p(x_i)$ denote the proportion of 1s in variable x_i.

Proposition 16.3. *Coefficients $S^{(k)}_{\text{McC}}$ and $S^{(k)*}_{\text{Kul}}$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S^{(k)}_{\text{McC}}$, we have

$$\frac{\frac{2}{k}\left[a^{(k)}_1\right]^{k-1} \sum_{i=1}^{k} p(x_i) - \prod_{i=1}^{k} p(x_i)}{\prod_{i=1}^{k} p(x_i)} > \frac{\frac{2}{k}\left[a^{(k)}_2\right]^{k-1} \sum_{i=1}^{k} p(y_i) - \prod_{i=1}^{k} p(y_i)}{\prod_{i=1}^{k} p(y_i)}$$

if and only if

$$\frac{\left[a^{(k)}_1\right]^{k-1} \sum_{i=1}^{k} p(x_i)}{\prod_{i=1}^{k} p(x_i)} > \frac{\left[a^{(k)}_2\right]^{k-1} \sum_{i=1}^{k} p(y_i)}{\prod_{i=1}^{k} p(y_i)}.$$

The same inequality is obtained for an arbitrary ordinal comparison with respect to $S^{(k)*}_{\text{Kul}}$. □

We end this section with two multivariate formulations of two measures presented in Sokal and Sneath (1963). These authors considered two coefficients (S_{SS3} and S_{SS4}) that can be defined as the arithmetic mean, respectively the square root of the geometric mean, of the quantities

$$\frac{a^{(2)}}{p_1},\ \frac{a^{(2)}}{p_2},\ \frac{d^{(2)}}{q_1} \quad \text{and} \quad \frac{d^{(2)}}{q_2}.$$

The arithmetic mean is given by

$$S^{(2)}_{\text{SS3}} = \frac{1}{4}\left[\frac{a^{(2)}}{p_1} + \frac{a^{(2)}}{p_2} + \frac{d^{(2)}}{q_1} + \frac{d^{(2)}}{q_2}\right].$$

A straightforward generalization of S_{SS3} is

$$S^{(k)}_{\text{SS3}} = \frac{1}{2k}\sum_{i=1}^{k} \frac{a^{(k)}}{p_i} + \frac{1}{2k}\sum_{i=1}^{k} \frac{d^{(k)}}{q_i}.$$

The square root of the geometric mean and a possible multivariate generalization are given by

$$S_{\text{SS4}}^{(2)} = \frac{a^{(2)}d^{(2)}}{[p_1p_2q_1q_2]^{1/2}}$$

and

$$S_{\text{SS4}}^{(k)} = \frac{a^{(k)}d^{(k)}}{\prod_{i=1}^{k}[p_iq_i]^{1/k}}.$$

16.3 Bounds

In this section it is shown that some multivariate coefficients are bounds with respect to each other. Proposition 16.4 is a straightforward generalization of Proposition 3.3.

Proposition 16.4. *It holds that* $S_{\text{GL2}}^{(k)}(\theta) \geq S_{\text{GL1}}^{(k)}(\theta)$.
Proof: $S_{\text{GL2}}^{(k)}(\theta) \geq S_{\text{GL1}}^{(k)}(\theta)$ if and only if $1 \geq a^{(k)} + d^{(k)}$.

Proposition 16.5 is a straightforward generalization of Proposition 3.6. Only the proof of inequality (i) is slightly more involved.

Proposition 16.5. *It holds that*

$$0 \leq S_{\text{Sorg}}^{(k)} \overset{(i)}{\leq} S_{\text{Jac}}^{(k)} \overset{(ii)}{\leq} S_{\text{BB}}^{(k)} \overset{(iii)}{\leq} S_{\text{Gleas}}^{(k)*} \overset{(iv)}{\leq} S_{\text{DK}}^{(k)} \overset{(v)}{\leq} S_{\text{Kul}}^{(k)} \overset{(vi)}{\leq} S_{\text{Sim}}^{(k)} \leq 1.$$

Proof: Inequality (i) holds if and only if

$$\prod_{i=1}^{k} p_i \geq \left[a^{(k)}\right]^{k-1}\left[1 - d^{(k)}\right].$$

First, it holds that

$$\prod_{i=1}^{k} p_i \geq \sum_{i=1}^{k}\left[a^{(k)}\right]^{k-1}\left[p_i - a^{(k)}\right] + \left[a^{(k)}\right]^{k} = \left[a^{(k)}\right]^{k-1}\left[\sum_{i=1}^{k} p_i - (k-1)a^{(k)}\right].$$

Because $\sum_{i=1}^{k} p_i - (k-1)a^{(k)} \geq 1 - d^{(k)}$, inequality (i) is true. Inequality (ii) holds if and only if $d^{(k)} + \max(p_1, p_2, ..., p_k) \leq 1$. Inequality (iii) holds if and only if

$$\max(p_1, p_2, ..., p_k) \geq \frac{1}{k}\sum_{i=1}^{k} p_i.$$

Inequalities (iv) and (v) are true because the harmonic mean of k numbers is equal or smaller than the geometric mean of the k numbers, which in turn is equal or smaller to the arithmetic mean of the numbers. Inequality (vi) holds if and only if

$$\frac{1}{k}\sum_{i=1}^{k} p_i \geq \min(p_1, p_2, ..., p_k). \qquad \square$$

16.4 Epilogue

In this chapter multivariate formulations of various two-way similarity coefficients for binary data were presented. Cox, Cox and Branco (1991) pointed out that multivariate resemblance measures, for example, three-way or four-way similarity coefficients instead of two-way similarity coefficients, may be used to detect possible higher-order relations between the objects. Consider the following data matrix for five binary strings on fourteen attributes.

objects	attributes													
1	1	1	1	1	1	1	0	0	0	0	0	0	0	1
2	1	1	1	0	0	0	1	1	1	1	0	0	0	0
3	1	0	0	1	1	0	1	1	0	0	1	1	0	0
4	0	1	0	0	1	1	1	0	1	0	1	0	1	0
5	0	0	1	1	0	1	1	0	0	1	0	1	1	0

The multivariate Jaccard (1912) coefficient was defined as

$$S_{\text{Jac}}^{(k)} = \frac{a^{(k)}}{1 - d^{(k)}}.$$

It can be verified for these data, that the ten two-way Jaccard coefficients between the five objects are all equal ($S_{\text{Jac}} = \frac{3}{11}$). In addition the ten three-way Jaccard coefficients are also all equal ($S_{\text{Jac}}^{(3)} = \frac{1}{13}$). Thus, no discriminative information about the five objects is obtained from either two-way or three-way Jaccard coefficient. However, the four-way Jaccard similarity coefficient between objects two, three, four and five ($S_{\text{Jac}}^{(4)} = \frac{1}{13}$) differs from the other four four-way Jaccard similarity coefficient ($S_{\text{Jac}}^{(4)} = 0$). The artificial example shows that higher-order information can put objects two, three, four and five in a group separated from object 1. Of course, one may also argue that the wrong two-way and three-way similarity coefficient has been specified.

Two major classes of multivariate formulations were distinguished. The first class is referred to as Bennani-Heiser similarity coefficients, which contains all measures that can be defined using only two dependent variables. Many of these Bennani-Heiser similarity coefficients are fractions, linear in both numerator and denominator. As it turned out, a second class was formed by coefficients that could be formulated as functions of association indices first presented in Dice (1945). These functions include the Pythagorean means (harmonic, arithmetic and geometric means).

Two multivariate formulations of S_{Gleas} were presented. The two multivariate formulations are given by

$$S_{\text{Gleas}}^{(k)} = \frac{2a^{(k)}}{1 + a^{(k)} - d^{(k)}} \quad \text{and} \quad S_{\text{Gleas}}^{(k)*} = \frac{k\, a^{(k)}}{\sum_{i=1}^{k} p_i}$$

where $S_{\text{Gleas}}^{(k)}$ is the Bennani-Heiser similarity coefficient.

The reader may have noted that we have failed to present multivariate versions of similarity coefficients that involve the covariance $(ad-bc)$ between two variables, for example

$$
\begin{aligned}
S_{\text{Phi}} &= \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \\
S_{\text{Cohen}} &= \frac{2(ad-bc)}{p_1q_2+p_2q_1} \\
S_{\text{Loe}} &= \frac{ad-bc}{\min(p_1q_2, p_2q_1)} \\
S_{\text{Yule1}} &= \frac{ad-bc}{ad+bc}.
\end{aligned}
$$

The definition of covariance between triples of objects is already quite complex and the topic is outside the scope of the present study. However, in the next chapter an alternative way of formulating k-way generalizations of bivariate coefficients is discussed. The approach in Chapter 17 may be used to generalize coefficients that involve the covariance.

CHAPTER 17

Multi-way coefficients based on two-way quantities

Similar to the Chapter 16, Chapter 17 is devoted to multivariate formulations of various similarity coefficients. In Chapter 16 an attempt was made to present multivariate formulations that reflect certain basic characteristics of, and have a similar interpretation as, their two-way versions. In this chapter multivariate formulations of resemblance measures are presented that preserve the properties presented in Chapter 4 on correction for similarity due to chance.

Suppose the two binary variables are the ratings of two judges, rating various people on the presence or absence of a certain trait. In this field, Scott (1955), Cohen (1960), Fleiss (1975), Krippendorff (1987), among others, have proposed measures that are corrected for chance. The best-known example is perhaps the kappa-statistic (Cohen, 1960; S_{Cohen}). A vast amount of literature exists on extensions of S_{Cohen}, including multivariate versions of the kappa-statistic (Fleiss, 1971; Light, 1971; Schouten, 1980; Popping, 1983a; Heuvelmans and Sanders, 1993). In a different domain of data analysis, a multivariate or multi-way coefficient was proposed by Mokken (1971). Mokken's multivariate index, referred to as coefficient H, is a measure of the degree of homogeneity among k test items (Sijtsma and Molenaar, 2002). Coefficient H can be used is the same context as coefficient alpha popularized by Cronbach (1951), which is the best-known measure from classical test theory (De Gruijter and Van der Kamp, 2008).

In this chapter the $\mathcal{L}$ family of bivariate coefficients of the form $\lambda+\mu x$ is extended to a family of multivariate coefficients. For reasons of notational convenience, only coefficients of the form $\lambda+\mu a$ (coefficients for binary data) are considered, although the extensions do apply to all coefficients in the $\mathcal{L}$ family. The new family of multivariate coefficients preserve various properties derived for the $\mathcal{L}$ family in Chapter 4. For various members the complete multivariate formulations are presented. In addition, it is shown how the multivariate coefficients presented in this chapter are related to the multivariate coefficients discussed in Chapter 16.

17.1 Multivariate formulations

In Section 3.3 a family $\mathcal{L}$ was introduced that consists of coefficients of the form $\lambda + \mu a$. Let a_{ij} denote the proportion of 1s that variables x_i and x_j share in the same positions. Furthermore, let p_i denote the proportion of 1s in variable x_i. Coefficients of the form $\lambda + \mu a$ can be extended to a k-way family of coefficients that are linear in the quantity

$$\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} a_{ij}. \tag{17.1}$$

Quantity (17.1) is equal to the sum of all a_{ij}, the proportion of 1s that variables x_i and x_j share in the same positions, obtained from all $k(k-1)/2$ pairwise fourfold tables. Coefficients in family $\mathcal{L}^{(k)}$ have a form

$$\lambda^{(k)} + \mu^{(k)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} a_{ij}$$

where $\lambda^{(k)}$ and $\mu^{(k)}$ are functions of the p_i only. For $k = 2$, we have $\lambda^{(2)} = \lambda$, $\mu^{(2)} = \mu$ and $\mathcal{L}^{(2)} = \mathcal{L}$. Before considering any properties of $\mathcal{L}^{(k)}$ family, we discuss some members of the family.

Coefficient S_{SM} can be written as

$$S_{\text{SM}} = a_{12} + d_{12}.$$

The three-way formulation of S_{SM}, such that the coefficient is linear in $(a_{12} + a_{13} + a_{23})$, is given by

$$S_{\text{SM}}^{(3)*} = \frac{a_{12} + d_{12}}{3} + \frac{a_{13} + d_{13}}{3} + \frac{a_{23} + d_{23}}{3}$$

where the asterisks $(*)$ is used to denote that this generalization of S_{SM} is different from the multivariate formulation presented in Chapter 16. The general multivariate formulation of S_{SM} is given by

$$\begin{aligned} S_{\text{SM}}^{(k)*} &= \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij}) \\ &= 1 + \frac{4}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} a_{ij} - \frac{2}{k} \sum_{i=1}^{k} p_i. \end{aligned} \tag{17.2}$$

The quantity $2/[k(k-1)]$ in (17.2) is used to ensure $0 \leq S_{\text{SM}}^{(k)*} \leq 1$.

Coefficient S_{Gleas} can be written as

$$S_{\text{Gleas}} = \frac{2a_{12}}{p_1 + p_2}.$$

The three-way formulation of S_{Gleas}, such that the coefficient is linear in $(a_{12} + a_{13} + a_{23})$, is given by

$$S_{\text{Gleas}}^{(3)**} = \frac{a_{12} + a_{13} + a_{23}}{p_1 + p_2 + p_3}$$

where the double asterisks $(**)$ are used to denote that this generalization of S_{Gleas} is different from the two multivariate formulations of S_{Gleas} presented in Chapter 16. The general multivariate formulation of S_{Gleas} is given by

$$S_{\text{Gleas}}^{(k)**} = \frac{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}}{(k-1)\sum_{i=1}^{k} p_i}.$$

The quantity $2/(k-1)$ ensures that the value $S_{\text{Gleas}}^{(k)**}$ is between 0 and 1.

Coefficient S_{Cohen} for two binary variables is given by

$$S_{\text{Cohen}} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1} = \frac{2(a_{12} - p_1 p_2)}{p_1 + p_2 - 2p_1 p_2}.$$

The three-way formulation of S_{Cohen} such that $S_{\text{Cohen}}^{(3)}$ is linear in $(a_{12} + a_{13} + a_{23})$, is given by

$$\frac{(a_{12} + a_{13} + a_{23}) - (p_1 p_2 + p_1 p_3 + p_2 p_3)}{(p_1 + p_2 + p_3) - (p_1 p_2 + p_1 p_3 + p_2 p_3)}.$$

The general multivariate generalization of S_{Cohen} is given by

$$\frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(a_{ij} - p_i p_j)}{2^{-1}(k-1)\sum_{i=1}^{k} p_i - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} p_i p_j}.$$

This multivariate formulation of Cohen's kappa can be found in Popping (1983a) and Heuvelmans and Sanders (1993).

17.2 Main results

In this section it is shown that $\mathcal{L}^{(k)}$ family is a natural generalization of $\mathcal{L}$ family with respect to correction for similarity due to chance. The main results from Chapter 4 are here generalized and formulated for multivariate coefficients. Proposition 17.1 is a generalization of Theorem 4.1, the powerful result by Albatineh et al. (2006).

Proposition 17.1. *Two members in $\mathcal{L}^{(k)}$ family become identical after correction* (4.1) *if they have the same ratio*

$$\frac{1-\lambda^{(k)}}{\mu^{(k)}}. \tag{17.3}$$

Proof:

$$E\left[S^{(k)}\right] = \lambda^{(k)} + \mu^{(k)} E\left(\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}\right)$$

and consequently the corrected coefficient $CS^{(k)}$ becomes

$$\begin{aligned} CS^{(k)} &= \frac{S^{(k)} - E(S^{(k)})}{1 - E(S^{(k)})} \\ &= \left[\frac{1-\lambda^{(k)}}{\mu^{(k)}} - E\left(\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}\right)\right]^{-1} \left[\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} - E\left(\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}\right)\right]. \end{aligned}$$

□

Corollary 17.1. *Coefficients $S^{(k)*}_{\mathrm{SM}}$, $S^{(k)**}_{\mathrm{Gleas}}$, and $S^{(k)}_{\mathrm{Cohen}}$ become equivalent after correction* (4.1).
Proof: Using the formulas of $\lambda^{(k)}$ and $\mu^{(k)}$ corresponding to each coefficient, ratio (17.3)

$$\frac{1-\lambda^{(k)}}{\mu^{(k)}} = \frac{k-1}{2}\sum_{i=1}^{k} p_i \tag{17.4}$$

for all three coefficients. □

Note that ratio (17.4) is a natural generalization of ratio (4.5). If it is assumed that expectation $E(a) = p_1 p_2$ is appropriate for all $[k(k-1)]/2$ bivariate fourfold tables, we obtain the multivariate formulation

$$E\left(\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}\right)_{\mathrm{Cohen}} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} p_i p_j. \tag{17.5}$$

The basic building block in (17.5) is the two-way expectation $E(a) = p_1 p_2$.

Proposition 17.2. *Let $S^{(k)}$ be a member in $\mathcal{L}^{(k)}$ family for which ratio* (17.4) *is characteristic. If $E(a) = p_1 p_2$ is the appropriate expectation for all bivariate fourfold tables, then $S^{(k)}$ becomes $S^{(k)}_{\text{Cohen}}$ after correction* (4.1).

17.3 Gower-Legendre families

The heuristics used for multivariate coefficients $S^{(k)*}_{\text{SM}}$, $S^{(k)**}_{\text{Gleas}}$ and $S^{(k)}_{\text{Cohen}}$, can also be applied to other coefficients. For this form of multivariate formulation to work, a multivariate coefficient need not necessarily belong to the $\mathcal{L}^{(k)}$ family, that is, be linear in (17.1). For instance, the corresponding multivariate formulation of $S_{\text{GL1}}(\theta)$ is given by

$$S^{(k)*}_{\text{GL1}}(\theta) = \left[(1-2\theta)\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} + \theta(k-1)\sum_{i=1}^{k} p_i\right]^{-1} \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}.$$

Members of family $S^{(k)*}_{\text{GL1}}(\theta)$ are

$$S^{(k)*}_{\text{GL1}}\left(\theta = \frac{1}{2}\right) = S^{(k)**}_{\text{Gleas}} = \frac{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}}{(k-1)\sum_{i=1}^{k} p_i}$$

$$\text{and} \qquad S^{(k)*}_{\text{GL1}}(\theta = 1) = S^{(k)*}_{\text{Jac}} = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}}{(k-1)\sum_{i=1}^{k} p_i - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}}.$$

Multivariate generalizations of other similarity coefficients may be formulated accordingly. Coefficient $S^{(k)**}_{\text{Gleas}}$ is in the $\mathcal{L}^{(k)}$ family, whereas $S^{(k)*}_{\text{Jac}}$ is not.

If two coefficients are globally order equivalent, they are interchangeable with respect to an analysis method that is invariant under ordinal transformations. Proposition 17.3 is, similar as Proposition 16.1, a straightforward generalization of Theorem 3.1.

Proposition 17.3. *Two members of $S^{(k)*}_{\text{GL1}}(\theta)$ are globally order equivalent.*
Proof: Let x_1 and x_2 denote two different versions of (17.1), and let y_1 and y_2 denote two different versions of the quantity $(k-1)\sum_{i=1}^{k} p_i$. For an arbitrary ordinal comparison with respect to $S^{(k)*}_{\text{GL1}}(\theta)$, we have

$$\frac{x_1}{(1-2\theta)x_1 + \theta y_1} > \frac{x_2}{(1-2\theta)x_2 + \theta y_2} \quad \text{if and only if} \quad \frac{x_1}{y_1} > \frac{x_2}{y_2}.$$

Since an arbitrary ordinal comparison with respect to $S^{(k)*}_{\text{GL1}}(\theta)$ does not depend on the value of θ, any two members of $S^{(k)*}_{\text{GL1}}(\theta)$ are globally order equivalent. □

A multivariate generalization of parameter family $S_{\text{GL2}}(\theta)$ is given by

$$S^{(k)*}_{\text{GL2}}(\theta) = \frac{2^{-1}k(k-1) + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} - (k-1)\sum_{i=1}^{k} p_i}{2^{-1}k(k-1) + 2(1-\theta)\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} + (\theta-1)(k-1)\sum_{i=1}^{k} p_i}.$$

Note that $S^{(k)*}_{\mathrm{GL2}}(\theta = 1) = S^{(k)*}_{\mathrm{SM}}$. Proposition 17.4 demonstrates the global order equivalence property for $S^{(k)*}_{\mathrm{GL2}}(\theta)$. The assertion is, similar as Proposition 16.2, a straightforward generalization of Theorem 3.2.

Proposition 17.4. *Two members of $S^{(k)*}_{\mathrm{GL2}}(\theta)$ are globally order equivalent.*
Proof: The proof is similar to the proof of Proposition 17.3. In addition to the quantities used in that proof, let $z = 2^{-1}k(k-1)$. For an arbitrary ordinal comparison with respect to $S^{(k)*}_{\mathrm{GL2}}(\theta)$, we have

$$\frac{z + 2x_1 - y_1}{z + 2(1-\theta)x_1 + (\theta - 1)y_1} > \frac{z + 2x_2 - y_2}{z + 2(1-\theta)x_2 + (\theta - 1)y_2}$$
$$2x_1 - y_1 > 2x_2 - y_2.$$

Since an arbitrary ordinal comparison with respect to $S^{(k)*}_{\mathrm{GL2}}(\theta)$ does not depend on the value of θ, any two members of $S^{(k)*}_{\mathrm{GL2}}(\theta)$ are globally order equivalent. □

Some multivariate coefficients are bounds with respect to each other. Proposition 17.5 is, similar to Proposition 16.4, a generalization of Proposition 3.3.

Proposition 17.5. *It holds that $S^{(k)*}_{\mathrm{GL2}}(\theta) \geq S^{(k)*}_{\mathrm{GL1}}(\theta)$.*
Proof: $S^{(k)*}_{\mathrm{GL2}}(\theta) \geq S^{(k)*}_{\mathrm{GL1}}(\theta)$ if and only if

$$\left[\frac{k(k-1)}{2} + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} - (k-1)\sum_{i=1}^{k} p_i\right]\left[(k-1)\sum_{i=1}^{k} p_i - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}\right] \geq 0.$$

The left part between brackets of the above inequality equals

$$\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} + \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} d_{ij}$$

whereas the right part between brackets is always positive. This completes the proof. □

17.4 Bounds

At this point it seems appropriate to compare some of the multivariate formulations presented in this chapter with the corresponding multivariate generalizations from the previous chapter. As it turns out, the different formulations are bounds of each other. In Proposition 17.6 the multivariate formulation $S^{(k)}_{\mathrm{GL2}}(\theta)$ of parameter family $S_{\mathrm{GL2}}(\theta)$ from Chapter 16, is compared to multivariate extension $S^{(k)*}_{\mathrm{GL2}}(\theta)$ presented in this chapter.

Proposition 17.6. *It holds that* $S^{(k)}_{\mathrm{GL2}}(\theta) \leq S^{(k)*}_{\mathrm{GL2}}(\theta)$.
Proof: $S^{(k)}_{\mathrm{GL2}}(\theta) \leq S^{(k)*}_{\mathrm{GL2}}(\theta)$ if and only if

$$\frac{k(k-1)}{2}\left[1 - a^{(k)} - d^{(k)}\right] \geq (k-1)\sum_{i=1}^{k} p_i - 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}. \tag{17.6}$$

Note that

$$\frac{k(k-1)}{2}\, a^{(k)} \leq \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} \tag{17.7}$$

is true, because any $a_{ij} \geq a^{(k)}$ (in words: the proportion of 1s that two variables share in the same positions is always equal or greater than the proportion of 1s that the two variables and $k-2$ other variables share in the same position). Using similar arguments it holds that

$$\frac{k(k-1)}{2}\left[1 - d^{(k)}\right] \geq \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} (1 - d_{ij}). \tag{17.8}$$

Since

$$(k-1)\sum_{i=1}^{k} p_i - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} (1 - d_{ij}) \tag{17.9}$$

it follows that, adding $-1 \times$ (17.7) and (17.8) gives (17.6). Since both (17.7) and (17.8) hold, (17.6) is true. This completes the proof. □

In Proposition 17.7 the multivariate formulation $S^{(k)}_{\mathrm{GL1}}(\theta)$ of parameter family $S_{\mathrm{GL1}}(\theta)$ from Chapter 16, is compared to multivariate extension $S^{(k)*}_{\mathrm{GL1}}(\theta)$ presented in this chapter. Some properties derived in the proof of Proposition 17.6 are used in the proof of Proposition 17.7.

Proposition 17.7. *It holds that* $S^{(k)}_{\mathrm{GL1}}(\theta) \leq S^{(k)*}_{\mathrm{GL1}}(\theta)$.
Proof: Using some algebra, we obtain $S^{(k)}_{\mathrm{GL1}}(\theta) \leq S^{(k)*}_{\mathrm{GL1}}(\theta)$ if and only if

$$\left[1 - d^{(k)}\right]\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} \leq a^{(k)}\left[(k-1)\sum_{i=1}^{k} p_i - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}\right]. \tag{17.10}$$

Using (17.9), (17.10) can be written as

$$\frac{1-d^{(k)}}{a^{(k)}} \geq \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(1-d_{ij})}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}a_{ij}}. \tag{17.11}$$

Equation (17.11) holds if (17.7) and (17.8) are true. This completes the proof. □

Proposition 17.6 and Proposition 17.7 consider two families of coefficients that are linear in both numerator and denominator. It follows from both assertions that for these rational functions the multivariate formulation from Chapter 16 is equal or smaller compared to the multivariate formulation of the same coefficient presented in this chapter.

Three different multivariate generalizations of S_{Gleas} may be found in Chapter 16 and 17. From Proposition 17.7 it follows that $S_{\text{Gleas}}^{(k)**} \geq S_{\text{Gleas}}^{(k)}$. Proposition 17.8 is used to show that multivariate formulation $S_{\text{Gleas}}^{(k)**}$ is also equal to or greater than $S_{\text{Gleas}}^{(k)*}$. Which is the largest of $S_{\text{Gleas}}^{(k)}$ or $S_{\text{Gleas}}^{(k)*}$ depends on the data.

Proposition 17.8. *It holds that* $S_{\text{Gleas}}^{(k)**} \geq S_{\text{Gleas}}^{(k)*}$.
Proof: $S_{\text{Gleas}}^{(k)**} \geq S_{\text{Dice}}^{(k)*}$ if and only if (17.7) holds. □

17.5 Epilogue

In Chapter 4 it was shown that various coefficients become equivalent after correction for similarity due to chance. Similar to Chapter 16, this chapter was used to present multivariate formulations of various similarity coefficients. First, family $\mathcal{L}$ of coefficients that are of the form $\lambda+\mu a$, was extended to a family $\mathcal{L}^{(k)}$ of multivariate coefficients. The new family of multivariate coefficients preserves the properties derived for the $\mathcal{L}$ family in Chapter 4. For example, multivariate formulation for S_{SM} presented in this chapter is given by

$$S_{\text{SM}}^{(k)*} = 1 + \frac{4}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}a_{ij} - \frac{2}{k}\sum_{i=1}^{k}p_i.$$

Coefficient $S_{\text{Gleas}}^{(k)**}$ and $S_{\text{SM}}^{(k)*}$ become $S_{\text{Cohen}}^{(k)}$ after correction for chance agreement.

The heuristic used for coefficients in the $\mathcal{L}^{(k)}$ family can also be used for coefficients not in the $\mathcal{L}^{(k)}$ family. For example, the multivariate extension of S_{Jac} is given by

$$S_{\text{Jac}}^{(k)*} = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}a_{ij}}{(k-1)\sum_{i=1}^{k}p_i - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}a_{ij}}.$$

A multivariate coefficient that can be found in Loevinger (1947, 1948), Mokken (1971) and Sijtsma and Molenaar (2002), which is also based on this heuristic, is given by

$$S_{\text{Loe}}^{(k)} = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} - p_i p_j)}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \min(p_j q_k, p_k q_j)}.$$

Coefficient $S_{\text{Loe}}^{(k)}$ is a multivariate version of the two-way coefficient S_{Loe}. The multivariate coefficient $S_{\text{Loe}}^{(k)}$ uses the same heuristic as the other coefficients in this chapter, and the coefficient may be used to measure the homogeneity of k test items. Note that the generalization of Proposition 5.4 to $S_{\text{Loe}}^{(k)}$ is straightforward.

In Section 17.4 we showed how the multivariate coefficients presented in this chapter are related to the multivariate coefficients discussed in Chapter 16. Proposition 17.6 and Proposition 17.7 consider two parameter families of coefficients that are linear in both numerator and denominator. It follows from both assertions that for these rational functions the multivariate formulation from Chapter 16 is equal to or smaller than the multivariate formulation of the same coefficient presented in this chapter.

In Section 17.2 a multivariate formulation of Cohen's kappa (S_{Cohen}) was presented. The multivariate kappa ($S_{\text{Cohen}}^{(k)}$) was formulated for the case of two categories. The extension to the case of two or more categories is straightforward. As it turns out, the formulation of $S_{\text{Cohen}}^{(k)}$ for two or more categories is also proposed in both Popping (1983a) and Heuvelmans and Sanders (1993). Both authors have some form of motivation for why this multivariate kappa should be preferred over other multivariate generalizations of Cohen's kappa. However, it appears that the properties of $S_{\text{Cohen}}^{(k)}$ presented here are the first to provide a convincing argument.

In Section 2.2 the equivalence between Cohen's kappa S_{Cohen} and the Hubert-Arabie adjusted Rand index S_{HA} was established. Note that $S_{\text{Cohen}}^{(k)}$ would be an appropriate multivariate formulation of the the adjusted Rand index. Then, when comparing partitions of three ($k = 3$) cluster algorithms we do not require the three-way matching table. Instead we need to obtain the three two-way matching tables and then summarize these matching tables in three fourfold tables. Each 2×2 contingency table contains the four different types of pairs from two clustering methods.

CHAPTER 18

Metric properties of multivariate coefficients

In Chapter 10 metric properties were studied of two-way dissimilarity coefficients corresponding to various similarity coefficients. The dissimilarity coefficients were obtained from the transformation $D = 1 - S$, D is the complement of S. In the present chapter metric properties of the multivariate formulations of the two-way coefficients from Chapter 10 are considered. Each dissimilarity coefficient of Chapter 10 satisfies the triangle inequality. In this chapter metric properties with respect to the polyhedral generalization of the triangle inequality noted by De Rooij (2001, p. 128) are studied. The polyhedral inequality is given by

$$(k-1) \times D(x_{1,k}) \leq \sum_{i=1}^{k} D(x_{1,k+1}^{-i}) \tag{18.1}$$

for $k \geq 3$. Inequality (18.1) is also presented in (12.4), (14.13) and (15.2). In Chapter 14 several functions were studied that satisfy polyhedral inequality (18.1).

In Chapter 10 only a few dissimilarities obtained from the transformations $D = 1 - S$ turned out to be metric, that is, satisfied the triangle inequality. The present chapter is limited to multivariate generalizations of two-way coefficients that satisfy the triangle inequality. Before considering any metric properties, the following notation is defined. Let $P\left(x^1_{1,k}\right)$ denote the proportion of 1s in variables x_1 to x_k. Furthermore, let $P\left(x^{1,0,1}_{1,i,k}\right)$ denote the proportion of 1s in variables x_1 to x_k and 0 in variable x_i. Moreover, denote by $P\left(x^{1,-,1}_{1,i,k}\right)$ the proportion of 1s in variables x_1 to x_k where x_i drops out. An important property of the proportions in this notation is that

$$P\left(x^{1,-,1}_{1,i,k}\right) = P\left(x^1_{1,k}\right) + P\left(x^{1,0,1}_{1,i,k}\right). \tag{18.2}$$

18.1 Russel-Rao coefficient

In this section the metric properties of two multivariate formulations of S_{RR} are studied. In Chapter 16 we encountered the Bennani-Heiser multivariate coefficient

$$S^{(k)}_{\text{RR}} = a^{(k)} = P\left(x^1_{1,k}\right).$$

The second multivariate formulation of S_{RR} can be obtained from the heuristics considered in Chapter 17. This multivariate coefficient is given by

$$S^{(k)*}_{\text{RR}} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} a_{ij}.$$

The quantity $2/k(k-1)$ in the definition of $S^{(k)*}_{\text{RR}}$ is used to ensure that $0 \leq S^{(k)*}_{\text{RR}} \leq 1$. Both Proposition 18.1 and 18.2 are generalizations of the first part of Theorem 10.1. In Proposition 18.1 the metric property of $1 - S^{(k)}_{\text{RR}}$ is considered. The proof is a generalization of the tool presented in Heiser and Bennani (1997, p. 197) for $k = 3$.

Proposition 18.1. *The function*

$$1 - S^{(k)}_{\text{RR}} = 1 - P\left(x^1_{1,k}\right)$$

satisfies (18.1).
Proof: Using $1 - S^{(k)}_{\text{RR}}$ in (18.1) we obtain

$$(k-1) - (k-1)P\left(x^1_{1,k}\right) \leq k - \sum_{i=1}^{k} P\left(x^{1,-,1}_{1,i,k+1}\right)$$

which equals

$$1 + (k-1)P\left(x^1_{1,k}\right) \geq \sum_{i=1}^{k} P\left(x^{1,-,1}_{1,i,k+1}\right). \tag{18.3}$$

Using the property in (18.2), (18.3) becomes

$$1 + (k-1)P\left(x_{1,k}^1, x_{k+1}^1\right) + (k-1)P\left(x_{1,k}^1, x_{k+1}^0\right) \geq kP\left(x_{1,k}^1\right) + \sum_{i=1}^{k} P\left(x_{1,i,k+1}^{1,0,1}\right)$$

which equals

$$1 + (k-1)P\left(x_{1,k}^1, x_{k+1}^0\right) \geq P\left(x_{1,k+1}^1\right) + \sum_{i=1}^{k} P\left(x_{1,i,k}^{1,0,1}\right). \tag{18.4}$$

The fact that 1 is equal or larger than the right part of inequality (18.4) completes the proof. □

In Proposition 18.2 the metric property of $1 - S_{\mathrm{RR}}^{(k)*}$ is considered. The first proof of the assertion is an application of Proposition 14.4 together with the first part of Theorem 10.1. The second proof is a direct proof of the assertion.

Proposition 18.2. *The function*

$$1 - S_{\mathrm{RR}}^{(k)*} = 1 - \frac{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij}}{k(k-1)}$$

satisfies (18.1).
Proof 1: By Proposition 14.4, the sum of $k(k-1)/2$ quantities $(1 - a_{ij})$ satisfies (18.1), if each quantity $(1 - a_{ij})$ satisfies the triangle inequality. The first part of Theorem 10.1 shows that this is the case.
Proof 2: Using $1 - S_{\mathrm{RR}}^{(k)*}$ in (18.1) we obtain the inequality

$$\frac{k(k-1)}{2} + \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} \geq (k-1)\sum_{i=1}^{k} a_{ik+1}. \tag{18.5}$$

It holds that

$$\begin{aligned}\frac{k(k-1)}{2} &\geq (k-1)\sum_{i=1}^{k} a_{ik+1} \\ &\quad - \left[\frac{k(k-1)}{2}\right] P\left(x_{1,k+1}^1\right) - \left[\frac{(k-1)(k-2)}{2}\right] P\left(x_1^0, x_{2,k+1}^1\right).\end{aligned}$$

Furthermore, it holds that

$$\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_{ij} \geq \left[\frac{k(k-1)}{2}\right] P\left(x_{1,k+1}^1\right) + \left[\frac{(k-1)(k-2)}{2}\right] P\left(x_1^0, x_{2,k+1}^1\right).$$

Thus, inequality (18.5) holds, which completes the proof. □

18.2 Simple matching coefficient

In this section the metric properties of two multivariate formulations of S_{SM} are studied. In Chapter 16 we encountered the Bennani-Heiser multivariate formulation of S_{SM} which is given by

$$S_{\text{SM}}^{(k)} = a^{(k)} + d^{(k)} = P\left(x_{1,k}^{1}\right) + P\left(x_{1,k}^{0}\right).$$

The second multivariate formulation of S_{SM} was presented in Chapter 17 and is given by

$$S_{\text{SM}}^{(k)*} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij}).$$

Both Proposition 18.3 and 18.4 are generalizations of the second part of Theorem 10.1. In Proposition 18.3 the metric property of $1 - S_{\text{SM}}^{(k)}$ is considered. The proof is a generalization of the tool presented in Heiser and Bennani (1997, p. 196) for $k = 3$.

Proposition 18.3. *The function*

$$1 - S_{\text{SM}}^{(k)} = 1 - P\left(x_{1,k}^{1}\right) - P\left(x_{1,k}^{0}\right)$$

satisfies (18.1).
Proof: Using $1 - S_{\text{SM}}^{(k)}$ in (18.1) gives

$$(k-1) - (k-1)P\left(x_{1,k}^{1}\right) - (k-1)P\left(x_{1,k}^{0}\right) \leq k - \sum_{i=1}^{k} P\left(x_{1,i,k+1}^{1,-,1}\right) - \sum_{i=1}^{k} P\left(x_{1,i,k+1}^{0,-,0}\right)$$

which equals

$$1 + (k-1)P\left(x_{1,k}^{1}\right) + (k-1)P\left(x_{1,k}^{0}\right) \geq \sum_{i=1}^{k} P\left(x_{1,i,k+1}^{1,-,1}\right) + \sum_{i=1}^{k} P\left(x_{1,i,k+1}^{0,-,0}\right). \tag{18.6}$$

Using (18.2), (18.6) becomes

$$\begin{aligned}(k-1)\left[P\left(x_{1,k}^{1}, x_{k+1}^{1}\right) + P\left(x_{1,k}^{1}, x_{k+1}^{0}\right) + P\left(x_{1,k}^{0}, x_{k+1}^{1}\right) + P\left(x_{1,k}^{0}, x_{k+1}^{0}\right)\right] + \\ 1 \;\geq\; kP\left(x_{1,k+1}^{1}\right) + kP\left(x_{1,k+1}^{0}\right) + \sum_{i=1}^{k} P\left(x_{1,i,k+1}^{1,0,1}\right) + \sum_{i=1}^{k} P\left(x_{1,i,k+1}^{0,1,0}\right)\end{aligned}$$

which equals

$$1 + (k-1)P\left(x^1_{1,k}, x^0_{k+1}\right) + (k-1)P\left(x^0_{1,k}, x^1_{k+1}\right) \geq$$
$$P\left(x^1_{1,k+1}\right) + P\left(x^0_{1,k+1}\right) + \sum_{i=1}^{k} P\left(x^{1,0,1}_{1,i,k}\right) + \sum_{i=1}^{k} P\left(x^{0,1,0}_{1,i,k}\right). \tag{18.7}$$

The fact that 1 is equal or larger than the right part of inequality (18.7) proves the assertion. □

The metric property of $1 - S^{(k)*}_{\mathrm{SM}}$ is presented in Proposition 18.4. The first proof of the assertion is an application of Proposition 14.4 together with the second part of Theorem 10.1. The second proof is a direct proof of the assertion.

Proposition 18.4. *The function*

$$1 - S^{(k)*}_{\mathrm{SM}} = 1 - \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij})$$

satisfies (18.1).
Proof 1: By Proposition 14.4, the sum of $k(k-1)/2$ quantities $(1-a_{ij}-d_{ij})$ satisfies (18.1), if each quantity $(1 - a_{ij} - d_{ij})$ satisfies the triangle inequality. The second part of Theorem 10.1 shows that this is the case.
Proof 2: Filling in $1 - S^{(k)*}_{\mathrm{SM}}$ in (18.1) we obtain the inequality

$$\frac{k(k-1)}{2} + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij}) \geq (k-1) \sum_{i=1}^{k} (a_{ik+1} + d_{ik+1}). \tag{18.8}$$

It holds that

$$\begin{aligned} \frac{k(k-1)}{2} \geq\ & (k-1) \sum_{i=1}^{k} (a_{ik+1} + d_{ik+1}) \\ & - \left[\frac{k(k-1)}{2}\right] \left[P\left(x^1_{1,k+1}\right) + P\left(x^0_{1,k+1}\right)\right] \\ & - \left[\frac{(k-1)(k-2)}{2}\right] \left[P\left(x^0_1, x^1_{2,k+1}\right) + P\left(x^1_1, x^0_{2,k+1}\right)\right]. \end{aligned}$$

Furthermore, it holds that

$$\begin{aligned} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (a_{ij} + d_{ij}) \geq\ & \left[\frac{k(k-1)}{2}\right] \left[P\left(x^1_{1,k+1}\right) + P\left(x^0_{1,k+1}\right)\right] \\ & + \left[\frac{(k-1)(k-2)}{2}\right] \left[P\left(x^0_1, x^1_{2,k+1}\right) + P\left(x^1_1, x^0_{2,k+1}\right)\right]. \end{aligned}$$

Thus, inequality (18.8) holds, which completes the proof. □

18.3 Jaccard coefficient

In this final section the metric properties of multivariate formulations of the Jaccard (1912) coefficient S_{Jac} and the parameter family $S_{\text{GL1}}(\theta)$ are studied. In Chapter 16 we encountered the Bennani-Heiser multivariate formulation of S_{Jac} given by

$$S_{\text{Jac}}^{(k)} = \frac{a^{(k)}}{1 - d^{(k)}} = \frac{P\left(x_{1,k}^{1}\right)}{1 - P\left(x_{1,k}^{0}\right)}.$$

In Proposition 18.5 the metric property of $1 - S_{\text{Jac}}^{(k)}$ is considered. The proof is a generalization of the proof used in the first part of Theorem 10.2. In the proof, the relation between multivariate coefficients $S_{\text{SM}}^{(k)}$ and $S_{\text{Jac}}^{(k)}$ given by

$$1 - S_{\text{SM}}^{(k)} = \left[1 - P\left(x_{1,k}^{0}\right)\right]\left[1 - S_{\text{Jac}}^{(k)}\right] \tag{18.9}$$

is used.

Proposition 18.5. *The function*

$$1 - S_{\text{Jac}}^{(k)} = 1 - \frac{P\left(x_{1,k}^{1}\right)}{1 - P\left(x_{1,k}^{0}\right)}$$

satisfies (18.1).
Proof: It holds that

$$1 \geq P\left(x_{1,k+1}^{1}\right) + \sum_{i=1}^{k+1} P\left(x_{1,i,k+1}^{1,0,1}\right) + P\left(x_{1,k+1}^{0}\right) + \sum_{i=1}^{k+1} P\left(x_{1,i,k+1}^{0,1,0}\right). \tag{18.10}$$

Note that for $k = 2$, inequality (18.10) becomes an equality. Adding

$$(k-1)\left[P\left(x_{1,k}^{1}, x_{k+1}^{0}\right) + P\left(x_{1,k}^{0}, x_{k+1}^{1}\right)\right]$$

to both sides of (18.10), the inequality can be written as

$$\sum_{i=1}^{k}\left[1 - S_{\text{SM}}^{(k)}\left(x_{1,k+1}^{-i}\right)\right] - (k-1)\left[1 - S_{\text{SM}}^{(k)}\left(x_{1,k}\right)\right] \geq k\left[P\left(x_{1,k}^{1}, x_{k+1}^{0}\right) + P\left(x_{1,k}^{0}, x_{k+1}^{1}\right)\right]. \tag{18.11}$$

Using (18.9) in (18.11) we obtain

$$\begin{aligned}\left[1 - P\left(x_{1,k+1}^{0}\right)\right] \times &\left(\sum_{i=1}^{k}\left[1 - S_{\text{Jac}}^{(k)}\left(x_{1,k+1}^{-i}\right)\right] - (k-1)\left[1 - S_{\text{Jac}}^{(k)}\left(x_{1,k}\right)\right]\right) \\ \geq & kP\left(x_{1,k}^{1}, x_{k+1}^{0}\right) + \sum_{i=1}^{k}\left[1 - S_{\text{Jac}}^{(k)}\left(x_{1,k+1}^{-i}\right)\right] P\left(x_{1,i,k+1}^{0,1,0}\right) \\ & + P\left(x_{1,k}^{0}, x_{k+1}^{1}\right)\left[1 + (k-1) S_{\text{Jac}}^{(k)}\left(x_{1,k}\right)\right].\end{aligned}$$

With respect to the first term of the inequality $P\left(x_{1,k+1}^{0}\right) \leq 1$. Hence, we conclude that $1 - S_{\text{Jac}}^{(k)}$ satisfies (18.1). □

We end this chapter with a generalization of Theorem 10.4. From Chapter 16 we obtain the multivariate formulation of parameter family $S_{\mathrm{GL1}}(\theta)$, which is given by

$$S_{\mathrm{GL1}}^{(k)}(\theta) = \frac{P\left(x_{1,k}^{1}\right)}{(1-\theta)P\left(x_{1,k}^{1}\right) + \theta\left[1 - P\left(x_{1,k}^{0}\right)\right]}.$$

In Proposition 18.6 the metric property of $1-S_{\mathrm{GL1}}^{(k)}(\theta)$ is considered. In order to proof the assertion, the result in Proposition 18.5 on $1 - S_{\mathrm{Jac}}^{(k)}$ is used. With respect to the proof of Proposition 18.6 it assumed that Conjecture 15.1, which is a generalization of Theorem 10.3, is true. We have the following metric property with respect to $1 - S_{\mathrm{GL1}}^{(k)}(\theta)$.

Proposition 18.6. *The function*

$$1 - S_{\mathrm{GL1}}^{(k)}(\theta) = 1 - \frac{P\left(x_{1,k}^{1}\right)}{(1-\theta)P\left(x_{1,k}^{1}\right) + \theta\left[1 - P\left(x_{1,k}^{0}\right)\right]} \tag{18.12}$$

satisfies (18.1) *for* $0 < \theta \leq 1$.
Proof: By Proposition 18.5 $1-S_{\mathrm{GL1}}^{(k)}(\theta = 1) = 1-S_{\mathrm{Jac}}^{(k)}$ satisfies (18.1). For $0 < \theta < 1$, let $\theta = (c+1)/c$ where c is a strictly positive real number. Equation (18.12) equals

$$\frac{\theta\left[1 - S_{\mathrm{SM}}^{(k)}\right]}{P\left(x_{1,k}^{1}\right) + \theta\left[1 - S_{\mathrm{SM}}^{(k)}\right]} = \frac{(c+1)\left[1 - S_{\mathrm{SM}}^{(k)}\right]}{cP\left(x_{1,k}^{1}\right) + (c+1)\left[1 - S_{\mathrm{SM}}^{(k)}\right]}. \tag{18.13}$$

Dividing both numerator and denominator of (18.13) by $1 - P\left(x_{1,k}^{0}\right)$ we obtain

$$1 - S_{\mathrm{GL1}}^{(k)}(\theta) = \frac{(c+1)\left[1 - S_{\mathrm{Jac}}^{(k)}\right]}{cS_{\mathrm{Jac}}^{(k)} + (c+1)\left[1 - S_{\mathrm{Jac}}^{(k)}\right]} = \frac{(c+1)\left[1 - S_{\mathrm{Jac}}^{(k)}\right]}{c + 1 - S_{\mathrm{Jac}}^{(k)}}. \tag{18.14}$$

Because $1 - S_{\mathrm{Jac}}^{(k)}$ satisfies (18.1) due to Proposition 18.5, the result follows if Conjecture 15.1 is valid.

18.4 Epilogue

In this chapter metric properties of several multivariate coefficients were presented. Each of the functions satisfies the strong polyhedral inequality (18.1), which is a generalization formulated by De Rooij (2001) of the tetrahedral inequality considered in Heiser and Bennani (1997). Although no well-established multi-way metric structure emerged from the study in Chapter 12, we have gathered several interesting properties of the polyhedral inequality in some of the chapters following Chapter 12. In Chapter 13 it was shown that the polyhedral inequality was the strongest multi-way metric implied by the an ultrametric. In Chapter 14 we formulated multi-way extensions of two three-way functions that satisfy this polyhedral inequality. In this particular chapter it was shown that several multivariate coefficients from Chapters 16 and 17 also satisfy the polyhedral inequality (18.1). So far, the preliminary results in these chapters suggest that the inequality is definitely the most interesting multi-way generalization of the triangle inequality.

CHAPTER 19

Robinson cubes

Robinson matrices were studied in Chapter 7. In this chapter the three-way generalization of the Robinson matrix is studied, which will be referred to as a Robinson cube. Whereas a matrix is characterized by rows and columns, a cube consists of rows, columns and pillars. A cube has six faces. The twelve rows, columns and pillars where two faces cross are called the edges. The eight entries where three edges meet are called the vertices of the cube. Some aspects of a cube are demonstrated in Figure 19.1.

First some definitions of a Robinson cube are presented. A similarity cube is called a Robinson cube if the highest entries within each row, column and pillar are on the main diagonal and moving away from this diagonal, the entries never increase. Next, it is considered what three-way functions and similarity coefficients satisfy these definitions.

[0]This chapter appeared in a slightly adapted version in Warrens, M.J. and Heiser, W.J. (2007), Robinson Cubes, in P. Brito, P. Bertrand, G. Cucumel and F. de Caravalho (Eds.), *Selected Contributions in Data Analysis and Classification*, 515–523, Berlin: Springer.

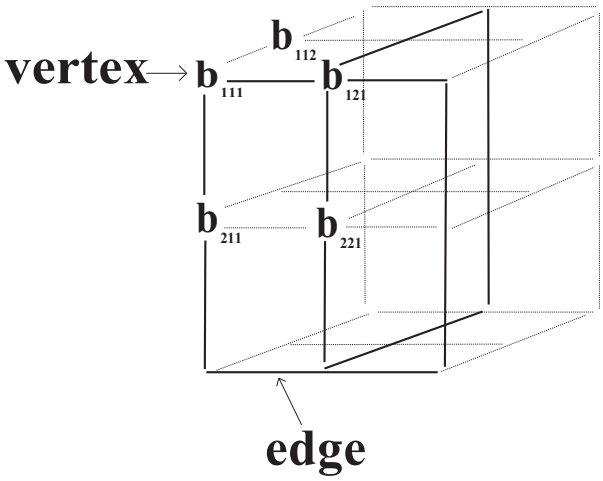

Figure 19.1: *Several aspects of a cube.*

19.1 Definitions

Before defining a Robinson cube we turn our attention to two natural requirements for cubes. Similar to a matrix, we may require that a similarity cube $\mathbf{S}^{(3)}$ satisfies three-way symmetry, that is,

$$\begin{aligned} S(x_1, x_2, x_3) = S(x_1, x_3, x_2) = S(x_2, x_1, x_3) \\ = S(x_2, x_3, x_1) = S(x_3, x_1, x_2) = S(x_3, x_2, x_1) \end{aligned}$$

for all x_1, x_2 and x_3. Another natural requirement for a similarity cube is the restriction

$$S(x_1, x_2, x_1) = S(x_1, x_2, x_2) \qquad \text{for all } x_1 \text{ and } x_2. \tag{19.1}$$

This requirement together with three-way symmetry implies the so-called diagonal-plane equality (Section 11.2; Heiser and Bennani, 1997, p. 191) which requires equality of the three matrices defined by the elements $S(x_1, x_1, x_2)$, $S(x_1, x_2, x_1)$ and $S(x_1, x_2, x_2)$, that are formed by cutting the cube diagonally, starting at one of the three edges joining at the vertex $S(1, 1, 1)$. A weak extension of the Robinson matrix is the following definition.

A similarity cube $\mathbf{S}^{(3)}$ is called a Robinson cube if the highest entries within each row, column and tube are on the main diagonal (elements $S(x_1, x_1, x_1)$) and moving away from this diagonal, the entries never increase.

Hence, $\mathbf{S}^{(3)}$ of size $m \times m \times m$ is a Robinson cube if

$$1 \leq x_1 < x_2 \leq m \;\Rightarrow\; \begin{cases} S(x_1, x_2, x_2) \leq S(x_1 + 1, x_2, x_2) \\ S(x_2, x_1, x_2) \leq S(x_2, x_1 + 1, x_2) \\ S(x_2, x_2, x_1) \leq S(x_2, x_2, x_1 + 1) \end{cases}$$

$$1 \leq x_2 < x_1 \leq m \;\Rightarrow\; \begin{cases} S(x_1, x_2, x_2) \geq S(x_1 + 1, x_2, x_2) \\ S(x_2, x_1, x_2) \geq S(x_2, x_1 + 1, x_2) \\ S(x_2, x_2, x_1) \geq S(x_2, x_2, x_1 + 1). \end{cases}$$

If the cube $\mathbf{S}^{(3)}$ satisfies the requirement in (19.1), then $\mathbf{S}^{(3)}$ is a Robinson matrix if we have

$$1 \leq x_1 < x_2 \leq m \;\Rightarrow\; \begin{cases} S(x_1, x_2, x_2) \leq S(x_1 + 1, x_2, x_2) \\ S(x_2, x_1, x_2) \leq S(x_2, x_1 + 1, x_2) \end{cases}$$

$$1 \leq x_2 < x_1 \leq m \;\Rightarrow\; \begin{cases} S(x_1, x_2, x_2) \geq S(x_1 + 1, x_2, x_2) \\ S(x_2, x_1, x_2) \geq S(x_2, x_1 + 1, x_2). \end{cases}$$

Moreover, if the cube $\mathbf{S}^{(3)}$ satisfies three-way symmetry, then $\mathbf{S}^{(3)}$ is a Robinson cube if we have

$$\begin{aligned} 1 \leq x_1 < x_2 \leq m &\;\Rightarrow\; S(x_1, x_2, x_2) \leq S(x_1 + 1, x_2, x_2) \\ 1 \leq x_2 < x_1 \leq m &\;\Rightarrow\; S(x_1, x_2, x_2) \geq S(x_1 + 1, x_2, x_2). \end{aligned}$$

For the definition of a dissimilarity cube $\mathbf{D}^{(3)}$ the roles of $\leq$ and $\geq$ in the comparisons involving cube elements must be interchanged. Note that, although this is perhaps suggested in the above arguments, a Robinson cube that satisfies three-way symmetry does not necessarily satisfy requirement (19.1). In the above definition of a Robinson cube not all entries are involved. More precisely, only those entries that are in a row, column or pillar with an entry of the main diagonal are involved. A stronger definition of a Robinson cube is the following.

A cube $\mathbf{S}^{(3)}$ is called a regular Robinson cube if

1. $\mathbf{S}^{(3)}$ is a Robinson cube

2. all matrices, which are formed by cutting the cube perpendicularly, where for each matrix $\mathbf{S}^{(2)}$ entry $S^{(2)}(1,1)$ is an element of one of the three edges joining at the vertex $S^{(3)}(1,1,1)$ (with $S^{(2)}(1,1) = S^{(3)}(1,1,1)$ if $S^{(2)}(1,1)$ is one of the three faces joining at the vertex $S^{(2)}(1,1,1)$), are Robinson matrices.

A regular Robinson cube has some interesting features. For example, if $\mathbf{S}^{(3)}$ is a regular Robinson cube then it satisfies both three-way symmetry and the diagonal-plane equality. These properties become clear from the following result on the composition of a regular Robinson cube.

Proposition 19.1. *Let* $x_4 = \min(x_1, x_2, x_3)$ *and* $x_5 = \max(x_1, x_2, x_3)$. *If* $\mathbf{S}^{(3)}$ *is a regular Robinson cube, then its entries* $S^{(3)}(x_1, x_2, x_3)$ *equal*

$$S^{(3)}(x_4, x_6, x_5) = S^{(3)}(x_6, x_4, x_5) = S^{(3)}(x_4, x_5, x_6) =$$
$$S^{(3)}(x_6, x_5, x_4) = S^{(3)}(x_5, x_4, x_6) = S^{(3)}(x_5, x_6, x_4) \quad \text{for } x_6 = x_4, ..., x_5.$$

Proof: First let $\mathbf{S}$ be the front face of the cube, where $S^{(2)}(1,1) = S^{(3)}(1,1,1)$. Since $S^{(3)}(2,2,1)$ is a diagonal element of $\mathbf{S}$, $\mathbf{S}$ is a Robinson matrix if $S^{(3)}(1,2,1) \leq S^{(3)}(2,2,1)$. Next let $\mathbf{S}$ be the cutting perpendicular on the front face of the cube, with $S^{(2)}(1,1) = S^{(3)}(1,2,1)$. Since $S^{(3)}(1,2,1)$ is a diagonal element of $\mathbf{S}$, the latter is a Robinson matrix if $S^{(3)}(1,2,1) \geq S^{(3)}(2,2,1)$. Thus, if $\mathbf{S}^{(3)}$ is a regular Robinson cube, then $S^{(3)}(1,2,1) = S^{(3)}(2,2,1)$ $(= S^{(3)}(2,1,1) = S^{(3)}(2,1,2) = S^{(3)}(1,1,2) = S^{(3)}(1,2,2))$. □

19.2 Functions

Let $D(x_1, x_2, x_3)$ denote a three-way dissimilarity. One of the more popular functions for three-way dissimilarities used in classification literature are the symmetric L_p-transforms defined as

$$D(x_1, x_2, x_3) = ([D(x_1, x_2)]^p + [D(x_1, x_3)]^p + [D(x_2, x_3)]^p)^{1/p}.$$

For instance, for $p = 1$ we have the perimeter function, for $p = 2$ the generalized Euclidean function. For $p = \infty$ we obtain the generalized dominance function or maximum distance (Section 14.4)

$$D(x_1, x_2, x_3) = \max[D(x_1, x_2), D(x_1, x_3), D(x_2, x_3)].$$

Somewhat lesser known is the variance function (De Rooij and Gower, 2003, p. 188)

$$\begin{aligned}[D(x_1, x_2, x_3)]^2 =& \text{var}[D(x_1, x_2), D(x_1, x_3), D(x_2, x_3)] \\ =& ([D(x_1, x_2)]^2 + [D(x_1, x_3)]^2 + [D(x_2, x_3)]^2) \\ & -\frac{1}{3}[D(x_1, x_2) + D(x_1, x_3) + D(x_2, x_3)]^2.\end{aligned}$$

The variance function is symmetric in x_1, x_2 and x_3.

Proposition 19.2. *Suppose* $D(x_1, x_2, x_3)$ *is defined as a* L_p*-transform or equals the variance function. Then the cube* $\mathbf{D}^{(3)}$ *with elements* $D(x_1, x_2, x_3)$ *is a Robinson cube if and only if the matrix* $\mathbf{D}$ *with elements* $D(x_1, x_2)$ *is a Robinson matrix.*
Proof: For $1 \leq x_1 < x_2 \leq m$ with respect to any L_p-transform, we have

$$D(x_1, x_2, x_2) = (2[D(x_1, x_2]^p)^{1/p} \geq (2D[x_1 + 1, x_2]^p)^{1/p} = D(x_1 + 1, x_2, x_2)$$

if and only if $D(x_1, x_2) \geq D(x_1 + 1, x_2)$.

For $1 \leq x_1 < x_2 \leq m$ with respect to the variance function, we have

$$\begin{aligned}[D(x_1, x_2, x_2)]^2 &= [2D(x_1, x_2)]^2 - \frac{1}{3}[2D(x_1, x_2)]^2 \\ &\geq [2D(x_1 + 1, x_2)]^2 - \frac{1}{3}[2D(x_1 + 1, x_2)]^2 \\ &= [D(x_1 + 1, x_2, x_2)]^2\end{aligned}$$

if and only if

$$\frac{2}{3}[D(x_1, x_2)]^2 \geq \frac{2}{3}[D(x_1, x_2)]^2 \quad \text{if and only if} \quad D(x_1, x_2) \geq D(x_1 + 1, x_2).$$

A similar property holds for $D(x_1, x_2, x_2) \leq D(x_1 + 1, x_2, x_2)$ for $1 \leq x_2 \leq x_1 < m$. □

A stronger result holds for the dominance function

$$D(x_1, x_2, x_3) = \max[D(x_1, x_2), D(x_1, x_3), D(x_2, x_3)] \quad \text{for dissimilarities}$$

or equivalently

$$S(x_1, x_2, x_3) = \min[S(x_1, x_2), S(x_1, x_3), S(x_2, x_3)] \quad \text{for similarities.}$$

Proposition 19.3. *Let* $\mathbf{S}$ *and* $\mathbf{S}^{(3)}$ *be respectively a similarity matrix and cube. If*

$$S(x_1, x_2, x_3) = \min[S(x_1, x_2), S(x_1, x_3), S(x_2, x_3)]$$

then $\mathbf{S}^{(3)}$ *is a regular Robinson cube if and only if* $\mathbf{S}$ *is a Robinson matrix.*
Proof: If $\mathbf{S}$ is a Robinson matrix then the minimum function satisfies

$$S(x_1, x_2, x_3) = \min\left[S(x_1, x_2), S(x_1, x_3), S(x_2, x_3)\right] = S(x_1, x_3)$$

for $1 \leq x_1 \leq x_2 \leq x_3 \leq m$, which demonstrates the second requirement of a regular Robinson cube. Moreover, we have

$$S(x_1, x_2, x_2) = S(x_1, x_2) \leq S(x_1 + 1, x_2) = S(x_1 + 1, x_2, x_2)$$

for $1 \leq x_1 < x_2 \leq m$, and

$$S(x_1, x_2, x_2) = S(x_1, x_2) \geq S(x_1 + 1, x_2) = S(x_1 + 1, x_2, x_2)$$

for $1 \leq x_2 \leq x_1 < m$, which demonstrates the first requirement of a regular Robinson cube. □

19.3 Coefficient properties

In this section it is shown for several three-way Bennani-Heiser similarity coefficients that the corresponding cube is a Robinson cube if and only if the matrix corresponding to the two-way similarity coefficient is a Robinson matrix. Let x_1, x_2 and x_3 be binary variables. Let $P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right)$ denote the proportion of 1s shared by x_1, x_2 and x_3 in the same positions. All matrices and cubes in this section are of the similarity kind. Yet, for all results below there exist an equivalent formulation in terms of dissimilarities.

Proposition 19.4 considers the Robinson property for the family $S_{\text{GL1}}(\theta)$ given by

$$S_{\text{GL1}}(\theta) = \frac{P\left(\overset{1}{x}_1, \overset{1}{x}_2\right)}{(1-\theta)P\left(\overset{1}{x}_1, \overset{1}{x}_2\right) + \theta\left[1 - P\left(\overset{0}{x}_1, \overset{0}{x}_2\right)\right]}.$$

The three-way generalization of $S_{\text{GL1}}(\theta)$ from Chapter 16 is given by

$$S^{(3)}_{\text{GL1}}(\theta) = \frac{P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right)}{(1-\theta)P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right) + \theta\left[1 - P\left(\overset{0}{x}_1, \overset{0}{x}_2, \overset{0}{x}_2\right)\right]}.$$

Proposition 19.4. *The cube $\mathbf{S}^{(3)}_{\text{GL1}}$ with elements $S^{(3)}_{\text{GL1}}(\theta)$ for some θ is a Robinson cube if and only if the matrix $\mathbf{S}_{\text{GL1}}$ with elements $S^{(2)}_{\text{GL1}}(\theta)$ using the same θ is a Robinson matrix.*

Proof: Due to Proposition 16.1, the proof can be limited to a specific value of θ. $S^{(2)}_{\text{Jac}}(x_1, x_2) = S^{(2)}_{\text{GL1}}(\theta = 1)$ and $S^{(3)}_{\text{Jac}}(x_1, x_2, x_3) = S^{(3)}_{\text{GL1}}(\theta = 1)$. $S^{(3)}_{\text{Jac}}(x_1, x_2, x_3)$ can be written as

$$S^{(3)}_{\text{Jac}} = \frac{P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right)}{1 - P\left(\overset{0}{x}_1, \overset{0}{x}_2, \overset{0}{x}_2\right)}.$$

The result then follows from the property

$$S^{(2)}_{\text{Jac}}(x_1, x_2) = \frac{P\left(\overset{1}{x}_1, \overset{1}{x}_2\right)}{1 - P\left(\overset{0}{x}_1, \overset{0}{x}_2\right)} = S^{(3)}_{\text{Jac}}(x_1, x_2, x_2). \qquad \square$$

Proposition 19.5 considers the Robinson property for the matrix $\mathbf{S}_{\text{RR}}$ with elements

$$S_{\text{RR}}(x_1, x_2) = P\left(\overset{1}{x}_1, \overset{1}{x}_2\right).$$

The three-way generalization of $\mathbf{S}_{\text{RR}}$ from Chapter 15 is the cube $\mathbf{S}^{(3)}_{\text{RR}}$ with elements

$$S^{(3)}_{\text{RR}}(x_1, x_2, x_3) = P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right).$$

Proposition 19.5. *The following statements are equivalent:*

1. $\mathbf{S}_{\mathrm{RR}}$ *is a Robinson matrix*
2. $\mathbf{S}_{\mathrm{RR}}^{(3)}$ *is a regular Robinson cube*
3. $S_{\mathrm{RR}}^{(3)}(x_1, x_2, x_3) = \min\left[S_{\mathrm{RR}}(x_1, x_2), S_{\mathrm{RR}}(x_1, x_3), S_{\mathrm{RR}}(x_2, x_3)\right]$.

Proof: The result follows from the fact that $P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_2\right) = P\left(\overset{1}{x}_1, \overset{1}{x}_2\right)$ and if $\mathbf{S}_{\mathrm{RR}}$ is a Robinson matrix, then $P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right)$ has the property, for $1 \leq x_1 \leq x_2 \leq x_3 \leq m$, we have

$$P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right) = \min\left[P\left(\overset{1}{x}_1, \overset{1}{x}_2\right), P\left(\overset{1}{x}_1, \overset{1}{x}_3\right), P\left(\overset{1}{x}_2, \overset{1}{x}_3\right)\right] = P\left(\overset{1}{x}_1, \overset{1}{x}_3\right). \quad \square$$

A sufficient condition for $\mathbf{S}_{\mathrm{RR}}$ in Proposition 19.5 is given in Theorem 7.1. It follows from Proposition 19.5 that this condition is then also sufficient for $\mathbf{S}_{\mathrm{RR}}^{(3)}$ to be a Robinson cube. Alternatively, it is also possible to generalize the second proof of Theorem 7.1.

Proposition 19.6. *If* $\mathbf{X}$ *is row Petrie then* $\mathbf{S}_{\mathrm{RR}}^{(3)}$ *is a regular Robinson cube.*
Proof: For the sake of an example let $\mathbf{X}$ be given by

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

where x_1, x_2 and x_3 identify the columns of $\mathbf{X}$. The proof is further depicted in Figure 19.2. The first six cubes are the similarity cubes with elements $P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right)$ corresponding to the six rows of $\mathbf{X}$. If a column has consecutive 1s, the similarity cube corresponding to this row, is a Robinson cube. The seventh and last cube in Figure 19.2 is the cube with elements $P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right)$ for the complete table $\mathbf{X}$. Figure 19.2 visualizes an interesting property of regular Robinson cubes, that is, the sum of regular Robinson cubes is again a regular Robinson cube. $\square$

19.4 Epilogue

A data array arranged in a cube in which rows, columns and pillars refer to the same objects has been called three-way one-mode, or triadic data. Such data have been studied in attempts to identify higher order interactions among objects (Heiser and Bennani, 1997). In this chapter, we have shown that we can recognize a simple order among the objects in three-way data, by a generalization of the Robinson property

for two-way data. We have discussed a general version of the Robinson cube, and a more specific one. Studying several definitions of three-way (dis)similarities, we found that in most cases, if a two-way (dis)similarity is Robinsonian, then the triadic (dis)similarity is Robinsonian too. A regular Robinson cube occurs only with the Russel and Rao (1940) coefficient calculated on an attribute matrix with the consecutive 1s property, and with the dominance metric for dissimilarities.

This chapter was limited to Robinson cubes. For the three-way case, two definitions of a Robinson cube may be adopted, one is a special case of the other. As it turns out, similar to the multi-way ultrametrics in Chapter 13, for the four-way case up to three definitions of a Robinson 4-cube or a Robinson tesseract can be given.

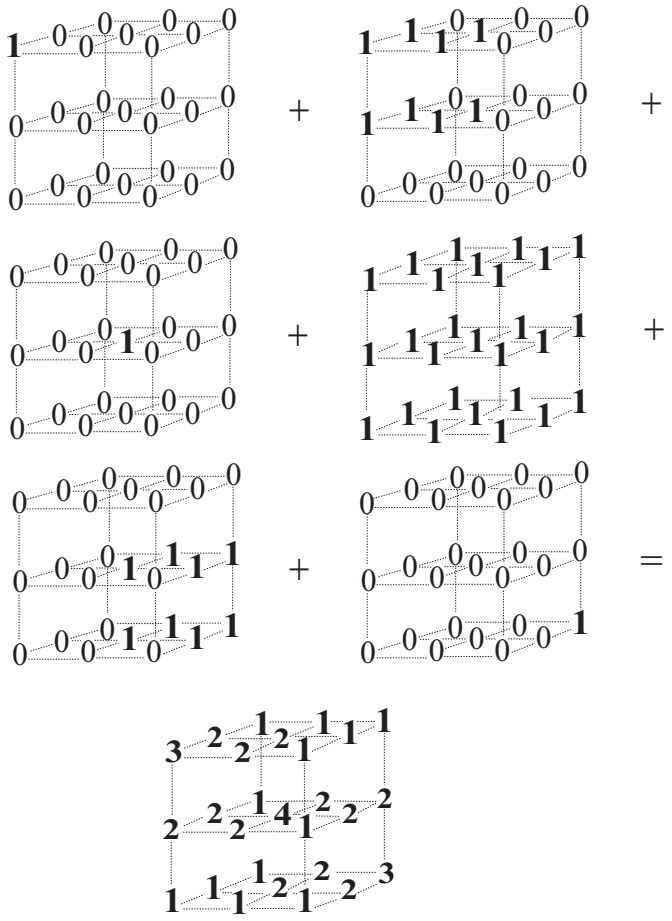

Figure 19.2: *The sum of the six regular Robinson cubes is a regular Robinson cube.*