Similarity Coefficients for Binary Data

Warrens, Matthijs Joost,
Similarity Coefficients for Binary Data. Properties of Coefficients, Coefficient Matrices, Multi-way Metrics and Multivariate Coefficients
Dissertation Leiden University – With References – With Summary in Dutch.

Subject headings: Association Measures; Correction for Chance; Correction for Maximum Value; Homogeneity Analysis; k-Way Metricity.

ISBN 978-90-8891-0524

Printed by Proefschriftmaken.nl, Oisterwijk

Similarity Coefficients for Binary Data

Properties of Coefficients, Coefficient Matrices, Multi-way Metrics and Multivariate Coefficients

PROEFSCHRIFT

ter verkrijging van de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 25 juni 2008
klokke 13.45 uur

door

Matthijs Joost Warrens

geboren te Rotterdam
in 1978

PROMOTIECOMMISSIE

Promotor	Prof. dr. W.J. Heiser
Co-promotor	Dr. D.N.M. de Gruijter
Referent	Prof. dr. J.C. Gower, The Open University, Milton Keynes, UK
Overige Leden	Prof. dr. J.J. Meulman Prof. dr. H.A.L. Kiers, University of Groningen Dr. M. de Rooij

To Sascha

Acknowledgments

There are several people I would like to thank for their help in one way or another during the time of writing this dissertation: Laurence Frank, for showing me around when I toke my first steps as a PhD student, and for the many conversations on the various aspects of research; and Marike Polak, with whom I shared a room for such a substantial period, for her companionship and the many talks on Dutch soccer and life in general. I also thank Marian Hickendorff and Susañña Verdel for just being there.

Contents

Prologue

A variety of data can be represented in strings of binary scores. In general, the binary scores reflect either the presence or absence of certain attributes of a certain object. For example, in psychology binary data may indicate if people do or do not posses a certain psychological trait; in ecology, the objects could be regions or districts in which certain species do or do not occur (or vice versa, the objects are two species that coexist in a number of locations); in archeology, binary data may reflect that particular artifact types were or were not found in a specific grave; finally, in chemical similarity searching, the objects may be target structures or queries and the attributes certain compounds in a database.

A vast amount of measures has been proposed that indicate how similar binary sequences are. A so-called similarity coefficient reflects in one way or another the association or resemblance of two or more binary variables. In various methods of data analysis, for example, multidimensional scaling or cluster analysis, the full information in the recorded binary variables is not required to perform the analysis. Often, the binary data are first summarized by a few coefficients or a coefficient matrix of pairwise resemblance measures. The information in the similarity coefficients is then used as input for the method of data analysis at hand.

Although the full information in comparing two binary variables is often not required, there are many different similarity coefficients that may be used to summarize the bivariate information. Preferring one coefficient over another may determine what information is summarized or what information is discarded. In order to choose the right coefficient, the different coefficients and their properties need to be better understood. Some properties of similarity coefficients for binary data are studied in this thesis. However, no attempt is made to be complete in the sense that all possible data-analytic applications of coefficients for binary data are covered. Instead, the thesis is centered around two theoretical issues.

The first issue is captured in the question, can the task of choosing the right coefficient be simplified? It may turn out that a coefficient may be placed in a group of coefficients all sharing a certain property. With respect to the property any coefficient in the group or family of coefficients can be used: one is as good as

the other. On the other hand, the property may also divide coefficients in different groups, coefficients that do posses the property and those that do not. For example, when comparing two binary variables it is not uncommon to be interested in the similarity between the variables corrected for possible similarity due to chance. It may turn out that some coefficients become equivalent after correction. The choice of coefficient can then be limited to coefficients that are not equivalent after correction for chance. As a second example, in cluster analysis several algorithms only make use of the ordinal information between the different coefficients, ignoring the numerical values. Coefficients can be grouped on the basis of what information they preserve with respect to an ordinal data analysis. The choice of coefficient can then be limited to coefficients that summarize different ordinal information.

As a second issue, a similarity coefficient must sometimes be considered in the context of the data-analytic study of which it is a part. Some method of data analysis may have certain prerequisites. If a coefficient possesses a specific property, it may be preferred over a coefficient which does not share this characteristic. For example, the outcome of metric data analysis methods like classical scaling, is better understood if the coefficient used in the analysis is metric, that is, satisfies the triangle inequality. As a bonus, the study of various properties of similarity coefficients provides a better understanding of the coefficients themselves. The insight obtained from how different coefficients are related, for example, one coefficient is the product of a transformation applied to a second coefficient, provides new ways of interpreting both coefficients.

The dissertation contains a mathematical approach to the analysis of resemblance measures for binary data. A variety of data-analytic properties are considered and for various coefficients it is established whether they possess the property or not. Counterexamples are sometimes used to show that a coefficient lacks a property. All mathematics are on the level of high school algebra and to read the thesis no 'higher' mathematical training is required. A statement is referred to as a proposition if it is believed to be a new result; a statement is called a theorem if the result is already known.

The first half of the dissertation (Part I and II) is devoted to what is basically two-way information. In the literature on data-analytic methods like, for example, cluster analysis, factor analysis, or multidimensional scaling, a distinction is made between two types of two-way information. Two-way similarity may be the bivariate information between two binary or dichotomous variables, that is, variables with two responses. Two-way similarity may also be the dyadic information between cases, persons, or objects. For the reader who is accustomed to this terminology it is important to note that in the present dissertation this (historical) distinction is largely ignored.

Some of the coefficients that are studied in the thesis have been proposed for comparing variables over cases, whereas others are primarily used to compare objects or cases over variables or attributes. Perhaps only a few coefficients are actually used in both the bivariate and dyadic case. Basically, similarity of two sequences of binary scores is referred to as two-way or bivariate information. The two terms are

considered interchangeable. To simplify the reading the sequences are referred to as variables. When considering a case by variable data matrix, the variables correspond to the columns. The latter notion is important in Part II on similarity matrices. A similarity matrix is obtained by calculating all two-way or pairwise coefficients between the columns of the case by variable data table. Finally, when two or more sequences are compared the words multi-way and multivariate are used.

This thesis consists of nineteen chapters divided into four parts. Part I and II are devoted to the bivariate case: a coefficient reflects the similarity of two variables at a time. Properties of individual coefficients are considered in Part I, whereas Part II focuses on properties that are studied in terms of coefficient matrices. Part III and IV are concerned with definitions and generalizations of various concepts from Part I and II to the multi-way case: a coefficient measures the resemblance of two or more binary variables. Part III is somewhat different from the other parts because no similarity coefficients are encountered in its chapters. Instead, various generalizations of the triangle inequality and other multi-way possibilities are studied in Part III. Some of the properties derived in Part III are used in Chapter 18 on metric properties of multi-way coefficients.

Part I consists of five chapters. Notation and some basic concepts concerning similarity coefficients are introduced in Chapter 1. We consider axioms for both similarity and dissimilarity coefficients. A first distinction is made between coefficients that do and coefficients that do not include the number of negative matches. A second distinction is made between coefficients that have zero value if the two variables are statistically independent and coefficients that have not. Also, some attention is paid to the problem of indeterminate values for coefficients that are fractions.

Chapter 2 is used to put the similarity coefficients for binary data into a broader perspective. The formulas considered in this thesis are often special cases that are obtained when more general formulas from various domains of data analysis are applied to dichotomous data. Furthermore, the same formulas may be encountered when two nominal variables are compared. For example, when comparing partitions from two cluster analysis algorithms or when measuring response agreement between two judges, a general approach is to count the four different types of pairs that can be obtained. The formulas defined on the four types of pairs may be equivalent to formulas defined on the four quantities obtained when comparing two binary variables.

In Chapter 3 it is shown that some resemblance measures belong to some sort of family of coefficients. Various relations between coefficients become apparent from studying their membership to a family. For most properties studied in Part I, greater generality is obtained if one works with (various types of) coefficient families. Linearity, another topic of this chapter, and metric properties (Chapter 10) are studied for families in which each coefficient is linear in both numerator and denominator.

Correction for chance agreement is the theme of Chapter 4. The chapter focuses on a coefficient family for which the study of correction for chance is relatively

simple. Several new properties on equivalences of coefficients after correction for chance irrespective of the choice of expectation are presented. In addition, a variety of properties of corrected coefficients are considered. Special interest is taken in a certain class of coefficients that become equivalent after correction. Also discussed is the relationship between the actual formula (coefficient) obtained after correction for chance and the particular choice of expectation.

The maximum value of various similarity coefficients is the topic of Chapter 5. Maximum values are studied in relationship to coefficient families that are power means. It is shown that different members of a specific family all have the same maximum value. New formulas are obtained if a coefficient is divided by its maximum value. Several results are presented that show what formulas are obtained after division by the maximum value. Two classes of coefficients are considered that become either a coefficient by Simpson (1943) or a coefficient by Loevinger (1947, 1948). Also, it is shown that Loevinger's coefficient is obtained if a general family of coefficients is corrected for both similarity due to chance and maximum value.

Part II consists of five chapters. In many applications of data analysis the data consist of more than two binary variables. In Part II various concepts and properties are considered that can only be studied when multiple variables (more than two) are considered. For example, multiple column vectors can be positioned next to each other to form a so-called data matrix. Given a binary data matrix, one may obtain a coefficient matrix by calculating all pairwise coefficients for any two columns of the data matrix. Different coefficient matrices are obtained, depending on the choice of similarity coefficient.

Chapter 6 focuses on how the 1s and 0s of the various column vectors of the data matrix may be related. For example, the 1s and 0s may be related in such a way that the data matrix exhibits certain patterns, possibly after a certain re-ordering or permutation of the columns, or after permuting both columns and rows of the data matrix. The 1s and 0s of the various column vectors may also be related in more complicated ways, not immediately clear from visual inspection. For example, some sort of probabilistic model can supposedly underlie the patterns of 1s and 0s of the various variables. Chapter 6 is used to describe some one-dimensional models and data structures that imply a certain ordering of the column vectors. These data structures are later on used in the remaining chapters of Part II for the study of various ordering properties of similarity matrices.

Chapter 7 is devoted to Robinson matrices. A square similarity matrix is called a Robinson matrix if the highest entries within each row and column are on the main diagonal and moving away from this diagonal, the entries never increase. A similarity matrix may or may not exhibit the Robinson property depending on the choice of resemblance measure. However, it seems to be a common notion in the classification literature that Robinson matrices arise naturally in problems where there is essentially a one-dimensional structure in the data. It is shown in Chapter 7 that the occurrence of a Robinson matrix is a combination of the choice of the similarity coefficient, and the specific one-dimensional structure in the data. Important coefficients in this chapter are the coefficient by Braun-Blanquet (1932) and Russel

and Rao (1940).

Eigendecompositions of several coefficient matrices are studied in Chapter 8. It is shown what information on the order of the model probabilities can be obtained from the eigenvector elements corresponding to the largest eigenvalues of various similarity matrices. It is therefore possible to uncover the correct ordering of several latent variable models considered in Chapter 6 using eigenvectors. The point to be made here is that the eigendecomposition of some similarity matrices, especially matrices corresponding to asymmetric coefficients, are more interesting compared to the eigendecomposition of other matrices. The important coefficients in this chapter have corresponding similarity matrices that are non-symmetrical. Also, the diverse matrix methodology of an eigenvalue method called homogeneity analysis is studied.

In Chapter 9, a systematic comparison of a one-dimensional homogeneity analysis and the item response theory approach is presented. It is shown how various item statistics from classical item analysis are related to the parameters of the 2-parameter logistic model from item response theory. Using these results, and the assumption that the homogeneity person score is a reasonable approximation of the latent variable, the functional relationships between the discrimination and location parameter of the 2-parameter logistic model and the two category weights of a homogeneity analysis applied to binary data are derived.

The study of metric properties is begun in Chapter 10, where metric properties of coefficients that are linear in both numerator and denominator are discussed. The chapter starts with an introduction of the concept of dissimilarity. Some tools are introduced here for the two-way case. Metric properties for multi-way coefficients are studied in Part IV. Because these tools are technically if not conceptually simpler for the two-way case, they are first presented here and later on generalized to the multi-way case in Chapters 15 and 18.

Part III consists of five chapters. Measures of resemblance play an important role in many domains of data analysis. However, similarity coefficients often only allow pairwise or bivariate comparison of variables or entities. An alternative to two-way resemblance measures is to formulate multivariate or multi-way coefficients. Before considering multi-way formulations of coefficients for binary data in Part IV, Part III is used to explore and extend some concepts from Chapter 10 and the literature on three-way data analysis to the multi-way case. Part III is devoted to possible generalizations and other related multi-way extensions of the triangle inequality, including the perimeter distance function, the maximum distance function, and multi-way ultrametrics.

Before extending the metric axioms, Chapter 11 is used to formulate more basic axioms for multi-way dissimilarities. Axiom systems for two-way and three-way dissimilarities are studied first. The dependencies between various axioms are reviewed to obtain axiom systems with a minimum number of axioms. The consistency and independence of several axiom systems is established by means of simple models. The remainder of Chapter 11 is used to explore how basic axioms for multi-way dissimilarities, like nonnegativity, minimality and symmetry, may be defined.

Chapter 12 explores how the two-way metric may be generalized to multi-way

metrics. A family of k-way metrics is formulated that generalize the two-way metric and the three-way metrics from the literature. Each inequality that defines a metric is linear in the sense that we have a single, possibly weighted, dissimilarity, which is equal to or smaller than an unweighted sum of dissimilarities. The family of inequalities gives an indication of the many possible extensions for introducing k-way metricity. It is shown how k-way metrics and k-way dissimilarities are related to their $(k-1)$-way counterparts.

Multi-way ultrametrics are explored in Chapter 13. In the literature two generalizations of the ultrametric inequality have been proposed for the three-way case. Continuing this line of reasoning three inequalities may be formulated for the four-way case. For the multi-way case $k-1$ inequalities may be defined. Some ideas on the three-way ultrametrics presented in the literature are explored in this chapter for multi-way dissimilarities. The multi-way ultrametrics as defined in this chapter imply a particular class of multi-way metrics.

In Chapter 14 it is explored how two particular three-way distance functions may be formulated for the multi-way case. The chapter is mostly about extensions of the three-way perimeter model. One section covers the maximum function, its multi-way extension, and a metric property of the generalization. The chapter contains both results on decompositions and on metric properties of two multi-way perimeter models. Chapter 15 is completely devoted to two generalizations of a particular theorem from Chapter 10. This result states that if d satisfies the triangle inequality, then so does the function $d/(c+d)$, where c is a positive real value. The result is extended to one family of multi-way metrics. An attempt is made to generalize the result to a class of stronger multi-way metrics.

Part IV consists of four chapters. In this final part, multivariate formulations of similarity coefficients are considered. Multivariate coefficients may 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 animals may or not may be encountered.

In Chapter 16 and 17 multivariate formulations (for groups of objects of size k) of various bivariate similarity coefficients (for pairs of objects) for binary data are presented. The multivariate coefficients in Chapter 16 are not functions of the bivariate similarity coefficients themselves. Instead, an attempt is made to present multivariate coefficients that reflect certain basic characteristics of, and have a similar interpretation as, their bivariate versions. The multivariate measures presented in Chapter 17 preserve the relations between various coefficients that were derived in Chapter 4 on correction for chance agreement. This chapter is also used to show how the multi-way formulations from the two chapters are related. In Chapter 18 metric properties of various multivariate coefficients with respect to the strong polyhedral generalization of the triangle inequality are studied. Finally, the Robinson matrices studied in Chapter 7 are extended to Robinson cubes in Chapter 19.

Part I

Similarity coefficients

CHAPTER 1

Coefficients for binary variables

Sequences of binary data are encountered in many different realms of research. For example, a rater may check whether or not a person possesses a certain psychological characteristic; it can be assessed if certain species types are encountered in a region or not; a person may fill in a test and can either fail or pass various items; it may be investigated if a certain object does possess or does not possess certain attributes or characteristics. Moreover, various types of quantitative data may be recoded and treated as binary. Noisy quantitative data may for instance be dichotomized. Quantitative data may also be dichotomized when the pertinent information for the problem at hand depends on a known threshold value.

A so-called similarity coefficient or association index reflects in one way or another the resemblance of two or more binary variables. Most coefficients have been proposed for the bivariate or two-way case, that is, the similarity of two sequences or variables of binary scores. In this first chapter a (brief) overview is presented of several of the bivariate coefficients for binary data that are available. The similarity coefficients may be considered both as population parameters as well as sample statistics. The formulations here will be the ones, utilized in the latter case. Following Sokal and Sneath (1963, p. 128) or more recently Albatineh, Niewiadomska-Bugaj and Mihalko (2006), 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.

A major distinction is made between coefficients that do and those that do not include a certain quantity d. If a binary variable is a coding of the presence or absence of a list of attributes, then d reflects the number of negative matches, which is generally felt not to contribute to similarity. A second distinction covers coefficients that have zero value if the two sequences are (statistically) independent and coefficients that have not.

Next to introducing various bivariate coefficients, the chapter is used to outline a common problem for coefficients for binary data. Since many similarity coefficients are defined as fractions, the denominator may become 0 in some cases. For these critical cases the value of the coefficient is undefined. This case of indeterminacy for some values of coefficients for binary data has been given surprisingly little attention. As it turns out, the number of critical cases differ with the coefficients.

1.1 Four dependent quantities

Suppose the data consist of two sequences of binary (1/0) scores, for example

$$\begin{bmatrix}1\\1\\0\\0\\1\\1\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}0\\1\\1\\0\\1\\0\end{bmatrix}.$$

Various data analysis techniques do not require the full information in the two binary sequences. A convenient way to summarize the information in the two vectors is by defining the four dependent quantities

a = proportion of 1s that the variables share in the same positions
b = proportion of 1s in the first variable and 0s in second variable in the same positions
c = proportion of 0s in the first variable and 1s in second variable in the same positions
d = proportion of 0s that both variables share in the same positions.

Together, the four quantities a, b, c, and d can be used to construct the 2×2 contingency table

	Variable two		
Variable one	Value 1	Value 0	Total
Value 1	a	b	p_1
Value 0	c	d	q_1
Total	p_2	q_2	1

where the marginal probabilities are given by

$$\begin{aligned} p_1 &= a + b \quad \text{proportion of 1s in the first variable} \\ p_2 &= a + c \quad \text{proportion of 1s in the second variable} \\ q_1 &= c + d \quad \text{proportion of 0s in the first variable} \\ q_2 &= b + d \quad \text{proportion of 0s in the second variable.} \end{aligned}$$

The information in the 2×2 contingency table can be summarized by an index, called here a coefficient of similarity (affinity, resemblance, association, coexistence). As a general symbol for a similarity coefficient the capital letter S will be used. An example of a similarity coefficient is the Phi coefficient, which is given by

$$S_{\text{Phi}} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

The measure S_{Phi} is sometimes attributed to Yule (1912), and is equivalent to the formula that is obtained when the Pearson's product-moment correlation derived for continuous data, is applied to binary data. See Zysno (1997) for a review on the literature on S_{Phi} and some of its modifications. The marginal proportions p_1, p_2, q_1, and q_2 can be used to obtain a shorter or more parsimonious formula for S_{Phi}, which is given by

$$S_{\text{Phi}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}}.$$

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 actually coefficient S_{Phi}. Sokal and Sneath (1963) (among others) make a major 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 reflects the number of negative matches, which is generally felt not to contribute to similarity. Sokal and Sneath (1963, p. 130) noted the following.

'Through reduction ad absurdum we can arrive at a universe of negative character matches purporting to establish the similarity between two entities.'

Sneath (1957) felt it was difficult to decide which negative features to include in a study and which to exclude.

'It is not pertinent to count "absence of feathers" when comparing two bacteria, but that this feature is applicable in comparing bacteria and birds.'

Sokal and Sneath (1963, p. 128, 130) also note that including negative matches may depend on what attributes or features are actually considered with respect to the species. They explain the difficulty as follows.

'It may be argued that basing similarity between two species on the mutual absence of a certain character is improper. The absence of wings, when observed among a group of distantly related organisms (such as a camel, louse and nematode), would surely be an absurd indication of affinity. Yet a positive character, such as the presence of wings (or flying organs defined without qualification as to kind of wing) could mislead equally when considered for a similarly heterogeneous assemblage (for example, bat, heron, and dragonfly).'

Examples (from the field of biological ecology) that do not include the quantity d are the coefficients given by

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

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2} \qquad \text{(Gleason, 1920; Dice, 1945; Sørenson, 1948)}$$

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

$$S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}} \qquad \text{(Driver and Kroeber, 1932; Ochiai, 1957).}$$

Coefficient S_{Jac} may be interpreted as the number of 1s shared by the variables in the same positions, divided by the total number of positions were 1s occur ($a + b + c = p_1 + p_2 - a$). Coefficient S_{Gleas} seems to be independently proposed by both Dice (1945) and Sørenson (1948) but is often contributed to the former. Bray (1956) noted that coefficient S_{Gleas} can already be found in Gleason (1920). The coefficient has also been proposed by various other authors, for example, Czekanowski (1932) and Nei and Li (1979). Coefficient S_{DK} by Driver and Kroeber (1932) is often attributed to Ochiai (1957). Coefficient S_{DK} is also proposed by Fowlkes and Mallows (1983) for the comparison of two clustering algorithms (see Section 2.2).

With respect to coefficient S_{Jac}, coefficient S_{Gleas} gives twice as much weight to a. The latter coefficient is regularly used with presence/absence data in the case that there are only a few positive matches relatively to the number of mismatches. In addition to S_{Jac} and S_{Gleas}, Sokal and Sneath (1963, p. 129) proposed a similarity measure that gives twice as much weight to the quantity $(b + c)$ compared to a, which is given by

$$S_{\text{SS1}} = \frac{a}{a + 2(b + c)}.$$

Coefficients S_{Jac}, S_{Gleas}, and S_{SS1} are rational functions which are linear in both numerator and denominator.

If a binary variable is a coding of a nominal variable, that is, one or the other of two mutually exclusive attributes (for example, correct and incorrect, or male and female), then the quantity a reflects the number of matches on the first attribute and d reflects the number of matches on the second one. In this case, it is often felt that the quantities a and d should be equally weighted.

Goodman and Kruskal (1954, p. 758) contend that, in general, the only reasonable coefficients are those based on $(a+d)$. Examples of coefficients that do include the quantity d are the coefficients given by

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \qquad \text{(Sokal and Michener, 1958; Rand, 1971)}$$

$$S_{\text{SS2}} = \frac{2(a+d)}{2a+b+c+2d} \qquad \text{(Sokal and Sneath, 1963)}$$

$$S_{\text{RT}} = \frac{a+d}{a+2(b+c)+d} \qquad \text{(Rogers and Tanimoto, 1960)}$$

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{p_1}+\frac{a}{p_2}+\frac{d}{q_1}+\frac{d}{q_2}\right) \qquad \text{(Sokal and Sneath, 1963)}$$

$$S_{\text{SS4}} = \frac{ad}{\sqrt{p_1 p_2 q_1 q_2}} \qquad \text{(Sokal and Sneath, 1963).}$$

Since a, b, c, and d are proportions, the simple matching coefficient $S_{\text{SM}} = a + d$. Coefficient S_{SM} can be interpreted as the number of 1s and 0s shared by the variables in the same positions, divided by the total length of the variables. Coefficient S_{SM} is also proposed by Rand (1971) for the comparison of two clustering algorithms and Brennan and Light (1974) for measuring agreement of two psychologists that rate people on categories not defined in advance (see Chapter 2). In addition to S_{SM} and S_{RT}, Sokal and Sneath (1963, p. 129) proposed coefficient S_{SS2}, which gives twice as much weight to the quantity $(a + d)$ compared to $(b + c)$. Moreover, Sokal and Sneath (1963) proposed coefficients S_{SS3} and S_{SS4} as alternatives (that include the quantity d) to coefficients S_{Kul} and S_{DK}. The coefficient by Rusel and Rao (1940), given by $S_{\text{RR}} = a/(a + b + c + d) = a$, is called hybrid by Sokal and Sneath (1963), since it includes the quantity d in the denominator but not in the numerator.

1.2 Axioms for (dis)similarities

Complementary to similarity or association is the concept of dissimilarity. As an alternative to a similarity measure, the fourfold table may also be summarized by some form of dissimilarity measure. A higher value of a similarity coefficient indicates there is more association between two binary variables, whereas a low value indicates that the two sequences are dissimilar. For a dissimilarity coefficient the interpretation is the other way around. A high value indicates great dissimilarity, whereas a low value indicates great resemblance. The capital letter D will be used as a general symbol for a dissimilarity coefficient in Parts I and IV. In Part III the symbol d is used.

Various authors presented more rigorous discussions on the concepts similarity and dissimilarity. A function can only be considered a similarity or dissimilarity if it satisfies certain requirements or axioms. Some interesting exposés and discussions on axioms for (dis)similarities can be found in Baroni-Urbani and Buser (1976), Baulieu (1989, 1997), Janson and Vegelius (1981) and Batagelj and Bren (1995), in the case of bivariate or two-way coefficients, and Heiser and Bennani (1997) and Joly and Le Calvé (1995), in the case of three-way or triadic coefficients. With respect to the latter, that is, three-way dissimilarities, see Chapter 11. In addition, Zegers (1986) presented an interesting overview of requirements for similarity coefficients for more general types of data.

An essential property of a similarity coefficient $S(x_1, x_2)$ that reflects the similarity between two variables x_1 and x_2, is the property that $S(x_1, x_1) \geq S(x_1, x_2)$ and $S(x_2, x_2) \geq S(x_1, x_2)$. Furthermore, it may be required that a coefficient is symmetric, that is, $S(x_1, x_2) = S(x_2, x_1)$. Examples of coefficients that are symmetric are

$$S_{\text{Phi}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}} \quad \text{and} \quad S_{\text{Jac}} = \frac{a}{a + b + c} = \frac{a}{p_1 + p_2 - a}.$$

Two-way similarity coefficients that do not satisfy the symmetry requirement are the functions that can be found in, among others, Dice (1945, p. 298), Wallace (1983), and Post and Snijders (1993), given by

$$S_{\text{Dice1}} = \frac{a}{a + b} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{a + c} = \frac{a}{p_2}.$$

Coefficient S_{Dice1} is the number of 1s that both sequences share in the same positions, relative to the total number of 1s in the first sequence. Both S_{Dice1} and S_{Dice2} can be interpreted as conditional probabilities.

If a variable is compared with itself, it may be required that the similarity equals the value 1, that is, $S(x_1, x_1) = 1$. Coefficients S_{Phi}, S_{Jac}, S_{Dice1}, and S_{Dice2} all satisfy this axiom. A coefficient that in general violates this requirement, is an interesting measure by Russel and Rao (1940), given by

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

In addition to the previous two axioms, it is sometimes required that a function has a certain range before it may be called a similarity. For similarities, it is sometimes required that the absolute value of a function is restricted from above by the value 1, that is, $|S(x_1, x_2)| \leq 1$. All coefficients that are investigated in this thesis satisfy this requirement. Coefficients that do not satisfy this axiom have quantities in the numerator that are not represented in the denominator. A coefficient that can be found in Kulczyński (1927), given by $a/(b + c)$, is an example of a coefficient that does not satisfy this requirement. Most similarity coefficients considered in this thesis satisfy the three above requirements.

Analogously to the requirements for similarities, there are axioms for the concept of dissimilarity. It is usual to require that a function $D(x_1, x_2)$ is referred to as a dissimilarity if it satisfies

$$\begin{aligned} & D(x_1, x_2) \geq 0 && \text{(nonnegativity)} \\ & D(x_1, x_2) = D(x_2, x_1) && \text{(symmetry)} \\ \text{and} \quad & D(x_1, x_1) = 0 && \text{(minimality).} \end{aligned}$$

A straightforward way to transform a similarity coefficient S into a dissimilarity coefficient D is taking the complement $D = 1 - S$. This transformation requires that $S(x_1, x_1) = 1$ in order to obtain $D = 0$. Another possible transformation, closely related to the Euclidean distance, is $D = \sqrt{1 - S}$ (Gower and Legendre, 1986): D is the square root of the complement of S. For several coefficients, transformation $D = 1 - S$ gives simple formulas. For example,

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

In order for coefficient D_{RR} to satisfy minimality, S_{RR} must be redefined as

$$S_{\text{RR}} = \begin{cases} 1 & \text{if } x_1 = x_2 \\ a & \text{otherwise.} \end{cases}$$

Dissimilarity coefficient D_{RR} is then given by

$$D_{\text{RR}} = \begin{cases} 0 & \text{if } x_1 = x_2 \\ 1 - a & \text{otherwise.} \end{cases}$$

With respect to a dissimilarity D various other requirements can be studied, which are usually not defined for a similarity coefficient S. For D to be a distance or metric, it must satisfy the metric axioms of symmetry and

$$D(x_1, x_2) = 0 \quad \text{if and only if } x_1 = x_2 \qquad \text{(definiteness)}$$

and foremost, the triangle inequality, which is given by

$$D(x_1, x_2) \leq D(x_1, x_3) + D(x_2, x_3).$$

Metric properties of various functions are studied (reviewed) in Chapter 10. In Chapter 12 various possible multi-way generalizations of the triangle inequality are studied. Another well-known inequality is the ultrametric inequality given by

$$D(x_1, x_2) \leq \max\left(D(x_1, x_3), D(x_2, x_3)\right).$$

If a dissimilarity $D(x_1, x_2)$ satisfies the ultrametric inequality, then it also satisfies the triangle inequality. Various multi-way generalizations of the ultrametric inequality are studied in Chapter 13. Axioms for multi-way or multivariate (dis)similarities are discussed in Chapter 11.

1.3 Uncorrelatedness and statistical independence

In probability theory two binary variables are called uncorrelated if they share zero covariance, that is, $ad - bc = 0$. The covariance between two binary variables is defined as the determinant of the 2×2 contingency table. In addition to being uncorrelated, two variables may be statistically independent, which is in general a stronger requirement compared to uncorrelatedness. The two concepts are equivalent if both variables are normally distributed. Probability theory tells us that two binary variables satisfy statistical independence if the odds ratio equals unity, that is

$$\frac{ad}{bc} = 1.$$

The odds ratio is defined as the ratio of the odds of an event occurring in one group (a/b) to the odds of it occurring in another group (c/d). These groups might be any other dichotomous classification. An odds ratio of 1 indicates that the condition or event under study is equally likely in both groups. An odds ratio greater than 1 indicates that the condition or event is more likely in the first group.

The value of the odds ratio lies between zero and infinity. Yule proposed two measures

$$S_{\text{Yule1}} = \frac{\frac{ad}{bc} - 1}{\frac{ad}{bc} + 1} = \frac{ad - bc}{ad + bc} \qquad \text{(Yule, 1900)}$$

and

$$S_{\text{Yule2}} = \frac{\frac{\sqrt{ad}}{\sqrt{bc}} - 1}{\frac{\sqrt{ad}}{\sqrt{bc}} + 1} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \qquad \text{(Yule, 1912)}$$

as alternatives to the odds ratio. Both coefficients S_{Yule1} and S_{Yule2} transform the odds ratio into a correlation-like scale with a range -1 to 1.

The odds ratio equals unity if $ad = bc$ which equals the case that $ad - bc = 0$. In this respect uncorrelatedness and independence are equivalent for two binary variables. For testing statistical independence, one may calculate the χ^2-statistic (Pearson and Heron, 1913; Pearson, 1947) for the 2×2 contingency table. Different opinions have been stated on what the appropriate expectations are for the fourfold table (see Chapter 4). In the majority of applications it is assumed that the data are a product of chance concerning two different frequency distribution functions underlying the two binary variables, each with its own parameter. The case of statistical independence for this possibility, conditionally on fixed marginal probabilities p_1, p_2, q_1, and q_2, is given by

	Variable two		
Variable one	Value 1	Value 0	Total
Value 1	p_1p_2	p_1q_2	p_1
Value 0	q_1p_2	q_1q_2	q_1
Total	p_2	q_2	1

The case of statistical independence visualized in this table is considered in Yule (1912), Pearson (1947), Goodman and Kruskal (1954) and Cohen (1960).

Let $E(a)$ denote the expectation of quantity a; the latter is the observed proportion of common 1s, whereas $E(a)$ is the expected proportion of common 1s. Under the assumption of two different frequency distribution functions, we have

$$\begin{aligned} a - E(a) &= a - p_1 p_2 = a(1 - a - b - c) - bc = ad - bc; \\ b - E(b) &= b - p_1 q_2 = bc - ad; \\ c - E(c) &= c - p_2 q_1 = bc - ad; \\ d - E(d) &= d - q_1 q_2 = ad - bc. \end{aligned}$$

The χ^2-statistic for the 2×2 contingency table is then given by

$$\chi^2 = \frac{n(ad - bc)^2}{p_1 p_2 q_1 q_2}$$

where n is the length of, or number of elements in, the binary variables. The quantity n is used to compensate for the fact that the entries in the fourfold table are proportions, not counts. The χ^2-statistic has one degree of freedom (Pearson, 1947; Fisher, 1922). The χ^2-statistic is related to the Phi coefficient by

$$S_{\text{Phi}} = \sqrt{\frac{\chi^2}{n}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}}.$$

Both χ^2 and S_{Phi} equal zero if $ad = bc$, that is, when the two binary variables have zero covariance or are statistically independent. Apart from coefficient S_{Phi} various other similarity coefficients are defined with the covariance $ad - bc$ in the numerator. An example is Cohen's kappa (Cohen, 1960), which in the case of two categories is given by

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

Coefficient S_{Cohen} is a measure that is corrected for similarity due to chance (see Section 2.1 and Chapter 4).

Various authors have studied the expected value and possible standard deviation of similarity coefficients (see, for example, Sokal and Sneath, 1963; Janson and Vegelius, 1981). An interesting overview of possible distributions and some new derivations for coefficients S_{SM}, S_{Jac}, and S_{Gleas}, is presented in Snijders, Dormaar, Van Schuur, Dijkman-Caes and Driessen (1990). Knowing a value of central tendency and a measure of the amount of likely dispersion for a coefficient, may be used for statistical inference. Next, it is possible to test the hypothesis whether a similarity coefficient is statistically different from the expected value or not.

1.4 Indeterminacy

In this section we work with a slightly adjusted definition of a similarity coefficient for two binary variables. Firstly, instead of proportions or probabilities, let a, b, c, and d be counts, and let $n = a + b + c + d$ denote the total number of attributes of the binary variables. Secondly, we define a presence/absence coefficient $S(a, b, c, d)$ or S to be a map $S : (\mathbb{Z}^+)^4 \to \mathbb{R}$ from the set, U, of all ordered quadruples of nonnegative integers into the reals (Baulieu, 1989).

Many similarity coefficients are defined as fractions. The denominator of these fractions may therefore become 0 for certain values of a, b, c and d. For example, it is well-known that if $d = n$, then the value of S_{Jac} given by

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

is not defined or indeterminate. As noted by Batagelj and Bren (1995, Section 4.2) this case of indeterminacy for some values of coefficients for binary data has been given surprisingly little attention. The critical case of S_{Jac} implies a situation in which two binary variables consist entirely of 0s. One may argue that it is highly unlikely that this occurs in practice. For example, in ecology it is unlikely to have an ordinal data table that has objects without species. Furthermore, the problem can be resolved by excluding zero vectors from the data. Although these may be valid arguments for S_{Jac}, it turns out that the number of cases in which the value of a coefficient is indeterminate, differs with the coefficients.

To compare the number of critical cases of two different coefficients, a domain of possible cases must be defined. Consider the set U of all ordered four-tuples (a, b, c, d) of nonnegative integers. Since $a + b + c + d = n$, the number of different quadruples for given n ($n \geq 1$) is given by the binomial coefficient

$$\binom{n+3}{3} = \frac{(n+3)!}{n!\,3!} = \frac{(n+3)(n+2)(n+1)}{6}$$

which is the number of different four-tuples one may obtain out of n objects. Thus, for $n = 1, 2, 3, 4, 5, \ldots$, the set U consists of $4, 10, 20, 35, 56, \ldots$ different four-tuples. For example, for $n = 2$ we have the ten unique four-tuples

$$\begin{array}{lll} (2,0,0,0) & (1,1,0,0) & (0,1,1,0) \\ (0,2,0,0) & (1,0,1,0) & (0,1,0,1) \\ (0,0,2,0) & (1,0,0,1) & (0,0,1,1) \\ (0,0,0,2). & & \end{array}$$

For each coefficient we may study for how many four-tuples or quadruples for fixed n the value of the coefficient is indeterminate. For twenty eight similarity coefficients for both nominal and ordinal data, the number of different quadruples

[0]Parts of this section are to appear in Warrens, M.J. (in press), On the indeterminacy of similarity coefficients for binary (presence/absence) data, *Journal of Classification*.

in U for which the denominator of the corresponding coefficient equals zero are presented in the following table

Ordinal data	Nominal data	4-tuples
S_{RR}	$S_{\text{SM}}, S_{\text{SS3}}, S_{\text{Mich}}, S_{\text{RT}}, S_{\text{Ham}}$	0
$S_{\text{Jac}}, S_{\text{Gleas}}, S_{\text{BUB}}, S_{\text{BB}}, S_{\text{SS1}}$		1
	$S_{\text{GK}}, S_{\text{Scott}}, S_{\text{Cohen}}, S_{\text{HD}}$	2
	S_{MP}	4
$S_{\text{Kul}}, S_{\text{DK}}, S_{\text{Sim}}, S_{\text{Sorg}}, S_{\text{McC}}$		$2n+1$
	$S_{\text{Phi}}, S_{\text{Yule1}}, S_{\text{Yule2}}, S_{\text{SS2}},$ $S_{\text{SS4}}, S_{\text{Fleiss}}, S_{\text{Loe}}$	$4n$

The formulas of all coefficients can be found in the appendix entitled "List of similarity coefficients". The above table may be read as follows. If $n = 5$, U has 56 elements and for 20 of these quadruples the value of the Phi coefficient S_{Phi} is indeterminate. Note that the coefficients are placed in groups with the same number of critical cases. For coefficients with the most critical cases ($4n$), the number of quadruples for which the value of the coefficient is indeterminate increases in a linear fashion as n becomes larger. Increases of the number of quadruples with the indeterminacy problem are not proportional to increases of n. Hence, the ratio

$$\frac{\text{number of critical cases in } U}{\text{total number of quadruples in } U} \quad \text{decreases as } n \text{ becomes larger.}$$

Furthermore, for most coefficients indeterminacy only occurs in the case that at least two elements of four-tuple (a, b, c, d) are zero.

As an alternative to excluding the vectors that result in zero denominators values, Batagelj and Bren (1995) proposed to eliminate the indeterminacies by appropriately defining values in critical cases. Some of the definitions presented in this section give the same results as definitions proposed in Batagelj and Bren (1995). The definitions presented here simplify the reading.

Let

$$K_y = \frac{a}{a+y} \quad \text{with} \quad y = b, c.$$

Coefficients S_{Gleas}, S_{DK}, S_{Kul} and

$$S_{\text{Sorg}} = \frac{a^2}{p_1 p_2}, \qquad S_{\text{BB}} = \frac{a}{\max(p_1, p_2)} \quad \text{and} \quad S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)}$$

are, respectively, the harmonic mean, geometric mean, arithmetic mean, product, minimum function, and maximum function of K_b and K_c.

Consider the arithmetic mean of K_b and K_c

$$S_{\text{Kul}} = \frac{K_b + K_c}{2} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right).$$

Suppose $a + c = 0$. Note that the value of S_{Kul} is indeterminate. If we set $K_c = 0$, then S_{Kul} becomes

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{a+b} + 0\right) = 0 \quad \text{since} \quad a = 0.$$

Alternatively, we may remove the part from the definition of S_{Kul} that causes the indeterminacy. Coefficient S_{Kul} becomes

$$S_{\text{Kul}} = \frac{a}{a+b} = 0 \quad \text{since} \quad a = 0.$$

Thus, either setting $K_c = 0$ or removing the indeterminate part from the definition of the coefficient, leads to the same conclusion: $S_{\text{Kul}} = 0$. We therefore define

$$S_{\text{Kul}} = \begin{cases} 0 & \text{if} \quad a + b = 0 \quad \text{or} \quad a + c = 0 \\ \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{DK}, S_{Sim}, and S_{Sorg}.

Coefficient

$$S_{\text{McC}} = \frac{a^2 - bc}{(a+b)(a+c)} = 2S_{\text{Kul}} - 1.$$

Suppose $a + c = 0$. The value of coefficient S_{McC} is indeterminate. Also the numerator $(a^2 - bc) = 0$. We define

$$S_{\text{McC}} = \begin{cases} 0 & \text{if} \quad a + b = 0 \quad \text{or} \quad a + c = 0 \\ \frac{a^2 - bc}{(a+b)(a+c)} & \text{otherwise.} \end{cases}$$

Consider the harmonic mean of K_b and K_c

$$S_{\text{Gleas}} = \frac{2}{K_b^{-1} + K_c^{-1}} = \frac{2a}{2a + b + c}.$$

Suppose $a + c = 0$. The value of K_c and K_c^{-1} is indeterminate. However, $2a/(2a + b + c) = 0$. Similar to S_{Kul} we define

$$S_{\text{Gleas}} = \begin{cases} 0 & \text{if} \quad d = n \\ 2a/(2a + b + c) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{Jac}, S_{SS2}, S_{BB}, and S_{BUB}.

Note that the definitions of S_{Kul} and S_{Gleas} presented here do not ensure that $S_{\text{Kul}} = 1$ or $S_{\text{Gleas}} = 1$ if variable x_1 is compared with itself. If $x_1 = x_2 = \overbrace{(0, 0, ..., 0)}^{n}$, that is, the two variables have nothing in common, $S_{\text{Kul}} = S_{\text{Gleas}} = 0$. Furthermore, if variable $x_1 = \overbrace{(0, 0, ..., 0)}^{n}$ is compared with itself, $S_{\text{Kul}} = S_{\text{Gleas}} = 0$. Since these coefficients are appropriate for ordinal data, it is a moot point what the value of the coefficient should be if variables x_1 and x_2, or just variable x_1 if x_2 is compared with itself, are zero vectors. From a philosophical point of view it might be better to leave the coefficients for ordinal data undefined for the critical case $d = n$.

Consider coefficient

$$S_{\text{HD}} = \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right) \qquad \text{(Hawkins and Dotson, 1968).}$$

The value of S_{HD} is indeterminate if either $a = n$ or $d = n$. If $a = n$ then variables x_1 and x_2 are unit vectors; if $d = n$ then variables x_1 and x_2 are zero vectors. If both variables are zero vectors or unit vectors, we may speak of perfect agreement if x_1 and x_2 are nominal variables. We therefore define

$$S_{\text{HD}} = \begin{cases} 1 & \text{if} \quad a = n \quad \text{or} \quad d = n \\ \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{Cohen}, S_{GK} and S_{Scott}. We also define

$$S_{\text{MP}} = \begin{cases} 1 & \text{if} \quad a = n \quad \text{or} \quad d = n \\ 0 & \text{if} \quad b = n \quad \text{or} \quad c = n \\ \frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)} & \text{otherwise.} \end{cases}$$

Consider the Phi coefficient

$$S_{\text{Phi}} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

The value of S_{Phi} is indeterminate if $a + b = 0$, $a + c = 0$, $b + d = 0$, or $c + d = 0$. For these critical cases the covariance $(ad - bc) = 0$. We define

$$S_{\text{Phi}} = \begin{cases} 1 & \text{if} \quad a = n \quad \text{or} \quad d = n \\ 0 & \text{if} \quad a + b = 0, \quad a + c = 0, \quad b + d = 0 \quad \text{or} \quad c + d = 0 \\ \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{SS4}, S_{Yule1}, S_{Yule2}, S_{Fleiss}, and S_{Loe}.

Let

$$K_y = \frac{a}{a+y} \quad \text{and} \quad K_y^* = \frac{d}{y+d} \quad \text{with} \quad y = b, c.$$

Consider the arithmetic mean of K_b, K_c, K_b^* and K_c^*

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right).$$

Suppose $c + d = 0$. Note that the value of K_c^* is indeterminate. To eliminate the critical case, we may set $K_c^* = 0$, and S_{SS3} becomes

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{a+b} + 1 + 0 + 0\right) = \frac{2a+b}{4(a+b)}. \tag{1.1}$$

Note that coefficient S_{SS3} in (1.1) has a range $[\frac{1}{4}, \frac{1}{2}]$. We may define

$$S_{\text{SS3}} = \begin{cases} \frac{2a+b}{4(a+b)} & \text{if} \quad c+d=0 \\ \frac{2a+c}{4(a+c)} & \text{if} \quad b+d=0 \\ \frac{b+2d}{4(b+d)} & \text{if} \quad a+c=0 \\ \frac{c+2d}{4(c+d)} & \text{if} \quad a+b=0 \\ \frac{1}{2} & \text{if} \quad a=n \quad \text{or} \quad d=n \\ 0 & \text{if} \quad b=n \quad \text{or} \quad c=n \\ \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) & \text{otherwise.} \end{cases}$$

As an alternative to the above robust definition of S_{SS3}, we propose to eliminate the critical case by removing the part from the definition of S_{SS3} that causes the indeterminacy. Suppose $c + d = 0$. The arithmetic mean of K_b, K_c and K_b^* is given by

$$S_{\text{SS3}}^* = \frac{1}{3}\left(\frac{a}{a+b} + 0 + 1\right) = \frac{2a+b}{3(a+b)}. \tag{1.2}$$

Note that coefficient S_{SS3}^* in (1.2) has a range $[\frac{1}{3}, \frac{2}{3}]$. We define

$$S_{\text{SS3}}^* = \begin{cases} \frac{2a+b}{3(a+b)} & \text{if} \quad c+d=0 \\ \frac{2a+c}{3(a+c)} & \text{if} \quad b+d=0 \\ \frac{b+2d}{3(b+d)} & \text{if} \quad a+c=0 \\ \frac{c+2d}{3(c+d)} & \text{if} \quad a+b=0 \\ 1 & \text{if} \quad a=n \quad \text{or} \quad d=n \\ 0 & \text{if} \quad b=n \quad \text{or} \quad c=n \\ \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) & \text{otherwise.} \end{cases}$$

1.5 Epilogue

In this first chapter basic notation and several concepts of similarity coefficients for binary data were introduced. A coefficient summarizes the two-way information in two sequences of binary (0/1) scores. A coefficient may be used to compare two variables over several cases or persons, two cases over variables, two objects over attributes, or two attributes over objects. Although the data analysis literature distinguishes between, for example, bivariate information between variables or dyadic information between cases, the terms bivariate and two-way are used for any two sequences of binary scores (the terms are considered interchangeable) in this dissertation.

Two distinctions between the large number of coefficients were made in this chapter. Coefficients may be divided in groups 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, then d reflects the number of negative matches. A second distinction was made between coefficients that have zero value if the two sequences are statistically independent and coefficients that have not. A full account of the possibilities of statistical testing with respect to the 2×2 contingency table can be found in Pearson (1947).

No attempt was made to present a complete overview of all proposed or all possible coefficients for binary data. An overview of bivariate coefficients for binary data from the literature can be found in the appendix entitled "List of similarity coefficients". To obtain some ideas of other possible coefficients, the reader is referred to other sources: Sokal and Sneath (1963), Cheetham and Hazel (1969), Baroni-Urbani and Buser (1976), Janson and Vegelius (1982), Hubálek (1982), Gower and Legendre (1986), Krippendorff (1987), Baulieu (1989) and Albatineh et al. (2006).

CHAPTER 2

Coefficients for nominal and quantitative variables

The main title ("Similarity coefficients for binary data") suggests that the thesis is about resemblance or association measures between objects characterized by two-state (binary) attributes. Many of the bivariate or two-way coefficients, however, were not proposed for use with binary variables only. The formulas considered in this thesis are often special cases that are obtained when more general formulas from various domains of data analysis are applied to dichotomous data. The general resemblance measures may, for example, be used for frequency data or other positive counts. Some coefficients based on proportions a, b, c, and d are special cases of not just one, but multiple coefficients. For example, coefficient

$$S_{\text{Gleas}} = \frac{2a}{2a+b+c} \quad \text{or its complement} \quad 1 - S_{\text{Gleas}} = \frac{b+c}{2a+b+c}$$

have been proposed for binary variables by Gleason (1920), Dice (1945), Sørenson (1948), Nei and Li (1979), and seem to have been popularized by Bray (1956) and Bray and Curtis (1957). Coefficient S_{Gleas} is a special case of, for example, a coefficient by Czekanowski (1932), a measure by Odum (1950), and a coefficient by Williams, Lambert and Lance (1966). The simple matching coefficient

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \quad \text{or its complement} \quad 1 - S_{\text{SM}} = \frac{b+c}{a+b+c+d}$$

can be obtained, for example, as a special case of a general coefficient by Gower (1971) or Cox and Cox (2000), the observed proportion of agreement of a bivariate table of two nominal variables, the city-block distance, or as a special case of a measure by Cain and Harrison (1958).

This chapter is used to present various interesting formulas for nominal and quantitative variables, accompanied by some measures used in set theory, of which some of the coefficients that will be frequently encountered in this thesis, like S_{Gleas} and S_{SM}, are special cases. This puts the coefficients for binary data in a more general context. In addition, from this chapter ideas or possibilities may be obtained for generalizing some of the results presented in this dissertation.

2.1 Nominal variables

When dealing with bivariate or two-way similarity coefficients for nominal variables two situations can be distinguished. The two nominal variables have either identical categories or they have different categories (Popping, 1983a; Zegers, 1986). The latter possibility is discussed in Section 2.3. Suppose that two psychologists each distribute m people among a set of k mutually exclusive categories. In addition suppose that the categories are defined in advance. To measure the agreement among the two psychologists, a first step is to obtain a contingency table or matching table $\mathbf{N}$ with elements n_{ij}, where n_{ij} indicates the number of persons placed in category i $(i = 1, 2, ..., I)$ by the first psychologist and in category j $(j = 1, 2, ..., J)$ by the second psychologist. Furthermore, let

$$n_{i+} = \sum_{j=1}^{J} n_{ij} \quad \text{and} \quad n_{+j} = \sum_{i=1}^{I} n_{ij}$$

denote the marginal counts (row and column totals) of the contingency table $\mathbf{N}$. Suppose that the categories of both nominal variables are in the same order, so that the diagonal elements of the square matrix $\mathbf{N}$ (n_{ii}) reflect the number of people put in the same category by both psychologists. If there are just two categories, then $m^{-1}\mathbf{N}$ equals the usual fourfold table. A straightforward measure of bivariate association is the observed proportion of agreement P_{o}, given by

$$P_{\text{o}} = \frac{1}{m} \sum_{i=1}^{k} n_{ii} = \frac{\text{tr}(\mathbf{N})}{m}.$$

If there are just two categories, for example, presence or absence of a psychological characteristic, then

$$P_{\text{o}} = \frac{a+d}{a+b+c+d} = S_{\text{SM}}.$$

Both Scott (1955) and Cohen (1960) proposed measures that incorporate correction for chance agreement. Both measures are corrected versions of P_{o}.

After correction a similarity coefficient S has a form

$$CS = \frac{S - E(S)}{1 - E(S)} \tag{2.1}$$

where $E(S)$ is conditional on the marginals of the contingency table of which S is the summary statistic. Furthermore, the constant 1 in the denominator of (2.1) may be replaced by the maximum value of a coefficient S (all coefficients that are studied in this thesis have a maximum value of unity). Expectation $E(S)$ depends on the marginal proportions, but the maximum value does not.

We note two expectations of P_{o}, which will be referred to as the expected proportion of agreement $E(P_{\mathrm{o}})$. Scott (1955) works with the assumption that the data are a product of chance of a single frequency distribution. To estimate the common parameters from the marginal counts, Scott (1955) uses

$$E(P_{\mathrm{o}})_{\mathrm{Scott}} = \frac{1}{4} \sum_{i=1}^{k} \left(\frac{n_{i+}}{m} + \frac{n_{+i}}{m} \right)^2 . \tag{2.2}$$

Alternatively, Cohen (1960) works with the assumption that the data are a product of chance of two different frequency distributions, one for each nominal variable. The expected proportion of agreement under statistical independence is given by

$$E(P_{\mathrm{o}})_{\mathrm{Cohen}} = \frac{1}{m^2} \sum_{i=1}^{k} n_{i+} n_{+i}. \tag{2.3}$$

Expectation (2.3) may be obtained by considering all permutations of the observations of one of the two variables, while preserving the order of the observations of the other variable. For each permutation the value of P_{o} can be determined. The arithmetic mean of these values is (2.3).

Using P_{o} and either (2.2) or (2.3) in (2.1), we obtain Scott's pi and Cohen's kappa, which are given by

$$S_{\mathrm{Scott}} = \frac{P_{\mathrm{o}} - E(P_{\mathrm{o}})_{\mathrm{Scott}}}{1 - E(P_{\mathrm{o}})_{\mathrm{Scott}}} \qquad \text{and} \qquad S_{\mathrm{Cohen}} = \frac{P_{\mathrm{o}} - E(P_{\mathrm{o}})_{\mathrm{Cohen}}}{1 - E(P_{\mathrm{o}})_{\mathrm{Cohen}}}$$

and become respectively

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

with binary variables. Other suitable measures for nominal variables with identical categories are discussed in Janson and Vegelius (1979).

2.2 Comparing two partitions

In cluster analysis one may be interested in comparing two clustering methods (Rand, 1971; Fowlkes and Mallows, 1983; Hubert and Arabie, 1985; Lerman, 1988; Steinley, 2004; Albatineh et al., 2006). Suppose we have two partitions of m data points. To compare these two clusterings, a first step is to obtain a so-called matching table $\mathbf{N}$ with elements n_{ij}, where n_{ij} indicates the number of data points placed in cluster i $(i = 1, 2, ..., I)$ according to the first clustering method and in cluster j $(j = 1, 2, ..., J)$ according to the second method.

The total number of points being clustered is given by $m = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. The cluster sizes in respective clusterings are the row and column totals of the matching table n_{i+} and n_{+j}. Furthermore, we define the quantity

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \binom{n_{ij}}{2} = \frac{1}{2} \left[\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}^2 - m \right]$$

which equals the number of object pairs that were placed in the same cluster according to both clustering methods, and the three quantities

$$P = \sum_{i=1}^{I} \binom{n_{i+}}{2}, \quad Q = \sum_{j=1}^{J} \binom{n_{+j}}{2} \quad \text{and} \quad N = \binom{m}{2}.$$

The quantity N equals the total number of pairs of objects given m points.

As a second step, one may calculate some sort of resemblance measure that summarizes the information in the matching table. A well-known measure for the similarity of two partitions is the Rand index (Rand, 1971), given by

$$S_{\text{Rand}} = \frac{N + 2T - P - Q}{N}.$$

Another measure of resemblance for comparing two partitions is the coefficient by Fowlkes and Mallows (1983), given by

$$S_{\text{FM}} = \frac{T}{\sqrt{PQ}}.$$

Similar to the proportion of observed agreement P_{o} from Section 2.1, coefficient S_{Rand} may be adjusted for agreement due to chance (Morey and Agresti, 1984; Hubert and Arabie, 1985; Albatineh et al., 2006). Fowlkes and Mallows (1983) and Hubert and Arabie (1985, p. 197) noted that, if the generalized hypergeometric distribution function is assumed appropriate for the matching table $\mathbf{N}$, then the expectation $E(T)$ under statistical independence is given by

$$E\left(T\right) = \frac{PQ}{N}. \tag{2.4}$$

[0]Parts of this section are to appear in Warrens, M.J. (in press), On the equivalence of Cohen's kappa and the Hubert-Ararbie adjusted Rand index, *Journal of Classification*.

Using (2.4), the expectation of S_{Rand} can be written as

$$E(S_{\text{Rand}}) = 1 + \frac{2PQ}{N^2} - \frac{P+Q}{N} \tag{2.5}$$

(Hubert and Arabie, 1985, p. 198). Using S_{Rand} and (2.5) in (2.1), we obtain the Hubert-Arabie adjusted Rand index, given by

$$CS_{\text{Rand}} = S_{\text{HA}} = \frac{T - PQ/N}{\frac{1}{2}(P+Q) - PQ/N} = \frac{2(NT - PQ)}{N(P+Q) - 2PQ}$$

(Hubert and Arabie, 1985, p. 198).

As noted in, for example, Steinley (2004) or Albatineh et al. (2006), the information in a matching table $\mathbf{N}$ of two clustering partitions on the same data points, can be summarized by a fourfold contingency table with quantities a, b, c, and d, where a is the number of object pairs that were placed in the same cluster according to both clustering methods, b (c) is the number of pairs that were placed in the same cluster according to one method but not according to the other, and d is the number of pairs that were not in the same cluster according to either of the methods. It then holds that $a + b + c + d = N$, where $a = T$, $b = P - T$, $c = Q - T$ and $d = N + T - P - Q$, and $p_1 = a + b = P$ and $q_1 = c + d = N - P$. The four different types of object pairs are also distinguished in Brennan and Light (1974), Hubert (1977), and Hubert and Arabie (1985, p. 194). However, the latter authors expressed their formulas in terms of the binomial coefficients in quantities T, P, Q, and N, instead of the quantities a, b, c, and d.

Expressing S_{Rand} in terms of the quantities a, b, c, and d we obtain S_{SM} (see, for example, Lerman, 1988; Steinley, 2004; Albatineh et al., 2006). Expressing S_{FM} in terms of the quantities a, b, c, and d we obtain S_{DK} (see, for example, Lerman, 1988; Albatineh et al., 2006). Expressing S_{HA} in these quantities, we obtain, following Steinley (2004, p. 388), the formula

$$S_{\text{HA}} = \frac{N(a+d) - [(a+b)(a+c) + (b+d)(c+d)]}{N^2 - [(a+b)(a+c) + (b+d)(c+d)]}. \tag{2.6}$$

The numerator of (2.6) can be written as

$$\begin{aligned} &N(a+d) - [(a+b)(a+c) + (b+d)(c+d)] \\ &= Na - p_1p_2 + Nd - q_1q_2 \\ &= 2(ad - bc) \end{aligned}$$

whereas the denominator of (2.6) equals

$$\begin{aligned} &N^2 - [(a+b)(a+c) + (b+d)(c+d)] \\ &= N^2 - p_1p_2 - q_1q_2 \\ &= (p_1+q_1)(p_2+q_2) - p_1p_2 - q_1q_2 \\ &= p_1q_2 + p_2q_1. \end{aligned}$$

Hence, expressing S_{HA} in terms of the quantities a, b, c, and d, the coefficient is equivalent to S_{Cohen}. Moreover, expectation $E(T)$ in (2.4) can be written as

$$E(T) = \frac{PQ}{N} = \frac{(a+b)(a+c)}{N} = \frac{p_1 p_2}{N}.$$

Hence, statistical independence under the generalized hypergeometric distribution function used in Hubert and Arabie (1985) for the matching table of two clusterings, is equivalent to the case of statistical independence under the binomial distribution function for the fourfold contingency table.

A practical conclusion is that we can calculate the Hubert-Arabie adjusted Rand index (S_{HA}) by first forming the fourfold contingency table counting the number of pairs of objects that were placed in the same cluster in both clusterings, in the same cluster in one clustering but in different clusters in the other clustering, and in different clusters in both, and then computing Cohen's kappa (S_{Cohen}) on this fourfold table.

2.3 Comparing two judges

A problem equivalent to that of comparing two partitions of two cluster algorithms may be encountered in psychology. In contrast to the case in Section 2.1, the categories are not defined in advance and the number of categories used by each psychologist may be different. Measures of agreement among judges in classifying answers to open-ended questions, or psychologists rating people, have been described by Brennan and Light (1974), Montgomery and Crittenden (1977), Hubert (1977), Janson and Vegelius (1982), and Popping (1983a). All these authors consider pairs of people and established for all N pairs formed from the m answers for both judges whether or not they were assigned to the same category. A comparison of the various measures is presented in Popping (1984).

We adopt the notation from Section 2.2, where quantities a, b, c, and d denote the four different types of pairs. Brennan and Light (1974) proposed the measure

$$S_{\text{BL}} = \frac{a+d}{a+b+c+d}$$

which equals the Rand index S_{Rand} and the simple matching coefficient S_{SM}. Montgomery and Crittenden (1977) proposed the measure

$$S_{\text{MC}} = \frac{ad-bc}{ad+bc}$$

which equals coefficient S_{Yule1} by Yule (1900). Hubert (1977) proposed a measure referred to as gamma, which is given by

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

Coefficient S_{Hub} is equal to a coefficient proposed by Hamann (1961) S_{Ham} and the $G-$index by Holley and Guilford (1964).

A discussion of properties of S_{Hub} and some adjustments to coefficient S_{Hub} can be found in Janson and Vegelius (1982). As an alternative to S_{Hub} these authors present a measure called the $J-$index. Popping (1983a, 1983b) proposed a measure based on the dot-product referred to as $D2$.

2.4 Quantitative variables

Let $\mathbf{x}_j$ and $\mathbf{x}_k$ be two column vectors of length n with positive entries, for example, counts or frequencies. In this section some examples of similarity coefficients formulated in terms of the elements of $\mathbf{x}_j$ and $\mathbf{x}_k$ are considered. Let x_{ij} denote the ith element of $\mathbf{x}_j$, and let x_{ik} denote the ith element of $\mathbf{x}_k$. In the terminology of Zegers (1986, p. 58) the measures considered in this section are coefficients for quantitative variables that consist of raw scores. These measures are either similarity functions or functions of the dissimilarity/distance type. Alternatively, one may formulate resemblance measures for normed raw scores, deviation scores, rank order scores, or combination of the previous scores. The reader is referred to Zegers (1986) and Gower and Legendre (1986) for more rigorous exposés on association coefficients for quantitative data.

The complement of the simple matching coefficient $1 - S_{\text{SM}}$ is a special case of the city-block or Manhattan distance

$$\frac{1}{n}\sum_{i=1}^{n}|x_{ij} - x_{ik}|.$$

The Jaccard (1912) coefficient

$$S_{\text{Jac}} = \frac{a}{a+b+c}$$

is obtained if in functions

$$\frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2 - \sum_{i=1}^{n} x_{ij}x_{ik}} \quad \text{or} \quad \frac{\sum_{i=1}^{n}\min(x_{ij}, x_{ik})}{\sum_{i=1}^{n}\max(x_{ij}, x_{ik})}$$

x_{ij} and x_{ik} take on values 1 and 0 only. The complement of the Jaccard coefficient S_{Jac} is a special case of

$$\frac{\sum_{i=1}^{n}|x_{ij} - x_{ik}|}{\sum_{i=1}^{n}\max(x_{ij}, x_{ik})} \quad \text{or} \quad \frac{\sum_{i=1}^{n}(x_{ij} - x_{ik})^2}{\sum_{i=1}^{n}\max(x_{ij}, x_{ik})}.$$

A member of a more general family of coefficients considered in Zegers and Ten Berge (1985) is given by

$$\frac{2\mathbf{x}_j^T\mathbf{x}_k}{\mathbf{x}_j^T\mathbf{x}_j + \mathbf{x}_k^T\mathbf{x}_k} = \frac{2\sum_{i=1}^{n} x_{ij}x_{ik}}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2}.$$

The latter coefficient is called the coefficient of identity and becomes S_{Gleas} if x_{ij} and x_{ik} take on values 1 and 0 only.

The measure

$$\frac{\sum_{i=1}^{n}|x_{ij}-x_{ik}|}{\sum_{i=1}^{n}(x_{ij}+x_{ik})} \quad \text{becomes} \quad 1-S_{\text{Gleas}}=\frac{b+c}{2a+b+c}$$

if x_{ij} and x_{ik} take on values 1 and 0 only, which is the complement of S_{Gleas} (Gower and Legendre, 1986, p. 27). Coefficient

$$\frac{\mathbf{x}_j^T\mathbf{x}_k}{(\mathbf{x}_j^T\mathbf{x}_j)^{1/2}(\mathbf{x}_k^T\mathbf{x}_k)^{1/2}}$$

is referred to as the coefficient of proportionality in Zegers and Ten Berge (1985), commonly known as Tucker's congruence coefficient (Tucker, 1951), also proposed by Burt (1948). The congruence coefficient for binary variables is given by $S_{\text{DK}} = a/\sqrt{p_j p_k}$. Three similarity coefficients, namely

$$\begin{aligned} S_{\text{Kul}} &= \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right) \\ S_{\text{Gleas}} &= \frac{2a}{p_j+p_k} \\ \text{and} \quad S_{\text{Sim}} &= \frac{a}{\min(p_j,p_k)} \end{aligned}$$

are sometimes attributed to Kulczyński (1927), Czekanowski (1932) and Simpson (1943). These authors proposed the coefficients for quantitative variables, which are given respectively by

$$\begin{aligned} S_{\text{Kul}} &= \frac{1}{2}\left[\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ij}}+\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ik}}\right] \\ S_{\text{Cze}} &= \frac{2\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}(x_{ij}+x_{ik})} \\ \text{and} \quad S_{\text{Sim}} &= \max\left[\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ij}},\frac{\sum_{i=1}^{n}\min(x_{ij},x_{ik})}{\sum_{i=1}^{n}x_{ik}}\right]. \end{aligned}$$

Sepkoski (1974) argues that, although similarity coefficients have been widely employed in cluster analysis, their use has been, for the most part, restricted to binary data. This author proposed quantified coefficients using basic rules like

$$\begin{aligned} a &= \frac{1}{n}\sum_{i=1}^{n}\min(x_{ij},x_{ik}) \\ b+c &= \frac{1}{n}\sum_{i=1}^{n}[\max(x_{ij},x_{ik})-\min(x_{ij},x_{ik})] \\ p_j &= \frac{1}{n}\sum_{i=1}^{n}x_{ij} \quad \text{and} \quad p_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik}. \end{aligned}$$

The similarity coefficient used by Robinson (1951) can be written as

$$S_{\text{Rob}} = 1 - \frac{1}{2}\sum_{i=1}^{n}\left|\frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} - \frac{x_{ik}}{\sum_{i=1}^{n} x_{ik}}\right|.$$

When the data are binary, S_{Rob} becomes

$$S_{\text{BB}} = \frac{a_{jk}}{\max(p_j, p_k)} \qquad \text{(Braun-Blanquet, 1932).}$$

Proposition 2.1. *If S_{Rob} is applied to binary (1/0) data, then $S_{\text{Rob}} = S_{\text{BB}}$.*
Proof: For $p_j \geq p_k$, S_{Rob} can be written as

$$S_{\text{Rob}} = 1 - \frac{1}{2}\left(\frac{a_{jk}}{p_k} - \frac{a_{jk}}{p_j} + \frac{p_j - a_{jk}}{p_j} + \frac{p_k - a_{jk}}{p_k}\right) = \frac{1}{2} - \frac{p_j - 2a_{jk}}{2p_j} = \frac{a_{jk}}{p_j}.$$

Furthermore, for $p_j \leq p_k$, S_{Rob} can be written as

$$S_{\text{Rob}} = 1 - \frac{1}{2}\left(\frac{a_{jk}}{p_j} - \frac{a_{jk}}{p_k} + \frac{p_j - a_{jk}}{p_j} + \frac{p_k - a_{jk}}{p_k}\right) = \frac{a_{jk}}{p_k}.$$

This completes the proof. □

2.5 Measures from set theory

Similarity and distance functions can also be defined on sets of arbitrary elements. The following notation is used. Let a set be denoted by A and let $\overline{A}$ denote its complement. Symbol $\cup$ denotes union or set sum, and $A \cup B$ is the set containing everything in either A or B or both. Also, $\cap$ denotes intersection or set product, and $A \cap B$ is the set containing just those elements common to both A and B. Furthermore, let $|A|$ denote the cardinality of set A, which is a measure of the number of elements of the set. Some examples of similarity coefficients for two sets A and B that are frequently used, are

$$\begin{aligned}
&\frac{2|A \cap B|}{|A| + |B|} && \text{Dice coefficient}\\
&\frac{|A \cap B|}{|A \cup B|} && \text{Jaccard coefficient}\\
&\frac{|A \cap B|}{|A|^{1/2}|B|^{1/2}} && \text{Cosine coefficient}\\
\text{and}\quad &\frac{|A \cap B|}{\min(|A|, |B|)} && \text{Overlap coefficient.}
\end{aligned}$$

Special cases of these measures are the respective similarity coefficients

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2},\ S_{\text{Jac}} = \frac{a}{a + b + c},\ S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}} \quad \text{and} \quad S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)}.$$

Restle (1959) studied the symmetric set difference

$$|(A \cup B) \cap (\overline{A \cap B})|$$

which is a more general form of the complement of the simple matching coefficient, $1 - S_{\text{SM}}$. Boorman and Arabie (1972) discuss several set-theoretical measures, including the minimum lattice-moves distance

$$|A| + |B| - 2|A \cap B|$$

which is equivalent to the above measure studied by Restle (1959), and the minimum set-moves distance which may be approximated by

$$|A \cap B| - \min(|A|, |B|).$$

2.6 Epilogue

In this second chapter, various general formulas from different domains of data analysis were considered. Some of the similarity coefficients for binary data considered throughout this thesis are special cases of these formulas. The chapter puts the coefficients for binary variables in a broader perspective. Furthermore, the more general formulas provide some ideas for possible generalizations of various results in this thesis. The thesis by Zegers (1986) is a good source for the vast amount of different contexts in which similarity coefficients may be considered.

It was shown that several similarity measures used in cluster analysis for the matching table of two clustering algorithms are in fact equivalent to similarity coefficients defined on the four dependent quantities from the 2×2 contingency table, after a simple recoding. Two well-known measures are the Rand index and the Hubert-Arabie adjusted Rand index, given respectively by

$$S_{\text{Rand}} = 1 - \frac{P + Q - 2T}{N} \quad \text{and} \quad S_{\text{HA}} = \frac{2(NT - PQ)}{N(P + Q) - 2PQ}.$$

Both measures are calculated using the information in the matching of two clusterings on the same data points. Coefficient S_{Rand} was also proposed by Brennan and Light (1974) for comparing ratings by two psychologists. If the Rand index S_{Rand} is formulated in terms of the quantities a, b, c, and d, it is equivalent to the simple matching coefficient S_{SM}. Furthermore, if the Hubert-Arabie adjusted Rand index S_{HA} is formulated in terms of the quantities a, b, c, and d, it is equivalent to Cohen's kappa for two categories (S_{Cohen}).

Interestingly, both Cohen (1960) and Hubert and Arabie (1985) proposed a similarity measure that has been, or still is, the preferred coefficient, or at least the best-known coefficient, in their particular domain of data analysis (respectively interrater reliability and cluster analysis). Moreover, both measures were proposed in response to, or as alternative to, earlier coefficients (Scott, 1955, in the case of Cohen, 1960; Morey and Agresti, 1984, in the case of Hubert and Arabie, 1985).

CHAPTER 3

Coefficient families

In this chapter it is shown how various similarity coefficients may be related. Similarity measures may be members of some sort of parameter family or can be related in the sense that several coefficients have a similar form. Various well-known coefficients belong to parameter families of which all members are fractions, linear in both numerator and denominator. A distinction is made between coefficients that do include the quantity d (representing negative matches), like

$$S_{\mathrm{SM}} = \frac{a+d}{a+b+c+d} \quad \text{and} \quad S_{\mathrm{Ham}} = \frac{a-b-c+d}{a+b+c+d} \quad \text{(Hamann, 1961)}$$

and those that do not include the quantity d, like

$$S_{\mathrm{Jac}} = \frac{a}{p_1+p_2-a} \quad \text{and} \quad S_{\mathrm{Gleas}} = \frac{2a}{p_1+p_2}.$$

A variety of similarity coefficients can be defined as some sort of mean value of two different quantities. For example, resemblance measures S_{Gleas} and

$$S_{\mathrm{DK}} = \frac{a}{\sqrt{p_1 p_2}} \quad \text{and} \quad S_{\mathrm{Kul}} = \frac{a(p_1+p_2)}{2p_1p_2}$$

are respectively the harmonic, geometric and arithmetic mean of the conditional probabilities $p_1^{-1}a$ and $p_2^{-1}a$.

Different types of coefficients may be obtained by considering abstractions of these Pythagorean means. One type of generalized mean that is considered in this chapter is the so-called power mean.

A very general family of coefficients is the class of all functions of the form $\lambda+\mu a$, where a is the proportion of 1s that two variables share in the same positions, and λ and μ are functions of p_1 and p_2 only. This family includes coefficients S_{Gleas}, S_{DK}, and S_{Kul} and various other measures. Properties of this family with respect to correction for similarity due to chance, are considered in Chapter 4.

There are some advantages to studying families of coefficients instead of individual coefficients. First of all, from the family formulation it is often apparent how different members are related. Coefficient properties like bounds are easily investigated using parameter families. Another advantage of studying parameter families instead of individual coefficients, is that often more general results can be obtained. As an example, results on linearity given in Hubálek (1982) for individual coefficients are here studied for families of coefficients.

3.1 Parameter families

Gower and Legendre (1986, p. 13) define two parameter families of which all members are linear in both numerator and denominator. They make a distinction between coefficients that do and do not include the quantity d. The first family for presence/absence data is given by

$$S_{\text{GL1}}(\theta) = \frac{a}{a+\theta(b+c)} = \frac{a}{\theta(p_1+p_2)+(1-2\theta)a}.$$

where $\theta > 0$ to avoid negative values. Members of $S_{\text{GL1}}(\theta)$ are

$$\begin{aligned} S_{\text{GL1}}(\theta=1) &= S_{\text{Jac}} = \frac{a}{p_1+p_2-a} \\ S_{\text{GL1}}(\theta=1/2) &= S_{\text{Gleas}} = \frac{2a}{p_1+p_2} \\ S_{\text{GL1}}(\theta=2) &= S_{\text{SS1}} = \frac{a}{a+2(b+c)} \quad \text{(Sokal and Sneath, 1963).} \end{aligned}$$

Members with $0 < \theta < 1$ give more weight to a. With presence/absence data this is regularly done in the case that there are only a few positive matches relatively to the number of mismatches, that is, a is much smaller than $(b+c)$. Similar arguments can be used for the opposite case and $\theta > 1$.

All members of $S_{\text{GL1}}(\theta)$ are bounded by 0 and 1, that is, $0 \le S_{\text{GL1}}(\theta) \le 1$. In addition, members are bounds of each other:

$$0 \le S_{\text{SS1}} \le S_{\text{Jac}} \le S_{\text{Gleas}} \le 1$$

or more generally

$$S_{\text{GL1}}(\theta_1) \le S_{\text{GL1}}(\theta_2) \quad \text{for} \quad \theta_1 > \theta_2 > 0.$$

The formulation of $S_{\mathrm{GL1}}(\theta)$ (and that of $S_{\mathrm{GL2}}(\theta)$ below) is closely related to the concept of global order equivalence (Sibson, 1972; Batagelj and Bren, 1995). Let $S(a, b, c, d)$ denote a function of the quantities a, b, c, and d. Two coefficients S and S^* are said to be globally order equivalent if

$$\begin{array}{rrcl} & S(a_1, b_1, c_1, d_1) & > & S(a_2, b_2, c_2, d_2) \\ \text{if and only if} & S^*(a_1, b_1, c_1, d_1) & > & S^*(a_2, b_2, c_2, d_2). \end{array}$$

If two coefficients are globally order equivalent, they are interchangeable with respect to an analysis method that is invariant under ordinal transformations (see, for example, Gower, 1986; Batagelj and Bren, 1995).

Theorem 3.1. *Two members of $S_{\mathrm{GL1}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL1}}(\theta)$, we have

$$\begin{aligned} \frac{a_1}{a_1 + \theta(b_1 + c_1)} &> \frac{a_2}{a_2 + \theta(b_2 + c_2)} \\ a_1 a_2 + a_1\theta(b_2 + c_2) &> a_1 a_2 + a_2\theta(b_1 + c_1) \\ \frac{a_1}{b_1 + c_1} &> \frac{a_2}{b_2 + c_2}. \end{aligned}$$

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

Janson and Vegelius (1981) pointed out an interesting relationship between various members of $S_{\mathrm{GL1}}(\theta)$. With respect to S_{Gleas}, S_{Jac}, and S_{SS1}, we have

$$S_{\mathrm{Jac}} = \frac{S_{\mathrm{Gleas}}}{2 - S_{\mathrm{Gleas}}} \qquad \text{and} \qquad S_{\mathrm{SS1}} = \frac{S_{\mathrm{Jac}}}{2 - S_{\mathrm{Jac}}}.$$

In general we have the following result.

Proposition 3.1. *It holds that*

$$S_{\mathrm{GL1}}(2\theta) = \frac{S_{\mathrm{GL1}}(\theta)}{2 - S_{\mathrm{GL1}}(\theta)}.$$

Proof: Define $x = a + \theta(b + c)$. Then

$$\frac{S_{\mathrm{GL1}}(\theta)}{2 - S_{\mathrm{GL1}}(\theta)} = \frac{x^{-1}a}{x^{-1}(2x - a)} = S_{\mathrm{GL1}}(2\theta). \quad \square$$

A parameter family closely related to $S_{\mathrm{GL1}}(\theta)$ may be obtained using the transformation $2S - 1$, that is,

$$S_{\mathrm{GL3}}(\theta) = \frac{2a}{a + \theta(b + c)} - 1 = \frac{a - \theta(b + c)}{a + \theta(b + c)}$$

with $\theta > 0$. A member of $S_{\mathrm{GL3}}(\theta)$ is

$$S_{\mathrm{GL3}}(\theta = 1/2) = S_{\mathrm{NS1}} = \frac{2a - b - c}{2a + b + c} \qquad \text{(No source)}.$$

Members with $0 < \theta < 1$ give more weight to a. All members of $S_{\mathrm{GL3}}(\theta)$ are bounded by -1 and 1, that is, $-1 \leq S_{\mathrm{GL3}}(\theta) \leq 1$. Parameter family $S_{\mathrm{GL3}}(\theta)$ is a transformation that preserves the scale of $S_{\mathrm{GL1}}(\theta)$ but uses a different range. The value zero for $S_{\mathrm{GL3}}(\theta)$ is equal to the value 0.5 for $S_{\mathrm{GL1}}(\theta)$ for fixed θ. For example, we have

$$S_{\mathrm{Gleas}} = \frac{2a}{2a + b + c} = 0.5 \quad \text{if and only if} \quad 2a = b + c$$

and

$$S_{\mathrm{NS1}} = \frac{2a - b - c}{2a + b + c} = 0 \quad \text{if and only if} \quad 2a = b + c.$$

The zero value case of coefficient S_{NS1} is not the same as the zero value case for coefficients with the covariance $ad - bc$ in the numerator. Two variables are not necessarily statistically independent if $S_{\mathrm{NS1}} = 0$ (Section 1.3). The formulation of $S_{\mathrm{GL3}}(\theta)$ is not completely arbitrary, because it is related to $S_{\mathrm{GL1}}(\theta)$ by the concept of global order equivalence.

Proposition 3.2. *Two members of $S_{\mathrm{GL3}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL3}}(\theta)$, we have

$$\frac{a_1 - \theta(b_1 + c_1)}{a_1 + \theta(b_1 + c_1)} > \frac{a_2 - \theta(b_2 + c_2)}{a_2 + \theta(b_2 + c_2)} \quad \text{if and only if} \quad \frac{a_1}{b_1 + c_1} > \frac{a_2}{b_2 + c_2}.$$

Since an ordinal comparison with respect to $S_{\mathrm{GL3}}(\theta)$ does not depend on the value of θ, any two members of $S_{\mathrm{GL3}}(\theta)$ are globally order equivalent. □

Corollary 3.1 *Members of $S_{\mathrm{GL1}}(\theta)$ and $S_{\mathrm{GL3}}(\theta)$ are globally order equivalent.*

The second family in Gower and Legendre (1986, p. 13), the counterpart of $S_{\mathrm{GL1}}(\theta)$ for nominal data, is given by

$$S_{\mathrm{GL2}}(\theta) = \frac{a + d}{a + \theta(b + c) + d} = \frac{1 + 2a - p_1 - p_2}{1 + (\theta - 1)(p_1 + p_2) + 2a(1 - \theta)}$$

where $\theta > 0$ to avoid negative values.

Members of $S_{\mathrm{GL2}}(\theta)$ are

$$S_{\mathrm{GL2}}(\theta = 1) = S_{\mathrm{SM}} = \frac{a+d}{a+b+c+d} = a+d$$

$$S_{\mathrm{GL2}}(\theta = 1/2) = S_{\mathrm{SS2}} = \frac{2(a+d)}{2a+b+c+2d} = \frac{2(a+d)}{1+a+d}$$

(Sokal and Sneath, 1963)

$$S_{\mathrm{GL2}}(\theta = 2) = S_{\mathrm{RT}} = \frac{a+d}{a+2(b+c)+d} = \frac{a+d}{1+b+c}$$

(Rogers and Tanimoto, 1960).

Similar to $S_{\mathrm{GL1}}(\theta)$, the members of $S_{\mathrm{GL2}}(\theta)$ are bounded by 0 and 1, that is, $0 \leq S_{\mathrm{GL2}}(\theta) \leq 1$. Also, members with $0 < \theta < 1$ give more weight to $(a+d)$.

Theorem 3.2. *Two members of $S_{\mathrm{GL2}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL2}}(\theta)$, we have

$$\frac{a_1+d_1}{a_1+\theta(b_1+c_1)+d_1} > \frac{a_2+d_2}{a_2+\theta(b_2+c_2)+d_2}$$
$$\frac{a_1+d_1}{b_1+c_1} > \frac{a_2+d_2}{b_2+c_2}.$$

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

Families $S_{\mathrm{GL1}}(\theta)$ and $S_{\mathrm{GL2}}(\theta)$ are related in the following way.

Proposition 3.3. *It holds that $S_{\mathrm{GL2}}(\theta) \geq S_{\mathrm{GL1}}(\theta)$.*
Proof: $S_{\mathrm{GL2}}(\theta) \geq S_{\mathrm{GL1}}(\theta)$ if and only if $\theta d(b+c) \geq 0$. □

Similar to S_{Gleas}, S_{Jac}, and S_{SS1}, we have with respect to S_{SS2}, S_{SM}, and S_{RT}

$$S_{\mathrm{SM}} = \frac{S_{\mathrm{SS2}}}{2 - S_{\mathrm{SS2}}} \qquad \text{and} \qquad S_{\mathrm{RT}} = \frac{S_{\mathrm{SM}}}{2 - S_{\mathrm{SM}}}.$$

In general we have the following result.

Proposition 3.4. *It holds that*

$$S_{\mathrm{GL2}}(2\theta) = \frac{S_{\mathrm{GL2}}(\theta)}{2 - S_{\mathrm{GL2}}(\theta)}.$$

Proof: Define $x = a + \theta(b+c) + d$. Then

$$\frac{S_{\mathrm{GL2}}(\theta)}{2 - S_{\mathrm{GL2}}(\theta)} = \frac{x^{-1}(a+d)}{x^{-1}(2x-a-d)} = S_{\mathrm{GL1}}(2\theta). \quad \square$$

A parameter family closely related to $S_{\mathrm{GL2}}(\theta)$ may be obtained using the transformation $2S - 1$,

$$S_{\mathrm{GL4}}(\theta) = \frac{2(a+d)}{a+\theta(b+c)+d} - 1 = \frac{a-\theta(b+c)+d}{a+\theta(b+c)+d}$$

with $\theta > 0$. A member of $S_{\mathrm{GL4}}(\theta)$ is

$$S_{\mathrm{GL4}}(\theta = 1) = S_{\mathrm{Ham}} = \frac{a-b-c+d}{a+b+c+d} = a-b-c+d \quad \text{(Hamann, 1961)}.$$

Members with $0 < \theta < 1$ give more weight to $(a+d)$. We have

$$S_{\mathrm{SM}} = a + d = 0.5 \quad \text{if and only if} \quad a + d = b + c$$

and

$$S_{\mathrm{Ham}} = a - b - c + d = 0 \quad \text{if and only if} \quad a + d = b + c.$$

The zero value case of coefficient S_{Ham} is not the same as the zero value case for coefficients with the covariance $ad - bc$ in the numerator (Section 1.3), nor the zero value case of S_{NS1}. Two variables are not necessarily independent if $S_{\mathrm{Ham}} = 0$. The formulation of $S_{\mathrm{GL4}}(\theta)$ is not completely arbitrary, since it is related to $S_{\mathrm{GL2}}(\theta)$ by the concept of global order equivalence.

Proposition 3.5. *Two members of $S_{\mathrm{GL4}}(\theta)$ are globally order equivalent.*
Proof: For an arbitrary ordinal comparison with respect to $S_{\mathrm{GL4}}(\theta)$, we have

$$\frac{a_1-\theta(b_1+c_1)+d_1}{a_1+\theta(b_1+c_1)+d_1} > \frac{a_2-\theta(b_2+c_2)+d_2}{a_2+\theta(b_2+c_2)+d_2}$$
$$\frac{a_1+d_1}{b_1+c_1} > \frac{a_2+d_2}{b_2+c_2}.$$

Since an ordinal comparison with respect to $S_{\mathrm{GL4}}(\theta)$ does not depend on the value of θ, any two members of $S_{\mathrm{GL4}}(\theta)$ are globally order equivalent. □

Corollary 3.2 *Members of $S_{\mathrm{GL2}}(\theta)$ and $S_{\mathrm{GL4}}(\theta)$ are globally order equivalent.*

3.2 Power means

There are several functions that may reflect the mean value of two real positive values x and y. The harmonic, geometric and arithmetic means, also known as the Pythagorean means, are given by respectively

$$\frac{2}{x^{-1}+y^{-1}}, \quad \sqrt{xy} \quad \text{and} \quad \frac{x+y}{2}.$$

Several coefficients can be expressed in terms of these Pythagorean means. For example, consider the quantities

$$S_{\text{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{p_2}$$

(Dice, 1945; Post and Snijders, 1993). The harmonic, geometric and arithmetic means of the quantities S_{Dice1} and S_{Dice2} are respectively

$$S_{\text{Gleas}} = \frac{2a}{p_1+p_2}, \; S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}} \quad \text{and} \quad S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1}+\frac{a}{p_2}\right).$$

Different types of coefficients may be obtained by considering abstractions of the Pythagorean means. One type of so-called generalized means is the power mean, sometimes referred to as the Hölder mean (see, for example, Bullen, 2003, Chapter 3). Let θ be a real value. The power mean $M_\theta(x,y)$ of x and y is then given by

$$M_\theta(x,y) = \left(\frac{x^\theta + y^\theta}{2}\right)^{1/\theta}.$$

Special cases of $M_\theta(x,y)$ are

$$\begin{aligned}
\lim_{\theta\to-\infty} M_\theta(x,y) &= \min(x,y) && \text{(minimum)}\\
M_{-1}(x,y) &= \frac{2}{x^{-1}+y^{-1}} && \text{(harmonic mean)}\\
\lim_{\theta\to 0} M_\theta(x,y) &= \sqrt{xy} && \text{(geometric mean)}\\
M_1(x,y) &= \frac{x+y}{2} && \text{(arithmetic mean)}\\
\lim_{\theta\to\infty} M_\theta(x,y) &= \max(x,y) && \text{(maximum).}
\end{aligned}$$

[0]Parts of this section are to appear in Warrens, M.J. (in press), Bounds of resemblance measures for binary (presence/absence) variables, *Journal of Classification*.

A variety of coefficients turn out to be special cases of a power mean. In terms of S_{Dice1} and S_{Dice2} we characterize the following coefficients from the literature.

$$\begin{aligned}
S_{\text{BB}} &= \frac{a}{\max(p_1, p_2)} && \text{(minimum; Braun-Blanquet, 1932)} \\
S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} && \text{(harmonic mean)} \\
S_{\text{DK}} &= \frac{a}{\sqrt{p_1 p_2}} && \text{(geometric mean; Driver and Kroeber, 1932)} \\
S_{\text{Kul}} &= \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) && \text{(arithmetic mean; Kulczyński, 1927)} \\
S_{\text{Sim}} &= \frac{a}{\min(p_1, p_2)} && \text{(maximum; Simpson, 1943).}
\end{aligned}$$

The product of the two quantities (or the square of the geometric mean S_{DK}) is not a special case of a power mean. It is given by

$$S_{\text{Sorg}} = \frac{a^2}{p_1 p_2} \quad \text{(Sorgenfrei, 1958; Cheetham and Hazel, p. 1131).}$$

Coefficient S_{Sorg} is sometimes referred to as the correlation ratio. The various coefficients for presence/absence data (without the quantity d) are related in the following way.

Proposition 3.6. *It holds that*

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

Proof: Inequality (i) holds if and only if $p_1 p_2 \geq a(a+b+c)$ if and only if $bc \geq 0$. Inequality (ii) holds if and only if $b + c \geq \max(b, c)$. The remaining inequalities follow from a property of a power mean:

$$M_{\theta_1}\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \leq M_{\theta_2}\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \text{for} \quad \theta_1 < \theta_2. \qquad \square$$

As a second example of a power mean, consider the quantities

$$S_{\text{Cole1}} = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad - bc}{p_2 q_1} \quad \text{(Cole, 1949).}$$

The quantity $(ad - bc)$ is known as the covariance between two binary vectors. If $p_1 \leq p_2$ then $p_1 q_2$ is the maximum value of the covariance $(ad - bc)$ given the marginal proportions. Note that the covariance may become negative and strictly speaking we have defined the power mean for two real positive values only. However, as it turns out, the power mean of two real negative values has very similar properties as the power mean of two positive values. As long as the two values have the same sign, the distinction between positive and negative values appears not to be important.

With respect to S_{Cole1} and S_{Cole2} we have the special cases

$$\begin{aligned} S_{\text{Cohen}} &= \frac{2(ad-bc)}{p_1q_2+p_2q_1} && \text{(harmonic mean)} \\ S_{\text{Phi}} &= \frac{ad-bc}{\sqrt{p_1p_2q_1q_2}} && \text{(geometric mean)} \\ S_{\text{Loe}} &= \frac{ad-bc}{\min(p_1q_2, p_2q_1)} && \text{(maximum; Loevinger, 1947, 1948).} \end{aligned}$$

Coefficient S_{Loe} is attributed to Loevinger (1947, 1948) by Mokken (1971) and Sijtsma and Molenaar (2002). However, Krippendorff (1987) reports Benini (1901) as probably the first to put forward this coefficient. Some new properties of this coefficient are considered in Chapter 5. Similar to Proposition 3.6, the next result follows from a property of power means, more specifically the harmonic-geometric mean inequality.

Proposition 3.7. *It holds that*

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

3.3 A general family

Albatineh et al. (2006) define yet another way on how various coefficients can be related. These authors study correction for chance with respect to a family $\mathcal{L}$ of the form $\lambda + \mu x$. Coefficients in the $\mathcal{L}$ family are linear functions of the quantity x, and the expectation of $S = \lambda + \mu x$ depends on the quantity x only, that is, $E(S) = \lambda + \mu E(x)$. Properties of the $\mathcal{L}$ family with respect to correction for chance are considered in the next chapter. For the moment it will be shown that $\mathcal{L}$ defines a very general family.

For example, coefficients in Section 2.1 belong to $\mathcal{L}$ family. Using $x = P_{\text{o}}$ we have

$$S_{\text{SM}} = P_{\text{o}} \;\rightarrow\; \lambda = 0 \quad \text{and} \quad \mu = 1$$

$$S_{\text{Scott}} \;\rightarrow\; \lambda = \frac{-E(P_{\text{o}})_{\text{Scott}}}{1 - E(P_{\text{o}})_{\text{Scott}}} \quad \text{and} \quad \mu = \frac{1}{1 - E(P_{\text{o}})_{\text{Scott}}}$$

$$\text{and} \quad S_{\text{Cohen}} \;\rightarrow\; \lambda = \frac{-E(P_{\text{o}})_{\text{Cohen}}}{1 -_E (P_{\text{o}})_{\text{Cohen}}} \quad \text{and} \quad \mu = \frac{1}{1 - E(P_{\text{o}})_{\text{Cohen}}}.$$

As a second example, take $x = a$, the proportion of 1s that two binary variables share in the same positions, and λ and μ are functions of p_1 and p_2 only. Then we

have

$$\begin{aligned} S_{\text{SM}} &= a + d \\ &= 1 + 2a - p_1 - p_2 && \rightarrow \lambda = 1 - p_1 - p_2,\ \mu = 2 \\ S_{\text{Ham}} &= a - b - c + d \\ &= 2a + 1 - 2p_1 - 2p_2 && \rightarrow \lambda = 1 - 2p_1 - 2p_2,\ \mu = 2 \\ \text{and} \quad S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} && \rightarrow \lambda = 0,\ \mu = \frac{2}{p_1 + p_2}. \end{aligned}$$

In Proposition 3.8 it is shown that the power mean of the quantities S_{Dice1} and S_{Dice2}, and the power mean of S_{Cole1} and S_{Cole2} are in the $\mathcal{L}$ family.

Proposition 3.8. *Power means*

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \textit{and} \quad M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_2q_1}\right)$$

are members of the $\mathcal{L}$ family.
Proof:

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) = \left[\frac{a^\theta(p_1^\theta + p_2^\theta)}{2p_1^\theta p_2^\theta}\right]^{1/\theta} = \frac{a}{p_1p_2}\left[\frac{p_1^\theta + p_2^\theta}{2}\right]^{1/\theta}.$$

Thus, for

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \text{we have} \quad \mu = \frac{1}{p_1p_2}\left(\frac{p_1^\theta + p_2^\theta}{2}\right)^{1/\theta}.$$

Similarly, for

$$M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_2q_1}\right)$$

we have

$$\mu = \frac{1}{p_1p_2q_1q_2}\left[\frac{(p_1q_2)^\theta + (p_2q_1)^\theta}{2}\right]^{1/\theta} \quad \text{and} \quad \lambda = -\frac{1}{q_1q_2}\left[\frac{(p_1q_2)^\theta + (p_2q_1)^\theta}{2}\right]^{1/\theta}$$

because $ad - bc = a - p_1p_2$. □

Let $f(p_1, p_2)$ be a function of the marginals p_1 and p_2. Then, all coefficients of the form

$$\frac{a}{f(p_1,p_2)} \quad \text{or} \quad \frac{ad-bc}{f(p_1,p_2)} = \frac{a - p_1p_2}{f(p_1,p_2)}$$

belong to the $\mathcal{L}$ family. Examples are

$$\begin{aligned} S_{\text{RR}} &= \frac{a}{a+b+c+d} \\ S_{\text{MP}} &= \frac{2(ad-bc)}{p_1q_1 + p_2q_2} && \text{(Maxwell and Pilliner, 1968)} \\ \text{and} \quad S_{\text{Fleiss}} &= \frac{(ad-bc)(p_1q_1 + p_2q_2)}{2p_1q_2p_2q_1} && \text{(Fleiss, 1975).} \end{aligned}$$

Moreover, if two coefficients $S_1 = \lambda_1 + \mu_1 a$ and $S_2 = \lambda_2 + \mu_2 a$ are in $\mathcal{L}$, then the arithmetic mean

$$\frac{S_1 + S_2}{2} = \frac{\lambda_1 + \mu_1 a + \lambda_2 + \mu_2 a}{2} = \frac{\lambda_1 + \lambda_2}{2} + \frac{a(\mu_1 + \mu_2)}{2}$$

is also in $\mathcal{L}$. Finally, if $S_1 = \lambda + \mu a$ is in the $\mathcal{L}$ family, then

$$S_2 = 2S_1 - 1 = 2\lambda - 1 + 2\mu a$$

also belongs to $\mathcal{L}$.

3.4 Linearity

Instead of proportions, let a, b, c, and d be the number of 1s and 0s that two binary variables may share or not share in the same positions. Furthermore, let $S(a)$ be short for $S(a, b, c, d)$ (S is a function of quantities a, b, c and d) and let $S(a+1)$ be short for $S(a+1, b-1, c-1, d+1)$. Hubálek (1982) gives the following definition of linearity. A function $S(a)$ is called linear if

$$S(a+1) - S(a) = S(a+2) - S(a+1),$$

or equivalently, if

$$2 \times S(a+1) = S(a+2) + S(a).$$

Using this definition of linearity, non-linearity can be defined in two ways. A function $S(a)$ is called convex if $2 \times S(a+1) < S(a+2) + S(a)$; $S(a, b, c, d)$ is called concave if $2 \times S(a+1) > S(a+2) + S(a)$.

Using numerical examples, Hubálek (1982) determined for various coefficients which ones are linear and which are non-linear. In this section the above definition of linearity is studied for several parameter families, instead of individual coefficients. The result below concerns coefficients that are rational functions, linear in both numerator and denominator.

Let $x = f(a, d)$ denote a linear function of a and d, and let $y = g(b, c)$ denote a linear function of b and c. Furthermore, let

$$u = \begin{cases} 1 & \text{if } x \text{ is a function of } a \text{ or } d \text{ only} \\ 2 & \text{if } x \text{ is a function of both } a \text{ and } d \end{cases}$$

and let

$$v = \begin{cases} 1 & \text{if } y \text{ is a function of } b \text{ or } c \text{ only} \\ 2 & \text{if } y \text{ is a function of both } b \text{ and } c. \end{cases}$$

Proposition 3.9. *Parameter families of the form*

$$(i) \qquad S(x,y) = \frac{x}{x+y} \qquad \left(\textit{with} \quad S(x+u,y-v) = \frac{x+u}{x+u+y-v} \right)$$

and

$$(ii) \qquad S(x,y) = \frac{x-y}{x+y} \qquad \left(\textit{with} \quad S(x+u,y-v) = \frac{x+u-y+v}{x+u+y-v} \right)$$

are convex for $u < v$, linear for $u = v$, and concave for $u > v$.
Proof: We consider (i) first. Using (i) in $2 \times S(a+1) \leq S(a+2) + S(a)$ we obtain

$$\frac{2(x+u)}{x+u+y-v} \leq \frac{x+2u}{x+2u+y-2v} + \frac{x}{x+y}. \tag{3.1}$$

Bringing all fractions under the same denominator, (3.1) becomes

$$\begin{aligned} (x+y)(2x+2u)(x+2u+y-2v) \;&\leq\; (x+y)(x+2u)(x+u+y-v) \\ &+\; x(x+u+y-v)(x+2u+y-2v) \end{aligned}$$

which, after some algebra, equals

$$(x+y)(x^2+3ux+xy-3vx+2u^2-2uv) \leq x(x+u+y-v)(x+2u+y-2v)$$

which, after some more algebra, can be written as $u^2y + uvx \leq uvy + v^2x$ if and only if $u \leq v$.

Next, we consider (ii). Parameter families (i) and (ii) are related by

$$\frac{x-y}{x+y} \;=\; \frac{2x}{x+y} - 1. \tag{3.2}$$

Using (3.2) in $2 \times S(a+1) \leq S(a+2) + S(a)$ we obtain

$$\frac{4(x+u)}{x+u+y-v} - 2 \;\leq\; \frac{2(x+2u)}{x+2u+y-2v} + \frac{2x}{x+y} - 2$$

which equals (3.1). □

Corollary 3.3. *Parameter families*

$$S_{\mathrm{GL1}}(\theta) = \frac{a}{a+\theta(b+c)} \quad \textit{and} \quad S_{\mathrm{GL3}}(\theta) = \frac{a-\theta(b+c)}{a+\theta(b+c)}$$

are convex for $\theta > \frac{1}{2}$, linear for $\theta = \frac{1}{2}$, and concave for $0 < \theta < \frac{1}{2}$.
Proof: With respect to these families we have $x = a$ and $y = \theta(b+c)$, and hence $u = 1$ and $v = 2\theta$. The family is then convex if $1 < 2\theta$. □

Corollary 3.4. *Parameter families*

$$S_{\mathrm{GL2}}(\theta) = \frac{a+d}{a+\theta(b+c)+d} \quad \textit{and} \quad S_{\mathrm{GL4}}(\theta) = \frac{a-\theta(b+c)+d}{a+\theta(b+c)+d}$$

are convex for $\theta > 1$, linear for $\theta = 1$, and concave for $0 < \theta < 1$.
Proof: For these families $u = 2$ and $v = 2\theta$. The families are then convex if $2 < 2\theta$. □

3.5 Epilogue

In this chapter it was shown how various similarity coefficients may be related. Similarity measures may be members of some sort of parameter family or can be related in the sense that several coefficients have a similar form. Various well-known coefficients belong to parameter families of which all members are rational functions, linear in both numerator and denominator. Some coefficients are members of more than one family. As an example, consider

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

Coefficient S_{Gleas} is the harmonic mean of

$$S_{\text{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{p_2}$$

and is therefore a special case of a power mean. In addition, S_{Gleas} is a member ($\theta = 1/2$) of the family given by

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

Due to this double membership, S_{Gleas} is a key coefficient in Chapter 16, where various multivariate formulations of coefficients are presented. In terms of linearity as defined by Hubálek (1982), S_{Gleas} is the linear coefficient in family $S_{\text{GL1}}(\theta)$. For other values than $\theta = 1/2$ we obtain either convex or concave coefficients. With respect to the linearity,

$$S_{\text{SM}} = \frac{a + d}{a + b + c + d} = a + d$$

is the linear coefficient in the second family of rational functions, $S_{\text{GL2}}(\theta)$. Similar to S_{Gleas}, S_{SM} can be introduced as a special case of a power mean. For example, S_{SM} is equal to the harmonic mean of the quantities

$$\frac{a + d}{p_1 + q_2} \quad \text{and} \quad \frac{a + d}{p_2 + q_1}.$$

Both S_{Gleas} and S_{SM} can be written as linear functions of the quantity a and are therefore members in the $\mathcal{L}$ family. Some of the consequences of this property are studied in the next chapter: S_{Gleas} and S_{SM} become equivalent after correction for chance. Moreover given a certain expectation of the quantity a, S_{Gleas} and S_{SM} become

$$S_{\text{Cohen}} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1} \qquad \text{(Cohen's kappa)}$$

after correction for similarity due to chance.

There are some properties in which S_{Gleas} and S_{SM} do differ. With respect to indeterminacy, S_{Gleas} has more critical cases compared to S_{SM}. Moreover, in Chapter 10 it is shown that $1 - S_{\text{SM}}$ is metric, that is, $1 - S_{\text{SM}}$ is a function that satisfies the triangle inequality, whereas the function $1 - S_{\text{Gleas}}$ does not.

Instead of using the power mean, new coefficients may be created by considering other type of means (Bullen, 2003). For example, the Heronian mean of

$$S_{\text{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{p_2}$$

is given by

$$\frac{1}{3}\left(\frac{a}{p_1} + \frac{a}{\sqrt{p_1 p_2}} + \frac{a}{p_2}\right)$$

whereas the Heinz mean is given by

$$\left(\frac{a}{p_1}\right)^u \left(\frac{a}{p_2}\right)^{1-u} + \left(\frac{a}{p_1}\right)^{1-u} \left(\frac{a}{p_2}\right)^u \quad \text{with} \quad 0 \leq u \leq \frac{1}{2}.$$

New coefficients can also be created by including the quantities

$$\frac{d}{b+d} = \frac{d}{q_2} \quad \text{and} \quad \frac{d}{c+d} = \frac{d}{q_1}.$$

For example, the function

$$\frac{4ad}{4ad + (a+d)(b+c)}$$

is the harmonic mean of conditional probabilities

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

CHAPTER 4

Correction for chance agreement

When comparing two variables some degree of similarity or agreement may be expected due to chance alone, except for the most extreme circumstances (either $p_1 = q_2 = 0$ or $p_2 = q_1 = 0$). Different opinions have been stated on the need to incorporate chance similarity. Goodman and Kruskal (1954, p. 758) contend that similarity due to chance in the measurement of resemblance need not be of much concern, since the observed degree of similarity may usually be assumed to be in excess of chance. In contrast, Zegers (1986) and Popping (1983a) find it quite natural that in absence of association between two variables, the value of a similarity coefficient is zero. Whether or not correction for chance is desirable, depends on the domain or field of data analysis that is considered.

Consider the situation where two variables are the ratings of m people by two observers on two mutually exclusive categories, for example, the observers rate various persons on the presence or absence of a certain trait. In this field, Scott (1955), Cohen (1960), Fleiss (1975), Krippendorff (1987), and Zegers (1986), among others, have proposed measures that are corrected for chance. The best-known example is perhaps the kappa-statistic (Cohen, 1960; S_{Cohen}). Alternatively, the quantities a, b, c, and d can be the result of a comparison between two clustering methods (Section 2.2). In cluster analysis it is general consensus that the popular coefficient S_{SM}, called the Rand index, should be corrected for chance agreement (Morey and Agresti, 1984; Hubert and Arabie, 1985), although there is some debate on what expectation is appropriate (Steinley, 2004; Albatineh et al., 2006).

With respect to correction for chance, various authors have reported results on equivalence of coefficients after correction for similarity due to chance (Fleiss, 1975; Zegers, 1986). Albatineh et al. (2006) studied correction for chance for a family $\mathcal{L}$ of coefficients of the form $S = \lambda + \mu x$ (Section 3.3). These authors appear to be the first to study correction for chance irrespective of the used expectation $E(S)$. The present chapter continues and extends this general approach. Furthermore, the results in this chapter unify various findings in Fleiss (1975), Zegers (1986) and Krippendorff (1987).

Clearly, not all coefficients studied in this thesis have been proposed for, or are used in, data-analytic circumstances where it is desirable to incorporate chance similarity. This practical limitation is however ignored in this chapter. Correction for chance is studied for a general family of coefficients, while ignoring the data-analytic context in which the individual members are usually applied. Using the powerful result from Albatineh et al. (2006), some additional properties of coefficients of the form $\lambda+\mu x$ with respect to correction for chance are presented. For both uncorrected and corrected similarity coefficients properties are derived. Some specific results are obtained by considering different expectations.

4.1 Some equivalences

A corrected similarity coefficient (denoted CS) has, after elimination of the effect of similarity due to chance, a form (2.1)

$$CS = \frac{S - E(S)}{1 - E(S)} \tag{4.1}$$

where S is the similarity coefficient, $E(S)$ the similarity coefficient under chance, and 1 embodies the maximum value of S regardless of the marginal proportions. Most coefficients in this thesis have maximum value unity. Albatineh et al. (2006) showed that correction (4.1) is relatively simple for members in $\mathcal{L}$ family.

Theorem 4.1 [Albatineh et al., 2006, p. 309]. *Two members in the $\mathcal{L}$ family become identical after correction* (4.1) *if they have the same ratio*

$$\frac{1-\lambda}{\mu}. \tag{4.2}$$

Proof: $E(S) = E(\lambda + \mu x) = \lambda + \mu E(x)$ and consequently the CS becomes

$$CS = \frac{S - E(S)}{1 - E(S)} = \frac{\lambda + \mu x - \lambda - \mu E(x)}{1 - \lambda - \mu E(x)} = \frac{x - E(x)}{\mu^{-1}(1-\lambda) - E(x)}. \tag{4.3}$$

□

Thus, the value of a similarity coefficient after correction for chance depends on ratio (4.2), where λ and μ characterize the particular measure within the $\mathcal{L}$ family. Two members in $\mathcal{L}$ become identical after correction (4.1) if they have the same ratio (4.2).

The following corollary concerns the coefficients from Section 2.1 that are linear in the observed proportion of agreement P_{o}.

Corollary 4.1. *Coefficients*

$$S_{\mathrm{SM}} = P_{\mathrm{o}}$$
$$S_{\mathrm{Scott}} = \frac{P_{\mathrm{o}} - E(P_{\mathrm{o}})_{\mathrm{Scott}}}{1 - E(P_{\mathrm{o}})_{\mathrm{Scott}}}$$
$$\textit{and} \qquad S_{\mathrm{Cohen}} = \frac{P_{\mathrm{o}} - E(P_{\mathrm{o}})_{\mathrm{Cohen}}}{1 - E(P_{\mathrm{o}})_{\mathrm{Cohen}}}$$

become equivalent after correction (4.1).
Proof: By Theorem 4.1 it suffices to look at ratio (4.2). Using the formulas of λ and μ corresponding to each coefficient (see Section 3.3), ratio (4.2)

$$\frac{1-\lambda}{\mu} = 1 \tag{4.4}$$

for all three coefficients. □

The next corollary extends Corollary 4.2 (i) in Albatineh et al. (2006) from three measures (S_{SM}, S_{Ham}, S_{Gleas}) to ten coefficients. All ten coefficients are linear in the quantity a.

Corollary 4.2. *Coefficients*

$$
\begin{aligned}
S_{\text{SM}} &= 1 + 2a - p_1 - p_2 \\
S_{\text{Ham}} &= 1 + 2a - 2p_1 - 2p_2 \\
S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} \\
S_{\text{GK}} &= \frac{2\min(a,d) - b - c}{2\min(a,d) + b + c} && \text{(Goodman and Kruskal, 1954)} \\
S_{\text{NS1}} &= \frac{2a - b - c}{2a + b + c} = \frac{4a - 2p_1 + 2p_2}{p_1 + p_2} && \text{(no source)} \\
S_{\text{NS2}} &= \frac{2d}{b + c + 2d} = \frac{2(a + q_1 + q_2 - 1)}{q_1 + q_2} && \text{(no source)} \\
S_{\text{NS3}} &= \frac{2d - b - c}{b + c + 2d} = \frac{4a + 3q_1 + 3q_2 - 4}{q_1 + q_2} && \text{(no source)} \\
S_{\text{RG}} &= \frac{a}{p_1 + p_2} + \frac{a + q_1 + q_2 - 1}{q_1 + q_2} && \text{(Rogot and Goldberg, 1966)} \\
S_{\text{Scott}} &= \frac{4a - (p_1 + p_2)^2}{4 - (p_1 + p_2)^2} \\
S_{\text{Cohen}} &= \frac{2(a - p_1 p_2)}{p_1 q_2 + p_2 q_1}
\end{aligned}
$$

become equivalent after correction (4.1).
Proof: By Theorem 4.1 it suffices to look at ratio (4.2). Using the formulas of λ and μ corresponding to each coefficient, ratio (4.2)

$$\frac{1 - \lambda}{\mu} = \frac{p_1 + p_2}{2} \tag{4.5}$$

for all ten coefficients. □

Note that ratio (4.5) is the arithmetic mean of marginal probabilities p_1 and p_2. The interpretation of (4.5) depends on how x was specified in in $\lambda + \mu x$, and ratio (4.5) is different from (4.4). Alternatively, we may formulate the ten coefficients as functions that are linear in the quantity $x = a + d$ instead of $x = a$. The result with respect to correction for chance agreement is of course the same, but ratio (4.6) now equals ratio (4.4).

Corollary 4.2b. *Coefficients*

$$
\begin{aligned}
S_{\text{SM}} &= a + d \\
S_{\text{Ham}} &= 2(a + d) - 1 \\
S_{\text{Gleas}} &= \frac{(a + d) - 1}{p_1 + p_2} + 1 \\
S_{\text{GK}} &= \frac{2(a + d) - 2}{\min(p_1 + p_2, q_1 + q_2)} + 1 \\
S_{\text{NS1}} &= \frac{2(a + d) - 2}{p_1 + p_2} + 1 \\
S_{\text{NS2}} &= \frac{(a + d) - 1}{q_1 + q_2} + 1 \\
S_{\text{NS3}} &= \frac{2(a + d) - 2}{q_1 + q_2} + 1 \\
S_{\text{RG}} &= \frac{(a + d) - 1}{2(p_1 + p_2)} + \frac{(a + d) - 1}{2(q_1 + q_2)} + 1 \\
S_{\text{Scott}} &= \frac{4(a + d) - (p_1 + p_2)^2 - (q_1 + q_2)^2}{4 - (p_1 + p_2)^2 - (q_1 + q_2)^2} \\
S_{\text{Cohen}} &= \frac{(a + d) - p_1 p_2 - q_1 q_2}{p_1 q_2 + p_2 q_1}
\end{aligned}
$$

become equivalent after correction (4.1).
Proof: By Theorem 4.1 it suffices to look at ratio (4.2). Using the formulas of λ and μ corresponding to each coefficient, ratio (4.2)

$$\frac{1 - \lambda}{\mu} = 1 \tag{4.6}$$

for all ten coefficients. □

Since $a = p_2 - q_1 + d$, probabilities a and d are also linear in $(a+d)$. Linear in $(a+d)$ is therefore equivalent to linear in a and linear in d. Furthermore, Albatineh et al. (2006) studied coefficients that are linear in $\sum\sum n_{ij}^2$, where n_{ij} is the number of data points placed in cluster i according to the first clustering method and in cluster j according to the second clustering method. Because $ma = (\sum\sum n_{ij}^2 - m)/2$, linear in $\sum\sum n_{ij}^2$ is equivalent to linear in a and equivalent to linear in $(a + d)$.

The corrected coefficient corresponding to the nine resemblance measures in Corollary 4.2 has a form

$$CS = \frac{(a + d) - E(a + d)}{1 - E(a + d)}. \tag{4.7}$$

Coefficient (4.7) may be obtained by using $(a+d)$, $E(a+d)$, and (4.6) in the extreme-right part of (4.3). Since expectation $E(a + d)$ is unspecified, coefficient (4.7) is a general corrected coefficient.

4.2 Expectations

A commonly used expectation was briefly considered in Section 1.3. Different opinions have been stated on what the appropriate expectations are for the 2×2 contingency table. Detailed discussions on the various ways of regarding data as the product of chance can be found in Krippendorff (1987), Mak (1988), Bloch and Kraemer (1989) and Pearson (1947). In cluster analysis it is general consensus that the popular coefficient S_{SM}, called the Rand index, should be corrected for agreement due to chance (Morey and Agresti, 1984; Hubert and Arabie, 1985), although there is some debate on what expectation is appropriate (Hubert and Arabie, 1985; Steinley, 2004; Albatineh et al., 2006). We consider five examples of $E(a+d)$.

Suppose it is assumed that the frequency distribution underlying the two variables in the 2×2 contingency table is the same for both variables (Scott, 1955; Krippendorff, 1987, p. 113). Coefficients used in this context are sometimes referred to as agreement indices. The common parameter p must be either known or it must be estimated from p_1 and p_2. Different functions may be used. For example, Scott (1955) and Krippendorff (1987) use the arithmetic mean

$$p = \frac{p_1 + p_2}{2}.$$

Following Scott (1955) and Krippendorff (1987, p. 113) we have

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

Let n denote the number of elements of the binary variables. Mak (1988) proposed the expectation

$$E(a+d)_{\text{Mak}} = 1 - \frac{n(p_1 + p_2)(q_1 + q_2) - (b+c)}{2(n-1)}$$

(see also, Blackman and Koval, 1993).

Instead of a single distribution function, it may be assumed that the data in the fourfold table are a product of chance concerning two different frequency distributions, each with its own parameter (Cohen, 1960; Krippendorff, 1987). Coefficients used in this context are sometimes referred to as association indices. The expectation of an entry in the 2×2 contingency table under statistical independence, is defined by the product of the marginal probabilities. We have

$$E(a+d)_{\text{Cohen}} = p_1 p_2 + q_1 q_2.$$

Expectation $E(a+d)_{\text{Cohen}}$ can be obtained by considering all permutations of the observations of one of the two variables, while preserving the order of the observations of the other variable. For each permutation the value of $(a+d)$ can be determined. The arithmetic mean of these values is $p_1 p_2 + q_1 q_2$.

A third possibility is that there are no relevant underlying continua. For this case two forms of $E(a+d)$ may be found in the literature. Goodman and Kruskal (1954, p. 757) use expectation

$$E(a+d)_{\mathrm{GK}} = \frac{\max(p_1+p_2, q_1+q_2)}{2}.$$

According to Krippendorff (1987, p. 114) an equity coefficient is characterized by expectation

$$E(a+d)_{\mathrm{Kripp}} = \frac{1}{2}.$$

Let us summarize the three situations. In the case of association the observations are regarded as ordered pairs. In the case of agreement the observations are considered as pairs without regard for their order; a mismatch is a mismatch regardless of the kind. In the case of equity one only distinguishes between matching and non-matching observations (cf. Krippendorff, 1987).

Proposition 4.1 below unifies and extends findings in Fleiss (1975) and Zegers (1986) on what coefficients become Cohen's kappa after correction for chance. Depending on what expectation $E(a+d)$ is used, the coefficients in Corollary 4.2 become, after correction for chance, either Scott's (1955) pi (S_{Scott}), Cohen's (1960) kappa (S_{Cohen}), Goodman and Kruskal's (1954) lambda (S_{GK}), Hamann's (1961) eta (S_{Ham}), or Mak's (1988) rho. The latter coefficient can be written as

$$S_{\mathrm{Mak}} = \frac{4nad - n(b+c)^2 + (b+c)}{n(p_1+p_2)(q_1+q_2) - (b+c)} \qquad \text{(Mak, 1988)}$$

where n is length of the binary variables. With respect to Proposition 4.1, let $\mathcal{L}$ family consists of functions $\lambda + \mu(a+d)$.

Proposition 4.1. *Let S be a member in $\mathcal{L}$ family for which ratio* (4.6) *holds. If the appropriate expectation is*

$$\begin{array}{rlll}
(i) & E(a+d)_{\mathrm{Scott}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Scott}} \\
(ii) & E(a+d)_{\mathrm{Mak}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Mak}} \\
(iii) & E(a+d)_{\mathrm{Cohen}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Cohen}} \\
(iv) & E(a+d)_{\mathrm{GK}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{GK}} \\
(v) & E(a+d)_{\mathrm{Kripp}}, & \textit{then } S \textit{ becomes} & S_{\mathrm{Ham}}
\end{array}$$

after correction (4.1).
Proof (i): Using $E(a+d)_{\mathrm{Scott}}$ in (4.7) we obtain an index with numerator

$$a + d - \left(\frac{p_1+p_2}{2}\right)^2 - \left(\frac{q_1+q_2}{2}\right)^2 = 2ad - \frac{(b+c)^2}{2} \tag{4.8}$$

and denominator

$$\frac{(p_1+p_2+q_1+q_2)^2 - (p_1+p_2)^2 - (q_1+q_2)^2}{4} = \frac{(p_1+p_2)(q_1+q_2)}{2}. \tag{4.9}$$

Dividing the right-hand part of (4.8) by the right-hand part of (4.9) we obtain

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

Proof (*ii*): Using $E(a+d)_{\text{Mak}}$ in (4.7) and multiplying the result by $2(n-1)$ we obtain an index with numerator

$$\begin{aligned} &2(a+d-1)(n-1)+n(p_1+p_2)(q_1+q_2)-(b+c)\\ =&n(2a+b+c)(b+c+2d)-2n(b+c)+(b+c) \end{aligned} \tag{4.10}$$

and denominator

$$n(p_1+p_2)(q_1+q_2)-(b+c). \tag{4.11}$$

We have

$$\begin{aligned} &(2a+b+c)(b+c+2d)-2(b+c)\\ =&4ad+(2a+2d)(b+c)+(b+c)^2-2(b+c)\\ =&4ad+(2a+2d-2)(b+c)+(b+c)^2\\ =&4ad-2(b+c)^2+(b+c)^2\\ =&4ad-(b+c)^2. \end{aligned} \tag{4.12}$$

Using (4.12), numerator (4.10) can be written as

$$n\left[4ad-(b+c)^2\right]+(b+c). \tag{4.13}$$

Dividing (4.13) by (4.11) we obtain coefficient S_{Mak}.
Proof (*iii*): Using $E(a+d)_{\text{Cohen}}$ in (4.7) we obtain

$$\frac{a+d-p_1p_2-q_1q_2}{(p_1+q_1)(p_2+q_2)-p_1p_2-q_1q_2} = \frac{2(ad-bc)}{p_1q_2+p_2q_1} = S_{\text{Cohen}}.$$

Proof (*iv*): Using $E(a+d)_{\text{GK}}$ in (4.7) we obtain

$$\frac{2[a+d-\max(a,d)]-b-c}{2-2\max(a,d)-b-c} = \frac{2\min(a,d)-b-c}{2\min(a,d)+b+c} = S_{\text{GK}}.$$

Proof (*v*): Using $E(a+d)_{\text{Kripp}}$ in (4.7) we obtain

$$2(a+d)-1 = a-b-c+d = S_{\text{Ham}}. \qquad \square$$

4.3 Two transformations

In this section we consider the two functions of similarity coefficients

$$S_2 = 2S_1 - 1 \qquad \text{and} \qquad S_3 = \frac{S_1 + S_2}{2}.$$

Both transformations may be used to construct new resemblance measures from existing similarity coefficients. It holds that $S_2 = 2S_1 - 1$ is in the $\mathcal{L}$ family if and only if S_1 is in $\mathcal{L}$, and if S_1 and S_2 are in $\mathcal{L}$, then $S_3 = (S_1 + S_2)/2$ is in $\mathcal{L}$. In this section it is shown how the new coefficients are related to the old coefficients in terms of correction for similarity due to chance. With respect to Proposition 4.2, let $\mathcal{L}$ consists of functions of the form $\lambda + \mu x$.

Proposition 4.2. *Let S_1 be a member of $\mathcal{L}$. S_1 and $S_2 = 2S_1 - 1$ become identical after correction* (4.1).
Proof: $S_2 = 2\lambda + 2\mu a - 1$ and $E(S_2) = 2\lambda - 1 + 2\mu E(x)$. Consequently the CS_2 becomes

$$\begin{aligned} CS_2 &= \frac{2\lambda + 2\mu x - 1 - 2\lambda - 2\mu E(x) + 1}{1 - 2\lambda - 2\mu E(x) + 1} = \frac{\lambda + \mu x - \lambda - \mu E(x)}{1 - \lambda - \mu E(x)} \\ &= \frac{S_1 - E(S_1)}{1 - E(S_1)} = CS_1. \qquad \square \end{aligned}$$

Similarity coefficients that are related by transformation $S_2 = 2S_1 - 1$ can be found in Corollary 4.2. Examples are

$$\begin{aligned} S_{\text{Ham}} &= 2S_{\text{SM}} - 1 \\ S_{\text{NS1}} &= 2S_{\text{Gleas}} - 1 \\ \text{and} \quad S_{\text{NS3}} &= 2S_{\text{NS2}} - 1. \end{aligned}$$

Another example is $S_{\text{McC}} = 2S_{\text{Kul}} - 1$, where

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) \qquad \text{and} \qquad S_{\text{McC}} = \frac{a^2 - bc}{p_1 p_2} \quad \text{(McConnaughey, 1964).}$$

The fact that coefficient S_{Kul} and S_{McC} become equivalent after correction (4.1) irrespective of the used expectation was already proved in Corollary 4.2 (ii) in Albatineh et al. (2006).

Proposition 4.3. *Let S_i for $i = 1, 2, ..., m$ be members in $\mathcal{L}$ family that become identical after correction* (4.1). *Then S_i for $i = 1, 2, ..., m$ and the arithmetic mean $S^* = m^{-1} \sum_{i=1}^{m} S_i$ coincide after correction* (4.1).
Proof:

$$E(S^*) = E\left(\frac{\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i x}{m}\right) = \frac{\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i E(x)}{m}.$$

Using arithmetic mean S^* in (4.1), we obtain

$$CS^* = \frac{x - E(x)}{y - E(x)} \qquad \text{where} \quad y = \frac{m - \sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{m} \mu_i}.$$

Let

$$z = \frac{1 - \lambda_1}{\mu_1} = \frac{1 - \lambda_2}{\mu_2} = ... = \frac{1 - \lambda_m}{\mu_m}.$$

It must be shown that ratio y equals ratio z. We have

$$y = \frac{\sum_{i=1}^{m}(1 - \lambda_i)}{\sum_{i=1}^{m} \mu_i} = \frac{\sum_{i=1}^{m} z\mu_i}{\sum_{i=1}^{m} \mu_i} = \frac{z \sum_{i=1}^{m} \mu_i}{\sum_{i=1}^{m} \mu_i} = z.$$

This completes the proof. □

Coefficient

$$S_{\text{RG}} = \frac{a}{2a + b + c} + \frac{d}{b + c + 2d} = \frac{S_{\text{Gleas}} + S_{\text{NS2}}}{2}$$

in Corollary 4.2, is the arithmetic mean of S_{Gleas} and S_{NS2}.

4.4 Corrected coefficients

The coefficients in Corollary 4.2 and Proposition 4.1 become either S_{Scott}, S_{Mak}, S_{Cohen}, S_{GK}, or S_{Ham}, depending on what expectation $E(a + d)$ is used. Note that corrected coefficients S_{Scott}, S_{Cohen}, S_{GK}, and S_{Ham} belong to the class of resemblance measures that is considered in Corollary 4.2 and Proposition 4.1. This suggests that corrected coefficients may have some interesting properties. The corrected coefficients and their properties are the topic of this section. If $E(S)$ in (4.1) depends on the marginal probabilities of the 2×2 contingency table, then CS in (4.1) belongs to $\mathcal{L}$. With respect to Proposition 4.4, let $\mathcal{L}$ consists of functions of the form $\lambda + \mu(a + d)$.

Proposition 4.4. *Let $E(S)$ in* (4.1) *depend on the marginal probabilities. If S is in $\mathcal{L}$ family, then CS in* (4.1) *is in $\mathcal{L}$.*
Proof: Expectation $E(S) = E[\lambda_1 + \mu_1(a + d)]$ is a function of the marginal probabilities. Thus $E(a + d)$, λ, and μ in (4.3) are functions of the marginal proportions. Equation (4.3) can therefore be written in a form $\lambda_2 + \mu_2(a + d)$ where

$$\lambda_2 = \frac{-E(a + d)}{\mu_1^{-1}(1 - \lambda_1) - E(a + d)} \qquad \text{and} \qquad \mu_2 = \frac{1}{\mu_1^{-1}(1 - \lambda_1) - E(a + d)}. \qquad \square$$

Examples of corrected coefficients that are in the $\mathcal{L}$ family are S_{Scott}, S_{Cohen}, S_{GK}, and S_{Ham}. These coefficients may be considered as corrected coefficients as well as ordinary coefficients that may be corrected for agreement due to chance. For example, S_{Scott}, S_{GK}, and S_{Ham} (and S_{Cohen}) become S_{Cohen} after correction (4.1) if expectation $E(a+d)_{\text{Cohen}}$ is used. Coefficient S_{Mak} cannot be written in a form $\lambda + \mu(a+d)$, and does therefore not belong to $\mathcal{L}$.

Next we consider the following problem. Suppose a coefficient S in $\mathcal{L}$ is corrected twice, using two different expectations, $E(a+d)$ and $E(a+d)^*$. Let the corrected coefficients be given by

$$CS = \frac{a+d-E(a+d)}{\mu^{-1}(1-\lambda)-E(a+d)} \qquad \text{and} \qquad CS^* = \frac{a+d-E(a+d)^*}{\mu^{-1}(1-\lambda)-E(a+d)^*}.$$

Note that $\mu^{-1}(1-\lambda)$ corresponding to coefficient S, is the same in both CS and CS^*. The problem is then as follows: if $E(a+d) \geq E(a+d)^*$, how are CS and CS^* related? Proposition 4.5 below is limited to coefficients in the $\mathcal{L}$ family of which the maximum value is unity, that is

$$\lambda + \mu(a+d) \leq 1 \qquad \text{if and only if} \qquad \frac{1-\lambda}{\mu} \geq (a+d).$$

It can be verified that most (if not all) similarity coefficients in this thesis satisfy this condition.

Proposition 4.5. *$CS \leq CS^*$ if and only if $E(a+d) \geq E(a+d)^*$.*
Proof: $CS \leq CS^*$ if and only if

$$E(a+d)\left[\frac{1-\lambda}{\mu}-(a+d)\right] \geq E(a+d)^*\left[\frac{1-\lambda}{\mu}-(a+d)\right].$$

The requirement $\lambda + \mu(a+d) \leq 1$ completes the proof. □

In the following, let $S = \lambda + \mu(a+d)$ be in $\mathcal{L}$ family and let

$$CS_{\text{Name}} = \frac{a+d-E(a+d)_{\text{Name}}}{\mu^{-1}(1-\lambda)-E(a+d)_{\text{Name}}}$$

be a corrected coefficient using expectation $E(a+d)_{\text{Name}}$. Using specific expectations $E(a+d)$ in combination with Proposition 4.5, we obtain the following result.

Proposition 4.6. *It holds that* $CS_{\mathrm{GK}} \overset{(i)}{\leq} CS_{\mathrm{Scott}} \overset{(ii)}{\leq} CS_{\mathrm{Cohen}}$.
Proof (i): Due to Proposition 4.5, it suffices to show that $E(a+d)_{\mathrm{GK}} \geq E(a+d)_{\mathrm{Scott}}$. Suppose $(p_1 + p_2) \geq (q_1 + q_2)$. We have

$$\begin{aligned}
E(a+d)_{\mathrm{GK}} &\geq E(a+d)_{\mathrm{Scott}} \\
\frac{p_1+p_2}{2} &\geq \left(\frac{p_1+p_2}{2}\right)^2 + \left(\frac{q_1+q_2}{2}\right)^2 \\
\frac{p_1+p_2}{2}\left(1 - \frac{p_1+p_2}{2}\right) &\geq \left(\frac{q_1+q_2}{2}\right)^2 \\
\frac{p_1+p_2}{2}\left(\frac{q_1+q_2}{2}\right) &\geq \left(\frac{q_1+q_2}{2}\right)^2 \\
(p_1+p_2) &\geq (q_1+q_2).
\end{aligned}$$

Proof (ii): It must be shown that $E(a+d)_{\mathrm{Scott}} \geq E(a+d)_{\mathrm{Cohen}}$. We have

$$\left(\frac{p_1+p_2}{2}\right)^2 \geq p_1 p_2 \tag{4.14}$$

if and only if

$$\frac{p_1+p_2}{2} \geq \sqrt{p_1 p_2}. \tag{4.15}$$

Furthermore, we have

$$\left(\frac{q_1+q_2}{2}\right)^2 \geq q_1 q_2 \tag{4.16}$$

if and only if

$$\frac{q_1+q_2}{2} \geq \sqrt{q_1 q_2}. \tag{4.17}$$

Because the arithmetic mean of two numbers is equal or greater than the geometric mean, inequalities (4.15) and (4.17) are true. Adding (4.14) and (4.16) we obtain the desired inequality. □

Blackman and Koval (1993, p. 216) derived the inequality $S_{\mathrm{Scott}} \leq S_{\mathrm{Cohen}}$. Note that this inequality follows from the more general result Proposition 4.6 by using a coefficient S for which (4.6) is characteristic.

4.5 Epilogue

Under the assumption that $E(a+d) = p_1p_2 + q_1q_2$ is the appropriate expectation, Fleiss (1975) showed that

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} = a+d \quad \text{and} \quad S_{\text{Gleas}} = \frac{2a}{p_1+p_2}$$

and S_{GK} and S_{RG} become S_{Cohen} after correction (4.1). Zegers (1986) showed that S_{SM}, S_{Gleas} and S_{Ham} become S_{Cohen} after correction (4.1). Albatineh et al. (2006) showed that S_{SM}, S_{Gleas} and S_{Ham} become equivalent irrespective of the used expectation. These results were extended and unified by Corollary 4.2 and Proposition 4.1. Corollary 4.2 specifies up to ten coefficients that become equivalent after correction (4.1) irrespective of expectation $E(a+d)$. The coefficients in Corollary 4.2 become either S_{Scott}, S_{Mak}, S_{Cohen}, S_{GK}, or S_{Ham}, depending on what expectation $E(a+d)$ is used. Moreover, two transformations from Section 4.3 may be used to construct an infinite amount of coefficients that become equivalent after correction (4.1).

Whether $E(a+d)_{\text{Cohen}}$ or another $E(a+d)$ is the appropriate expectation depends on the context of the data analysis. However, since a large number of coefficients are defined with the covariance

$$\frac{a+d-E(a+d)_{\text{Cohen}}}{2} = \frac{(a-p_1p_2)+(d-q_1q_2)}{2} = ad-bc$$

in the numerator, it appears that $E(a+d)_{\text{Cohen}}$ is the preferred (or most appropriate) expectation in many cases.

The quantities

$$S_{\text{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{p_2}$$

and

$$S_{\text{Cole1}} = \frac{ad-bc}{p_1q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad-bc}{p_2q_1} \quad \text{(Cole, 1949)}$$

where used in the previous chapter to construct power means

$$M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) \quad \text{and} \quad M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_2q_1}\right).$$

As it turns out, if the expectation of a is $E(a) = p_1 p_2$, several members of the two power means corresponding to the same θ are related. We have, for example,

$$S_{\text{Dice1}} = \frac{a}{p_1} \qquad \text{becomes} \quad S_{\text{Cole1}} = \frac{ad - bc}{p_1 q_2}$$

$$S_{\text{Dice2}} = \frac{a}{p_2} \qquad \text{becomes} \quad S_{\text{Cole2}} = \frac{ad - bc}{p_2 q_1}$$

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2} \qquad \text{becomes} \quad S_{\text{Cohen}} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}$$

$$\text{and} \quad S_{\text{Sim}} = \frac{a}{\min(p_1, p_2)} \qquad \text{becomes} \quad S_{\text{Loe}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)}.$$

CHAPTER 5

Correction for maximum value

The proportions a, b, c, and d in the fourfold table

a	b	p_1
c	d	q_1
p_2	q_2	1

are constrained by the marginal proportions p_1, p_2, q_1, and q_2. The coefficients based on these quantities are therefore also constrained by the marginals, so that maximum and minimum values are sometimes untenable. Guilford (1965), Cureton (1959) and Davenport and El-Sanhurry (1991) consider the maximum of S_{Phi} given marginals p_1 and p_2, denoted by $[S_{\text{Phi}}]_{\text{max}}$. Loevinger (1947, 1948) suggested using the ratio

$$\frac{S_{\text{Phi}}}{[S_{\text{Phi}}]_{\text{max}}}$$

since this procedure allows the corrected value to become unity. As noted by Loevinger (1947, 1948), Sijtsma and Molenaar (2002) and Davenport and El-Sanhurry (1991), coefficients S_{Phi}, S_{Cohen} and S_{Loe} are related by

$$S_{\text{Loe}} = \frac{S_{\text{Phi}}}{[S_{\text{Phi}}]_{\text{max}}} = \frac{S_{\text{Cohen}}}{[S_{\text{Cohen}}]_{\text{max}}}.$$

The relations between similarity coefficients for two binary variables suggested in this equality are the topic of this chapter.

The maximum and minimum of various coefficients and several equivalences are studied first. The maximum of a coefficient is determined by applying the formula to the case of two Guttman items (Section 6.3; Mokken, 1971; Guilford, 1965). Furthermore, it is shown what families of coefficients become equivalent after correction

$$\frac{S}{[S]_{\max}}. \tag{5.1}$$

5.1 Maximum value

In this section we derive the maximum value for a variety of coefficients. We focus on coefficients that are special cases of a power mean. Following Guilford (1965) and Cureton (1959), the maximum value of a coefficient is obtained if either quantity b, c, or both equal zero. Hence, with unequal marginal proportions $p_1 \neq p_2$, the 2×2 contingency table has the form

a	0	p_1
c	d	q_1
p_2	q_2	1

for example $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

if $b = 0$, or

a	b	p_1
0	d	q_1
p_2	q_2	1

for example $\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

if $c = 0$. Note that the maximum is obtained if the two binary variables being compared are so-called Guttman items (Section 6.3; Mokken, 1971). The maximum value of proportion a given the marginals p_1 and p_2, denoted by $a_{\max}$, is given by

$$a_{\max} = \begin{cases} p_1 & \text{if} \quad b = 0 \\ p_2 & \text{if} \quad c = 0 \end{cases} \qquad \text{or} \quad a_{\max} = \min(p_1, p_2).$$

Thus, without correction for maximum value, quantity a can only reach its maximum value if $p_1 = p_2$. The maximum value of measures for binary variables that do not include quantity d, may be obtained by replacing probability a by $a_{\max}$. Assuming $p_1 \neq p_2$ we obtain

$$[S_{\text{GL1}}(\theta)]_{\max} = \frac{\min(p_1, p_2)}{\theta(p_1 + p_2) + (1 - 2\theta)\min(p_1, p_2)}$$

with

$$[S_{\mathrm{GL1}}(1)]_{\max} = [S_{\mathrm{Jac}}]_{\max} = \frac{\min(p_1, p_2)}{\max(p_1, p_2)} < 1$$

$$[S_{\mathrm{GL1}}(1/2)]_{\max} = [S_{\mathrm{Gleas}}]_{\max} = \frac{2\min(p_1, p_2)}{p_1 + p_2} < 1.$$

With respect to the inequalities

$$S_{\mathrm{Sorg}} = \frac{a^2}{p_1 p_2} \;\leq\; S_{\mathrm{Jac}} = \frac{a}{p_1 + p_2 - a} \;\leq\; S_{\mathrm{BB}} = \frac{a}{\max(p_1, p_2)}$$

we obtain the equality

$$[S_{\mathrm{Sorg}}]_{\max} = [S_{\mathrm{Jac}}]_{\max} = [S_{\mathrm{BB}}]_{\max} = \frac{\min(p_1, p_2)}{\max(p_1, p_2)}.$$

With respect to the power mean of the quantities

$$S_{\mathrm{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\mathrm{Dice2}} = \frac{a}{p_2}$$

the equality $a_{\max} = \min(p_1, p_2)$ leads to

$$\left[M_\theta \left(\frac{a}{p_1}, \frac{a}{p_2} \right) \right]_{\max} = M_\theta \left(1, \frac{\min(p_1, p_2)}{\max(p_1, p_2)} \right).$$

where

$$\frac{\min(p_1, p_2)}{\max(p_1, p_2)} = [S_{\mathrm{BB}}]_{\max}.$$

Thus, the maximum value of a coefficient that is a special case of the power mean of S_{Dice1} and S_{Dice2}, is equal to the coefficient corresponding to the same θ of the value 1 and $[S_{\mathrm{BB}}]_{\max}$, where the latter is the maximum value of the minimum function of S_{Dice1} and S_{Dice2}. Hence, only for the maximum function, that is, $S_{\mathrm{Sim}} = a/\min(p_1, p_2)$, it holds that

$$[S_{\mathrm{Sim}}]_{\max} = \lim_{\theta \to \infty} M_\theta \left(1, \frac{\min(p_1, p_2)}{\max(p_1, p_2)} \right) = \max \left(1, \frac{\min(p_1, p_2)}{\max(p_1, p_2)} \right) = 1.$$

Next, we consider the maximum value of the covariance $(ad - bc)$ of two binary variables. The maximum covariance given the marginals p_1 and p_2, denoted $(ad - bc)_{\max}$, is given by

$$(ad - bc)_{\max} = \begin{cases} p_1 q_2 & \text{if} \quad b = 0 \\ p_2 q_1 & \text{if} \quad c = 0 \end{cases} \qquad \text{or} \quad (ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1).$$

We may obtain the maximum value of measures for binary variables that use the covariance in the numerator by replacing covariance $(ad-bc)$ by $(ad-bc)_{\max}$. With respect to the power mean of the quantities

$$S_{\text{Cole1}} = \frac{ad-bc}{p_1q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad-bc}{p_2q_1} \quad \text{(Cole, 1949)}$$

the equality $(ad-bc)_{\max} = \min(p_1q_2, p_2q_1)$ leads to

$$\left[M_\theta\left(\frac{ad-bc}{p_1q_2}, \frac{ad-bc}{p_1q_2}\right)\right]_{\max} = M_\theta\left(1, \frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}\right).$$

Thus, the maximum value of a coefficient that is a special case of the power mean of S_{Cole1} and S_{Cole2}, is equal to the coefficient corresponding to the same θ of the value 1 and the quantity

$$\frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}.$$

Hence, only for the maximum function, that is, S_{Loe}, it holds that

$$[S_{\text{Loe}}]_{\max} = \lim_{\theta\to\infty} M_\theta\left(1, \frac{\min(p_1q_2, p_2q_1)}{\max(p_1q_2, p_2q_1)}\right) = 1.$$

5.2 Correction for maximum value

Let x/y and x/z be two real positive values, of which the maximum depends on x only, that is

$$\left[\frac{x}{y}\right]_{\max} = \frac{x_{\max}}{y} \quad \text{and} \quad \left[\frac{x}{z}\right]_{\max} = \frac{x_{\max}}{z}.$$

Examples of x/y and x/z are S_{Dice1} and S_{Dice2}. For example, $x = a$ or $x = ad-bc$ and y and z are functions of p_1 and p_2 only. It turns out that division of the power mean of x/y and x/z by its maximum value given quantities y and z, does not depend on the choice of θ. Moreover, the outcome of the division does not depend on the definitions of y and z.

Proposition 5.1. *Let x/y and x/z be two real positive values defined as above. Then*

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) \Big/ \left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\max} = \frac{x}{x_{\max}}.$$

Proof:

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) = \left[\frac{1}{2}\left(\frac{x}{y}\right)^\theta + \frac{1}{2}\left(\frac{x}{z}\right)^\theta\right]^{1/\theta} = \left[\frac{x^\theta(y^\theta+z^\theta)}{2y^\theta z^\theta}\right]^{1/\theta} = \frac{x}{yz}\left[\frac{y^\theta+z^\theta}{2}\right]^{1/\theta}$$

and

$$\left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\max} = \frac{x_{\max}}{yz}\left[\frac{y^\theta+z^\theta}{2}\right]^{1/\theta}. \quad \square$$

An interesting consequence of Proposition 5.1 is the following property. Dividing the power mean of x/y and x/z by its maximum value gives the maximum function of x/y and x/z.

Corollary 5.1. *Let x/y and x/z be defined as above. If $x_{\max} = \min(y, z)$, then*

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) \Big/ \left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\max} = \lim_{\theta\to\infty} M_\theta\left(\frac{x}{y}, \frac{x}{z}\right).$$

As a first example, consider the power mean of

$$x = \frac{a}{p_1} \quad \text{and} \quad y = \frac{a}{p_2}.$$

Because $a_{\max} = \min(p_1, p_2)$, we have

$$\frac{M_\theta(x, y)}{[M_\theta(x, y)]_{\max}} = \lim_{\theta\to\infty} M_\theta\left(\frac{a}{p_1}, \frac{a}{p_2}\right) = \frac{a}{\min(p_1, p_2)} = S_{\text{Sim}}.$$

As a second example, consider the power mean of

$$x = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_1}.$$

Since $(ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1)$, we have

$$\frac{M_\theta(x, y)}{[M_\theta(x, y)]_{\max}} = \lim_{\theta\to\infty} M_\theta\left(\frac{ad - bc}{p_1 q_2}, \frac{ad - bc}{p_2 q_1}\right) = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)} = S_{\text{Loe}}.$$

As a third example, consider the power mean of the quantities

$$x = \frac{ad - bc}{p_1 q_1} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_2} \qquad \text{(see Peirce, 1884).}$$

Then

$$\begin{aligned}
M_{-1}(x, y) &= \frac{2(ad - bc)}{p_1 q_1 + p_2 q_2} = S_{\text{MP}} && \text{(harmonic mean)} \\
\lim_{\theta\to 0} M_\theta(x, y) &= \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}} = S_{\text{Phi}} && \text{(geometric mean)} \\
M_1(x, y) &= \frac{(ad - bc)(p_1 q_1 + p_2 q_2)}{2 p_1 q_2 p_2 q_1} = S_{\text{Fleiss}} && \text{(arithmetic mean).}
\end{aligned}$$

In light of Corollary 5.1, because $(ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1)$, which is different from $\min(p_1 q_1, p_2 q_2)$, we have

$$\begin{aligned}
\frac{M_\theta(x, y)}{[M_\theta(x, y)]_{\max}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)} &\neq \lim_{\theta\to\infty} M_h\left(\frac{ad - bc}{p_1 q_1}, \frac{ad - bc}{p_2 q_2}\right) \\
&= \frac{ad - bc}{\min(p_1 q_1, p_2 q_2)}.
\end{aligned}$$

Thus, the power mean of these x and y becomes S_{Loe}, although the latter coefficient is not a special case of the power mean.

Instead of considering power means, correction (5.1) can also be approached from a different angle. Below, two assertions are presented with respect to coefficients S_{Sim} and S_{Loe}.

Proposition 5.2. *Let $S = a/x$ with x a function of p_1 and p_2. Then*

$$S/\left[S\right]_{\max} = \frac{a}{\min(p_1, p_2)} = S_{\text{Sim}}.$$

Proof:

$$\left[S\right]_{\max} = \left[\frac{a}{x}\right]_{\max} = \frac{a_{\max}}{x} = \frac{\min(p_1, p_2)}{x}. \quad \text{Hence} \quad S/\left[S\right]_{\max} = S_{\text{Sim}}. \quad \square$$

Proposition 5.3. *Let $S = (ad - bc)/x$ with x a function of p_1 and p_2. Then $S/\left[S\right]_{\max} = S_{Loe}$.*
Proof:

$$\left[S\right]_{\max} = \left[\frac{ad - bc}{x}\right]_{\max} = \frac{(ad - bc)_{\max}}{x} = \frac{\min(p_1 q_2, p_2 q_1)}{x}.$$

Hence $S/\left[S\right]_{\max} = S_{\text{Loe}}$. $\square$

5.3 Correction for minimum value

In addition to the maximum value $\left[S\right]_{\max}$ of a coefficient S, one may study the minimum value $\left[S\right]_{\min}$. For coefficients that are special cases of the power mean of the quantities

$$S_{\text{Dice1}} = \frac{a}{p_1} \qquad \text{and} \qquad S_{\text{Dice2}} = \frac{a}{p_2}$$

the minimum value 0 is obtained if $a = 0$. Similarly, coefficients of the form a/x where x is a function of p_1 and p_2, equal 0 whenever $a = 0$. Thus, for this type of coefficients the minimum value is not constrained by the marginals. The section is therefore restricted to the minimum value of coefficients with the covariance $(ad-bc)$ in the numerator. For this class of coefficients the minimum value is obtained if either quantity a, d, or both equal zero. Hence, with unequal marginals $p_1 \neq q_1$, the 2×2 contingency table has the form

0	b	p_1
c	d	q_1
p_2	q_2	1

for example $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

if $a = 0$,

or

a	b	p_1
c	0	q_1
p_2	q_2	1

for example $\begin{bmatrix}1\\1\\1\\1\\0\\0\end{bmatrix}$ and $\begin{bmatrix}0\\0\\0\\1\\1\\1\end{bmatrix}$

if $d = 0$. The minimum covariance of two binary variables given marginal proportions p_1 and p_2, denoted $(ad - bc)_{\min}$, is thus given by

$$(ad - bc)_{\min} = \begin{cases} -p_1 p_2 & \text{if} \quad a = 0 \\ -q_1 q_2 & \text{if} \quad d = 0 \end{cases}$$

which equals

$$(ad - bc)_{\min} = \max(-p_1 p_2, -q_1 q_2) = -\min(p_1 p_2, q_1 q_2).$$

Thus, the minimum value of the covariance can only be obtained if $p_1 p_2 = q_1 q_2$ if and only if $p_1 + p_2 = 1$.

With correction for the minimum value the following issue must be taken into consideration. Because the quantity $(ad - bc)_{\min}$ is negative, division of a coefficient by $(ad - bc)_{\min}$ results in a change of sign. However, the minimum value of -1 can be obtained if the quantity $\min(p_1 p_2, q_1 q_2)$ is used instead of $-\min(p_1 p_2, q_1 q_2)$.

Similar as in the previous section, let x/y and x/z be two real positive values, of which the minimum depends on x only, that is

$$\left[\frac{x}{y}\right]_{\min} = \frac{x_{\min}}{y} \quad \text{and} \quad \left[\frac{x}{z}\right]_{\min} = \frac{x_{\min}}{z}.$$

Similar to $S/\left[S\right]_{\max}$, the outcome of $S/\left[S\right]_{\min}$ does not depend on the definitions of y and z with respect to power means. The proof of the next result is similar to the proof of Proposition 5.1.

Proposition 5.4. *Let x/y and x/z be two real positive values defined as above. Then*

$$M_\theta\left(\frac{x}{y}, \frac{x}{z}\right) / \left[M_\theta\left(\frac{x}{y}, \frac{x}{z}\right)\right]_{\min} = \frac{x}{|x_{\min}|}.$$

As a first example, consider the power mean of

$$x = \frac{ad - bc}{p_1 q_1} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_2}.$$

We have

$$\frac{M_\theta(x, y)}{|[M_\theta(x, y)]_{\min}|} = \lim_{\theta \to \infty} M_\theta \left(\frac{ad - bc}{p_1 q_1}, \frac{ad - bc}{p_2 q_2} \right) = \frac{ad - bc}{\min(p_1 p_2, q_1 q_2)}$$

which is a special case of the power mean. As a second example, consider the power mean of

$$x = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad y = \frac{ad - bc}{p_2 q_1}.$$

Again, we obtain

$$\frac{M_\theta(x, y)}{|[M_\theta(x, y)]_{\min}|} = \lim_{\theta \to \infty} M_\theta \left(\frac{ad - bc}{p_1 q_2}, \frac{ad - bc}{p_2 q_1} \right) = \frac{ad - bc}{\min(p_1 p_2, q_1 q_2)}$$

which is not a special case of this power mean.

We end this chapter with an argument made in Davenport and El-Sanhurry (1991). These authors argue that studying the minimum of $(ad - bc)$ is somewhat trivial. The minimum problem can be turned into a maximum problem at any time, simply by recoding the values of one of the binary variables. Maximum and minimum of $(ad - bc)$ are given by

$$(ad - bc)_{\max} = \min(p_1 q_2, p_2 q_1) \quad \text{and} \quad (ad - bc)_{\min} = -\min(p_1 p_2, q_1 q_2).$$

Suppose that the observations of the second variable are recoded, $1 \to 0$ and $0 \to 1$, for example

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \to \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Note that the recoding changes the sign of the covariance $(ad - bc)$ between the two binary vectors. Furthermore, for the second vector $p_2 \to q_2$ and $q_2 \to p_2$. Multiplying $(ad - bc)_{\min}$ by -1 and changing the roles of p_2 and q_2 in $(ad - bc)_{\min}$, we obtain $(ad - bc)_{\max}$.

5.4 Epilogue

In this chapter it was shown that various coefficients become equivalent if they are divided by their maximum value given fixed marginal probabilities p_1 and p_2. For example, the power mean of the quantities

$$S_{\text{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{p_2}$$

has as special cases

$$\begin{aligned} S_{\text{BB}} &= \frac{a}{\max(p_1, p_2)} \\ S_{\text{Gleas}} &= \frac{2a}{p_1 + p_2} \\ S_{\text{DK}} &= \frac{a}{\sqrt{p_1, p_2}} \\ \text{and} \quad S_{\text{Kul}} &= \frac{1}{2}\left[\frac{a}{p_1} + \frac{a}{p_2}\right]. \end{aligned}$$

By Proposition 5.1, S_{BB}, S_{Gleas}, S_{DK} and S_{Kul} coincide after correction for maximum value. Furthermore, by Corollary 5.1 all special cases of the power mean become equivalent to the maximum function (also a special case) of the two quantities. For example, S_{BB}, S_{Gleas}, S_{DK} and S_{Kul} become

$$S_{\text{Sim}} = \max\left(\frac{a}{p_1}, \frac{a}{p_2}\right) = \frac{a}{\min(p_1, p_2)}$$

after correction (5.1). As a second example, by Proposition 5.1 and Corollary 5.1,

$$S_{\text{Cohen}} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1} \quad \text{and} \quad S_{\text{Phi}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}}$$

are special cases of the power mean of

$$S_{\text{Cole1}} = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad - bc}{p_2 q_1}.$$

Coefficient S_{Cohen} and S_{Phi} become

$$S_{\text{Loe}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)}$$

after correction for maximum value. Moreover, by Proposition 5.3, S_{Cole1}, S_{Cole2},

$$S_{\text{MP}} = \frac{2(ad - bc)}{p_1 q_1 + p_2 q_2} \quad \text{and} \quad S_{\text{Fleiss}} = \frac{(ad - bc)(p_1 q_1 + p_2 q_2)}{2 p_1 q_2 p_2 q_1}$$

also become equivalent to S_{Loe}, after division by their maximum value given fixed marginals p_1 and p_2.

5.5 Loevinger's coefficient

Correction for chance and correction for maximum value were treated separately in Chapters 4 and 5. This section is used to show two properties of

$$S_{\text{Loe}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)}$$

the coefficient by Loevinger (1947, 1948), with respect to correction for chance and correction for maximum value simultaneously. With respect to both properties it is assumed that $E(a)_{\text{Cohen}} = p_1 p_2$ is the appropriate expectation.

First of all, if $E(a) = p_1 p_2$ and $a_{\max} = \min(p_1, p_2)$, then coefficient S_{Loe} can be defined as

$$S_{\text{Loe}} = \frac{a - E(a)}{a_{\max} - E(a)}$$

or dually

$$S_{\text{Loe}} = \frac{d - E(d)}{d_{\max} - E(d)}$$

where $E(d) = q_1 q_2$ and $d_{\max} = \min(q_1, q_2)$. Furthermore, under the same conditions, any coefficient in the $\mathcal{L}$ family (of the form $\lambda + \mu a$) becomes S_{Loe} after correction for maximum value and correction for chance. Moreover, the result does not depend on what correction is considered first.

Proposition 5.5. *A coefficient of the form $\lambda + \mu a$ becomes S_{Loe} after correction* (4.1) *and* (5.1).

Proof: Dividing coefficient $\lambda + \mu a$ by its maximum value given fixed marginals p_1 and p_2, we obtain

$$\frac{\lambda + \mu a}{\lambda + \mu \min(p_1, p_2)}. \tag{5.2}$$

The expectation of (5.2) is given by

$$E\left[\frac{\lambda + \mu a}{\lambda + \mu \min(p_1, p_2)}\right] = \frac{\lambda + \mu E(a)}{\lambda + \mu \min(p_1, p_2)} = \frac{\lambda + \mu p_1 p_2}{\lambda + \mu \min(p_1, p_2)}. \tag{5.3}$$

Using (5.2) and (5.3) in (4.1), and multiplying by $\lambda + \mu \min(p_1, p_2)$, we obtain

$$\frac{\lambda + \mu a - \lambda - \mu p_1 p_2}{\lambda + \mu \min(p_1, p_2) - \lambda - \mu p_1 p_2} = \frac{a - p_1 p_2}{\min(p_1, p_2) - p_1 p_2} = S_{\text{Loe}}.$$

Alternatively, Using $\lambda + \mu a$ and the corresponding expectation

$$\lambda + \mu p_1 p_2$$

in (4.1), we obtain

$$\frac{\lambda + \mu a - \lambda - \mu p_1 p_2}{1 - \lambda - \mu p_1 p_2} = \frac{a - p_1 p_2}{(1 - \lambda)/\mu - p_1 p_2}. \tag{5.4}$$

The maximum value of (5.4) given fixed marginals p_1 and p_2, is given by

$$\frac{\min(p_1, p_2) - p_1 p_2}{(1 - \lambda)/\mu - p_1 p_2}. \tag{5.5}$$

Dividing (5.4) by (5.5), we obtain

$$\frac{a - p_1 p_2}{\min(p_1, p_2) - p_1 p_2} = S_{\text{Loe}}.$$

This completes the proof. □

Zero value under statistical independence, and maximum value unity independent of the marginal distributions, are two properties or desiderata that similarity coefficients may have in general. Proposition 5.5 shows that the linear transformations that set the value under independence at zero (4.1) and the maximum value at unity (5.1), transform all coefficients in $\mathcal{L}$ family (of the form $\lambda + \mu a$) into the same underlying coefficient. This coefficient happens to be S_{Loe}.

Part II

Similarity matrices

CHAPTER 6

Data structures

In this chapter the basic notation that will be used in Part II is introduced. In Part I the data consisted of two binary sequences or variables. In Part II the data are collected in a data matrix $\mathbf{X}$ of m column vectors. In this chapter we do not consider individual coefficients but coefficient matrices. Given a $n \times m$ data matrix $\mathbf{X}$, one may obtain a $m \times m$ coefficient matrix $\mathbf{S}$ by calculating all pairwise coefficients S_{jk} for two columns j and k from $\mathbf{X}$. Different coefficient matrices are obtained, depending on the choice of similarity coefficient.

Chapter 6 is used to introduce several data structures that are either reflected in the data matrix or that can be assumed to underlie the data matrix. In the latter case, matrix $\mathbf{X}$ may contain the realizations, 0 or 1, generated by a latent variable model. The latent variable models presented in this chapter are discussed in terms of item response theory (De Gruijter and Van der Kamp, 2008; Van der Linden and Hambleton, 1997; Sijtsma and Molenaar, 2002).

Suppose the data matrix $\mathbf{X}$ contains the responses of n persons on m binary items. Item response theory is a psychometric approach that enables us to study these data in terms of item characteristics and persons' propensities to endorse different items. A subfield of item response theory, so-called nonparametric item response theory (Sijtsma and Molenaar, 2002), is concerned with identifying modeling properties that follow from basic assumptions like a single latent variable or local independence. Often, if a particular model holds for the data at hand, then the columns of the data matrix can be ordered such that certain structure properties become apparent.

In addition to several probabilistic models, various possible patterns of 1s and 0s are described in this chapter. These data structures are referred to as Guttman items and Petrie matrices, and, if the data matrix is not too big, can be confirmed by visual inspection. The theoretical conditions considered and derived in this chapter are used in the remaining chapters of Part II as possible sufficient conditions for coefficient matrices to exhibit or not exhibit certain ordinal properties.

6.1 Latent variable models

Suppose the binary data are in a matrix $\mathbf{X}$ of size $n \times m$. For example, the data may be the responses of n persons on m binary items. Let ω denote a single latent variable or trait and let $p_j(\omega)$ denote the response function corresponding to the response 1 in column vector j, with $0 \leq p_j(\omega) \leq 1$. The response 0 on j is modeled by the function $1 - p_j(\omega)$. Moreover, let $L(\omega)$ denote the distribution function of the latent variable ω. The unconditional probability of a score 1 on vector j is given by

$$p_j = \int_{\mathbb{R}} p_j(\omega) dL(\omega)$$

where $\mathbb{R}$ denotes the set of reals. We also define the quantity $q_j = 1 - p_j$.

At this point assume local independence, that is, conditionally on ω the responses of a person on the m items are stochastically independent. The joint probability of items j and k for a value of ω is then given by $p_j(\omega)p_k(\omega)$. The corresponding unconditional probability can be obtained from

$$a_{jk} = \int_{\mathbb{R}} p_j(\omega)p_k(\omega) dL(\omega).$$

In item response theory (De Gruijter and Van der Kamp, 2008; Van der Linden and Hambleton, 1997; Sijtsma and Molenaar, 2002) a distinction is made between so-called parametric and nonparametric models. In a parametric model a specific shape of the response function is assumed. An example of a parametric model is the 2-parameter model. The normal ogive formulation of the 2-parameter model comes from Lord (1952). Birnbaum (1968) later on proposed the logistic form of the 2-parameter model. A response function of the latter formulation is given by

$$p_j(\omega) = \frac{\exp[\delta_j(\omega - \beta_j)]}{1 + \exp[\delta_j(\omega - \beta_j)]}$$

where δ_j controls the slope of the response function and β_j controls the location of the response function.

In nonparametric models no shapes of the response function are assumed, only a general tenor for a set of functions. For example, all functions may be non-increasing in the latent variable, or they are unimodal functions. An example of a nonparametric model is the following model. Suppose that the response functions of all m items are monotonically increasing on ω, that is

$$p_j(\omega_1) \leq p_j(\omega_2) \quad \text{for} \quad 1 \leq j \leq m \quad \text{and} \quad \omega_1 < \omega_2. \tag{6.1}$$

The case in (6.1) (together with the assumptions of a single latent variable and local independence) describes the monotone homogeneity model in Sijtsma and Molenaar (2002, p. 22). A well-known result is that if (6.1) holds, then all binary items are positively dependent. The result follows from the fact that

$$a_{jk} - p_j p_k = \frac{1}{2} \int \int_{\mathbb{R}^2} [p_j(\omega_2) - p_j(\omega_1)] \, [p_k(\omega_2) - p_k(\omega_1)] \, dL(\omega_2) dL(\omega_1) > 0.$$

A stronger nonparametric model is the following model. In addition to (6.1), suppose that the items can be ordered such that the corresponding response functions are non-intersecting, that is,

$$p_j(\omega) \geq p_k(\omega) \quad \text{for} \quad 1 \leq j < k \leq m. \tag{6.2}$$

The case that assumes (6.1) and (6.2) (together with the assumptions of local independence and a single latent variable) is called the double monotonicity model in Sijtsma and Molenaar (2002, p. 23). A well-known result is that, if the double monotonicity model holds, then the items can be ordered such that

$$p_j \geq p_{j+1} \quad \text{for} \quad 1 \leq j < m \tag{6.3}$$

and

$$a_{jk} \geq a_{j+1k} \quad \text{for fixed } k \ (\neq j+1) \quad \text{and} \quad 1 \leq j < m. \tag{6.4}$$

Thus, under the double monotonicity model the item ordering can directly be obtained by inspecting the p_j. A parametric model that satisfies both requirement (6.1) and (6.2) is the 1-parameter logistic model or Rasch model (Rasch, 1960). The response function of the Rasch model is given by

$$p_j(\omega) = \frac{\exp[\omega - \beta_j]}{1 + \exp[\omega - \beta_j]}$$

where β_j controls the location of the individual response function. Note that the Rasch (1960) model is a special case of the 2-parameter logistic model.

Instead of a monotonically increasing function, let $p_j(\omega)$ be a unimodal function, that is

$$\begin{aligned} & p_j(\omega_1) \leq p_j(\omega_2) \quad \text{for} \quad \omega_1 < \omega_2 \leq \omega_0 \\ \text{and} \quad & p_j(\omega_1) \geq p_j(\omega_2) \quad \text{for} \quad \omega_0 \geq \omega_1 < \omega_2 \end{aligned}$$

where $p_j(\omega)$ obtains its maximum at ω_0. The class of models with unimodal response functions includes models with monotone response functions, since the latter can be interpreted as unimodal functions of which the maximum lies at plus or minus infinity.

Apart from being monotone or unimodal, response functions may also satisfy various orders of total positivity (Karlin, 1968; Post and Snijders, 1993). If a set of response functions is totally positive of order 2, then the items can be ordered such that

$$p_j(\omega_1) p_k(\omega_2) - p_j(\omega_2) p_k(\omega_1) \geq 0 \quad \text{for} \quad \omega_1 < \omega_2 \quad \text{and} \quad 1 \leq j < k \leq m. \tag{6.5}$$

Schriever (1986, p. 125) derived the following result for functions that are both monotonically increasing and satisfy total positivity of order 2.

Theorem 6.1 [Schriever, 1986]. *If m response functions are ordered such that* (6.1) *and* (6.5) *hold, then the items satisfy*

$$1 \leq j < m, \quad 1 \leq k \leq m \quad \Rightarrow \quad \frac{a_{jk}}{p_j} \leq \frac{a_{j+1k}}{p_{j+1}} \quad \text{for fixed } k \ (\neq j+1). \tag{6.6}$$

Proof: $p_j^{-1} p_j(\omega)$ can be interpreted as a density with respect to the measure $dL(\omega)$, which by (6.5), is totally positive of order 2 and satisfies

$$\int_{\mathbb{R}} p_j^{-1} p_j(\omega) dL(\omega) \ = \ 1.$$

Since by (6.1), $p_k(\omega)$ is increasing in ω for each $k = 1, ..., m$, it follows from Proposition 3.1 in Karlin (1968, p. 22) that

$$p_j^{-1} a_{jk} = \int_{\mathbb{R}} p_j^{-1} p_j(\omega) p_k(\omega) dL(\omega) \quad \text{is increasing in } j. \quad \square$$

6.2 Petrie structure

Coombs (1964) describes a model in which the unimodal response functions consists of two step functions. Characteristic of the Coombs scale is that the columns of $\mathbf{X}$ can be ordered such that all rows of the data matrix $\mathbf{X}$ contain consecutive 1s, that is, all the 1s in a row are bunched together. If the data matrix $\mathbf{X}$ is a reordered subject by attribute table with consecutive 1s in each row, all subjects have single-peaked preference functions, that is, they always check contiguous stimuli. If all runs of ones have the same length, the table has a parallelogram structure as defined by Coombs (1964, Chapter 4).

A (0,1)-table with consecutive 1s may also be interpreted as an intuitively meaningful and simple archaeological model. An artifact comes into use at a certain point in time, it remains in use for a certain period, and after some time it goes out of use. In an archaeological context, matrices with consecutive 1s were studied by Sir Flinders Petrie (Kendall, 1971, p. 215; Heiser, 1981, Section 3.2). Matrices with consecutive 1s in the rows will be called row Petrie. Column Petrie is defined in a similar way. A matrix is called double Petrie if it is both row Petrie and column Petrie. Examples of Petrie matrices are

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \mathbf{X}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$\mathbf{X}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Matrix $\mathbf{X}_1$ is row Petrie, whereas $\mathbf{X}_2$, $\mathbf{X}_3$ and $\mathbf{X}_4$ are double Petrie.

Determinants of any square 2×2 submatrix of a double Petrie matrix are positive. A double Petrie matrix is therefore totally positive of order 2 (Karlin, 1968). This property is used in Proposition 6.1, where $\mathbf{X}^T$ denote the transpose of the matrix $\mathbf{X}$. Moreover, let $\mathbf{S}_{\mathrm{RR}}$ denote the $m \times m$ similarity matrix containing all pairwise coefficients $S_{\mathrm{RR}} = a_{jk}$, calculated from the columns of $\mathbf{X}$.

Proposition 6.1. *If* $\mathbf{X}$ *is double Petrie, then*

$$\mathbf{S}_{\mathrm{RR}} = m^{-1}\mathbf{X}^T\mathbf{X}$$

is totally positive of order 2.

Proof: Because all possible second order-determinants of a double Petrie matrix, that is

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

their transposes, and

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

are either 1 or 0, a double Petrie matrix is (at least) totally positive of order 2. Since the product of two totally positive matrices of order h is again totally positive of order h (Gantmacher and Krein, 1950, p. 86), it follows that the matrix $\mathbf{S}_{\mathrm{RR}}$ is (at least) totally positive of order 2. □

We have a particular reason for studying Petrie matrices. It turns out that the data table $\mathbf{X}$ being row Petrie or double Petrie is manifested in the quantities

$$\begin{aligned} a_{jk} &= \text{the proportion of 1s shared by columns } j \text{ and } k \\ &\quad \text{in the same positions} \\ p_j &= \text{the proportion of 1s in column } j \\ \text{and } \quad p_k &= \text{the proportion of 1s in column } k. \end{aligned}$$

We present various properties in this section of quantities a_{jk}, p_j and p_k that hold if $\mathbf{X}$ reflects some sort of Petrie structure. We first consider the case that $\mathbf{X}$ is row Petrie. In Proposition 6.2 it is derived what pattern a_{jk} exhibits when $\mathbf{X}$ is row Petrie.

Proposition 6.2. *If* $\mathbf{X}$ *is row Petrie, then*

$$\begin{aligned} & a_{jk} \geq a_{j+1k} \quad \textit{for} \quad 1 \leq k \leq j < m \\ \textit{and} \quad & a_{jk} \leq a_{j+1k} \quad \textit{for} \quad 1 \leq j < k \leq m. \end{aligned} \tag{6.7}$$

Proof: We only consider the proof of (6.7). If $\mathbf{X}$ is row Petrie then columns k, j and $j+1$ of $\mathbf{X}$ can form the two types of row profiles

k	j	$j+1$	freq.
1	1	0	u_1
1	1	1	u_2

with frequencies u_1 and u_2. Thus u_1 is the number of row profiles that contain a 1 for columns k and j and a 0 for column $j+1$. Equation (6.7) is true if

$$\begin{aligned} & a_{jk} \geq a_{j+1k} \\ & u_1 + u_2 \geq u_2 \\ & u_1 \geq 0. \end{aligned}$$

The assertion is true because u_1 is a positive number. □

In the remainder of the section we consider the case that $\mathbf{X}$ is double Petrie. We present several properties of quantities a_{jk}, p_j and p_k for the case that $\mathbf{X}$ is double Petrie.

Proposition 6.3. *If* $\mathbf{X}$ *is double Petrie, then*

$$\begin{aligned} & \frac{a_{jk}}{p_j} \geq \frac{a_{j+1k}}{p_{j+1}} \quad \textit{for} \quad 1 \leq k \leq j < m \\ \textit{and} \quad & \frac{a_{jk}}{p_j} \leq \frac{a_{j+1k}}{p_{j+1}} \quad \textit{for} \quad 1 \leq j < k \leq m. \end{aligned} \tag{6.8}$$

Proof: We only consider the proof of (6.8). If $\mathbf{X}$ is double Petrie, we may distinguish two situations with respect to the types of row profiles of columns j, $j+1$, and k. Firstly, we have

k	j	$j+1$	freq.
1	1	0	u_1
0	1	0	u_2
0	1	1	u_3
0	0	1	u_4

with frequencies u_1 and u_4. In this case there are no row profiles with a 1 in both column k and $j+1$. Equation (6.8) is true if

$$\begin{aligned} & \frac{a_{jk}}{p_j} \geq \frac{a_{j+1k}}{p_{j+1}} \\ & \frac{u_1}{u_1 + u_2 + u_3} \geq \frac{0}{u_3 + u_4} \\ & u_1 \geq 0. \end{aligned}$$

Since u_1 is a positive number, (6.8) holds for the first situation. Secondly, we may have

k	j	$j+1$	freq.
1	1	0	u_1
1	1	1	u_2
0	1	1	u_3
0	0	1	u_4

with frequencies u_1 and u_4. With respect to the second case, (6.8) is true if

$$\begin{aligned}
&\frac{a_{jk}}{p_j} \geq \frac{a_{j+1k}}{p_{j+1}} \\
&\frac{u_1+u_2}{u_1+u_2+u_3} \geq \frac{u_2}{u_2+u_3+u_4} \\
&u_1u_2+u_1u_3+u_1u_4+u_2u_2+u_2u_3+u_2u_4 \geq u_1u_2+u_2u_2+u_2u_3 \\
&u_1u_3+u_1u_4+u_2u_4 \geq 0.
\end{aligned}$$

This completes the proof of the assertion. □

Proposition 6.4. *If* $\mathbf{X}$ *is double Petrie, then*

$$\frac{a_{jk}}{p_j+p_k} \geq \frac{a_{j+1k}}{p_{j+1}+p_k} \quad \text{for} \quad 1 \leq k \leq j < m \tag{6.9}$$

$$\text{and} \qquad \frac{a_{jk}}{p_j+p_k} \leq \frac{a_{j+1k}}{p_{j+1}+p_k} \quad \text{for} \quad 1 \leq j < k \leq m.$$

Proof: We only consider the proof of (6.9). Since $\mathbf{X}$ is double Petrie, we have

$$p_{j+1}a_{jk} \geq a_{j+1k}p_j \quad \text{for} \quad 1 \leq k \leq j < m \tag{6.10}$$

by Proposition 6.3 and

$$p_k a_{jk} \geq p_k a_{j+1k} \quad \text{for} \quad 1 \leq k \leq j < m \tag{6.11}$$

by Proposition 6.2. Adding (6.10) and (6.11) we obtain (6.9). □

6.3 Guttman items

The simplest data structure considered in this chapter is the Guttman or perfect scale (Guttman, 1950, 1954), named after the person who popularized the model with the method of scalogram analysis. A scalogram matrix is a special type of double Petrie matrix, for which all pairs of columns are Guttman items. Let p_j (q_j) denote the proportion of 1s (0s) of variable j, and let a_{jk} denote the proportion of 1s that vector j and k share in the same positions. Two binary variables are Guttman items if the number of 1s that variables j and k share in the same positions equals the total amount of 1s in one of the vectors, that is,

$$a_{jk} = \min(p_j, p_k) \quad \text{for} \quad 1 \leq j \leq m \quad \text{and} \quad 1 \leq k \leq m. \tag{6.12}$$

Matrix $\mathbf{X}_4$ (Section 6.2) satisfies condition (6.12). Furthermore, the columns of $\mathbf{X}_4$ are ordered such that (6.3) holds. If the columns of $\mathbf{X}$ satisfy both (6.12) and (6.3), $\mathbf{X}$ is sometimes referred to as a scalogram. Scalogram matrices are totally positive, that is, the determinant of any square submatrix, including the minors, is positive (Karlin, 1968).

Various coefficients have specific properties if the data consist of Guttman items. If (6.12) holds, then the matrices $\mathbf{S}_{\text{Sim}} = \mathbf{S}_{\text{Loe}}$ have elements $S_{\text{Sim}} = S_{\text{Loe}} = 1$. For example, $\mathbf{S}_{\text{Sim}} = \mathbf{S}_{\text{Loe}}$ corresponding to matrix $\mathbf{X}_4$ is given by

$$\mathbf{S}_{\text{Sim}} = \mathbf{S}_{\text{Loe}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Furthermore, if (6.12) and (6.3) hold, then the elements of the similarity matrices $\mathbf{S}_{\text{Dice1}} = \{a_{jk}/p_j\}$ and $\mathbf{S}_{\text{Dice2}} = \{a_{jk}/p_k\}$ have the form

$$S_{\text{Dice1}} = \begin{cases} p_j^{-1} p_k & \text{for } j < k \\ 1 & \text{for } j \geq k \end{cases}$$

and

$$S_{\text{Dice2}} = \begin{cases} 1 & \text{for } j \leq k \\ p_k^{-1} p_j & \text{for } j > k. \end{cases}$$

For example, coefficient matrices $\mathbf{S}_{\text{Dice1}}$ and $\mathbf{S}_{\text{Dice2}}$ corresponding to data matrix $\mathbf{X}_4$ in Section 6.2, are given by

$$\mathbf{S}_{\text{Dice1}} = \begin{bmatrix} 1 & .8 & .4 & .2 \\ 1 & 1 & .5 & .25 \\ 1 & 1 & 1 & .5 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{S}_{\text{Dice2}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ .8 & 1 & 1 & 1 \\ .4 & .5 & 1 & 1 \\ .2 & .25 & .5 & 1 \end{bmatrix}.$$

Similarly, the elements of the similarity matrices $\mathbf{S}_{\text{Cole1}}$ and $\mathbf{S}_{\text{Cole2}}$ have the form

$$S_{\text{Cole1}} = \begin{cases} (p_j q_k)^{-1} p_k q_j & \text{for } j < k \\ 1 & \text{for } j \geq k \end{cases}$$

and

$$S_{\text{Cole2}} = \begin{cases} 1 & \text{for } j \leq k \\ (p_k q_j)^{-1} p_j q_k & \text{for } j > k. \end{cases}$$

A matrix $\mathbf{S}$ is said to be a Green's matrix (Karlin, 1968, p. 110) if its elements can be expressed in the form

$$S_{jk} = u_{\min(j,k)} v_{\max(j,k)} = \begin{cases} u_j\, v_k & \text{for } j \leq k \\ u_k\, v_j & \text{for } j \geq k \end{cases}$$

where u_j and v_k for $j, k = 1, 2, ..., m$ are real constants. Green's matrices are totally positive, that is, the determinant of any square submatrix, including the minors, is positive. These matrices have a variety of interesting properties (cf. Karlin, 1968). Various similarity matrices corresponding to different coefficients become Green's matrices if the data are Guttman items.

Proposition 6.5. *If the columns of* $\mathbf{X}$ *are ordered such that* (6.12) *and* (6.3) *hold, then* $\mathbf{S}_{\text{RR}}$, $\mathbf{S}_{\text{DK}}$, $\mathbf{S}_{\text{BB}} = \mathbf{S}_{\text{Jac}} = \mathbf{S}_{\text{Sorg}}$ and $\mathbf{S}_{\text{Phi}}$ *are Green's matrices.*
Proof: If $a_{jk} = \min(p_j, p_k)$ and $p_j \geq p_{j+1}$, then

$$S_{\text{RR}} = \begin{cases} p_k & \text{for } j \leq k \\ p_j & \text{for } j \geq k \end{cases}$$

$$S_{\text{DK}} = \begin{cases} p_j^{-1/2} p_k^{1/2} & \text{for } j < k \\ 1 & \text{for } j = k \\ p_k^{-1/2} p_j^{1/2} & \text{for } j > k \end{cases}$$

$$S_{\text{BB}} = S_{\text{Jac}} = S_{\text{Sorg}} = \begin{cases} p_j^{-1} p_k & \text{for } j < k \\ 1 & \text{for } j = k \\ p_k^{-1} p_j & \text{for } j > k \end{cases}$$

$$S_{\text{Phi}} = \begin{cases} (p_j q_k)^{-1/2} (p_k q_j)^{1/2} & \text{for } j < k \\ 1 & \text{for } j = k \\ (p_k q_j)^{-1/2} (p_j q_k)^{1/2} & \text{for } j > k. \end{cases} \quad \square$$

6.4 Epilogue

This chapter was used to introduce several data structures that are either reflected in the data matrix or that can be assumed to underlie the data matrix. In the latter case, data matrix $\mathbf{X}$ may contain the realizations, 0 or 1, generated by a latent variable model. It was shown that if $\mathbf{X}$ exhibits some sort of Petrie structure or if a certain latent variable model can be assumed to underlie data matrix $\mathbf{X}$, then this data structure is manifested in the quantities

$$\begin{aligned} a_{jk} &= \text{the proportion of 1s shared by columns } j \text{ and } k \\ &\quad \text{in the same positions} \\ p_j &= \text{the proportion of 1s in column } j \\ \text{and } \quad p_k &= \text{the proportion of 1s in column } k. \end{aligned}$$

The properties of the manifest probabilities derived in this chapter are used in the later chapters of the Part II as possible sufficient conditions for coefficient matrices to exhibit or not certain ordinal properties.

CHAPTER 7

Robinson matrices

Given a $n \times m$ data matrix $\mathbf{X}$ one may obtain a $m \times m$ coefficient matrix by calculating all pairwise coefficients for two columns j and k of $\mathbf{X}$. Different similarity matrices are obtained depending on the choice of similarity coefficient. Various matrix properties of coefficient matrices may be studied. The topic of this chapter is Robinson matrices.

A square similarity matrix $\mathbf{S}$ is called a Robinson matrix (after Robinson, 1951) if the highest entries within each row and column of $\mathbf{S}$ are on the main diagonal (elements S_{jj}) and moving away from this diagonal, the entries never increase. The Robinson property of a (dis)similarity matrix reflects an ordering of the objects, but also constitutes a clustering system with overlapping clusters. Such ordered clustering systems were introduced under the name pyramids by Diday (1984, 1986) and under the name pseudo-hierarchies by Fichet (1984). The CAP algorithm to find an ordered clustering structure was described in Diday (1986) and Diday and Bertrand (1986), and later extended to deal with symbolic data by Brito (1991) and with missing data by Gaul and Schader (1994). Chepoi and Fichet (1997) describe several circumstances in which Robinson matrices are encountered. For an in-depth review of overlapping clustering systems the reader is referred to Barthélemy, Brucker and Osswald (2004).

A similarity matrix may or may not exhibit the Robinson property depending on the choice of resemblance measure. It seems to be a common notion in the classification literature that Robinson matrices arise naturally in problems where there is essentially a one-dimensional structure in the data (see, for example, Critchley, 1994, p. 174). As will be shown in this chapter, the occurrence of a Robinson matrix is a combination of the choice of the similarity coefficient, and the specific one-dimensional structure in the data. Here, the data structures from Chapter 6 come into play. In this chapter it is specified in terms of sufficient conditions what data structure must be reflected in the data matrix $\mathbf{X}$ for a corresponding similarity matrix to exhibit the Robinson property. The Robinson property is primarily studied for coefficient matrices that are symmetric. Chapter 19 is devoted to a three-way generalization of Robinson matrix, called a Robinson cube.

7.1 Auxiliary results

When studying symmetric coefficient matrices, it is convenient to work with the following definition of a Robinson matrix. A symmetric matrix $\mathbf{S} = \{S_{jk}\}$ is called a Robinson matrix if we have

$$S_{jk} \leq S_{j+1k} \qquad \text{for } 1 \leq j < k \leq m \tag{7.1}$$

$$S_{jk} \geq S_{j+1k} \qquad \text{for } 1 \leq k \leq j < m. \tag{7.2}$$

In this first section we present several auxiliary results without proof. These results may be used to establish Robinson properties for other coefficients once a property has been established for some resemblance measures.

Proposition 7.1. *Coefficient matrix* $\mathbf{S}$ *with elements* S_{jk} *is a Robinson matrix if and only if the coefficient matrix with elements* $2S_{jk} - 1$ *is a Robinson matrix.*

Coefficients that are related by the formula in Proposition 7.1 are $S_{\text{Ham}} = 2S_{\text{SM}} - 1$ where

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \quad \text{and} \quad S_{\text{Ham}} = \frac{a-b-c+d}{a+b+c+d}$$

(Hamann, 1961) and $S_{\text{McC}} = 2S_{\text{Kul}} - 1$ where

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \quad \text{and} \quad S_{\text{McC}} = \frac{a^2-bc}{(a+b)(a+c)}$$

(McConnaughey, 1964).

Proposition 7.2. *If* $\mathbf{S}_i$ *for* $i = 1, 2, ..., n$ *are* n *Robinson matrices of order* $m \times m$*, then their sum (or their arithmetic mean) is also a Robinson matrix.*

Proposition 7.3. *If* $\mathbf{S} = \{S_{jk}\}$ *and* $\mathbf{S}^* = \{S^*_{jk}\}$ *are Robinson matrices of order* $m \times m$*, then matrix* $\mathbf{T}$ *with elements* $T_{jk} = S_{jk} \times S^*_{jk}$ *is a Robinson matrix.*

Proposition 7.4. *Let* $\mathbf{S} = \{S_{jk}\}$ *be a Robinson matrix, and let* $f()$ *be a monotonic function. Then matrix* $\mathbf{T}$ *with elements* $T_{jk} = f(S_{jk})$ *is a Robinson matrix.*

We also consider two propositions that are specific to parameter families $S_{\mathrm{GL1}}(\theta)$ and $S_{\mathrm{GL2}}(\theta)$.

Proposition 7.5. *Let* $\mathbf{S}$ *and* $\mathbf{S}^*$ *be coefficient matrices corresponding to any two members of* $S_{\mathrm{GL1}}(\theta)$. $\mathbf{S}$ *is a Robinson matrix if and only if* $\mathbf{S}^*$ *is a Robinson matrix.*
Proof: Due to Theorem 3.1, (7.1) and (7.2) for any member of $S_{\mathrm{GL1}}(\theta)$ become

$$\frac{a_{jk}}{p_j + p_k} \geq \frac{a_{j+1k}}{p_{j+1} + p_k} \quad \text{for} \quad 1 \leq k \leq j < m$$

$$\text{and} \qquad \frac{a_{jk}}{p_j + p_k} \leq \frac{a_{j+1k}}{p_{j+1} + p_k} \quad \text{for} \quad 1 \leq j < k \leq m. \quad \square$$

Proposition 7.6. *Let* $\mathbf{S}$ *and* $\mathbf{S}^*$ *be coefficient matrices corresponding to any two members of* $S_{\mathrm{GL2}}(\theta)$. $\mathbf{S}$ *is a Robinson matrix if and only if* $\mathbf{S}^*$ *is a Robinson matrix.*
Proof: Due to Theorem 3.2, (7.1) and (7.2) for any member of $S_{\mathrm{GL2}}(\theta)$ become

$$2a_{jk} - p_j \geq 2a_{j+1k} - p_{j+1} \quad \text{for} \quad 1 \leq k \leq j < m$$

$$\text{and} \qquad 2a_{jk} - p_j \leq 2a_{j+1k} - p_{j+1} \quad \text{for} \quad 1 \leq j < k \leq m. \quad \square$$

7.2 Braun-Blanquet + Russel and Rao coefficient

Coefficient

$$S_{\mathrm{BB}} = \frac{a_{jk}}{\max(p_j, p_k)} \qquad \text{(Braun-Blanquet, 1932)}$$

is one of the few interesting measures with respect to the Robinson property. It was shown in Chapter 2 that S_{BB} is a special case of a coefficient used by Robinson (1951) (Proposition 2.1). The Robinson property of coefficient S_{BB} is related to latent variable models with monotonically increasing response functions. The coefficient matrix corresponding to $\mathbf{S}_{\mathrm{BB}}$ is a Robinson matrix if $p_j \geq p_{j+1}$ (6.3), $a_{jk} \geq a_{j+1k}$ (6.4), and $p_j^{-1} a_{jk} \geq p_{j+1}^{-1} a_{j+1k}$ (6.6) hold. Condition (6.4) holds under the double monotonicity model (Sijtsma and Molenaar, 2002). Condition (6.6) was derived by Schriever (1986) for increasing response function that are totally positive of order 2.

Proposition 7.7. *Suppose the* m *columns of* $\mathbf{X}$ *are ordered such that* (6.3), (6.4) *and* (6.6) *hold. Then* $\mathbf{S}_{\mathrm{BB}}$ *with* $S_{\mathrm{BB}} = a_{jk}/\max(p_j, p_k)$ *is a Robinson matrix.*
Proof: Suppose (6.3) holds. Using S_{BB} in (7.1) and (7.2) we obtain

$$\frac{a_{jk}}{p_j} \leq \frac{a_{j+1k}}{p_{j+1}} \quad \text{for } 1 \leq j < k \leq m \quad \text{and} \quad a_{jk} \geq a_{j+1k} \quad \text{for } 1 \leq k \leq j < m.$$

The conditions are satisfied if (6.6) and (6.4) hold. $\square$

The coefficient by Russel and Rao (1940) $S_{\text{RR}} = a_{jk}$ is by far the simplest coefficient for binary data considered in this thesis. Nevertheless, S_{RR} is an interesting coefficient which possesses an interesting Robinson property. The result is not new, but can already be found in Wilkinson (1971). Coefficient matrix $\mathbf{S}_{\text{RR}}$ is a Robinson matrix if $\mathbf{X}$ is row Petrie.

Theorem 7.1 [Wilkinson, 1971, p. 279]. *If* $\mathbf{X}$ *is row Petrie, then* $\mathbf{S}_{\text{RR}}$ *with elements* S_{RR} *is a Robinson matrix.*
Proof 1: The result follows from Proposition 6.2.
Proof 2: Let $\mathbf{x}_i$ be the ith row of $\mathbf{X}$ and let $\mathbf{x}_i^T$ denotes its transpose. The matrix $\mathbf{S}_{\text{RR}}$ equals

$$\mathbf{S}_{\text{RR}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i.$$

If $\mathbf{X}$ is row Petrie, then each $\mathbf{x}_i^T \mathbf{x}_i$ is a Robinson matrix. Due to Proposition 7.2, the arithmetic mean of Robinson matrices is again a Robinson matrix. □

7.3 Double Petrie

A variety of coefficient matrices are Robinson matrices when $\mathbf{X}$ is double Petrie. Proposition 7.8 covers this Robinson property for parameter family $S_{\text{GL1}}(\theta)$. Proposition 7.9 concerns asymmetric coefficients S_{Dice1} and S_{Dice2}, whereas Proposition 7.10 concerns S_{Kul} and S_{DK}.

Proposition 7.8. *If* $\mathbf{X}$ *is double Petrie, then the coefficient matrix corresponding to any member of* $S_{\text{GL1}}(\theta)$ *is a Robinson matrix.*
Proof: The result follows from Proposition 7.5 and Proposition 6.4. □

Proposition 7.9. *If* $\mathbf{X}$ *is double Petrie, then* $\mathbf{S}_{\text{Dice1}}$ *and* $\mathbf{S}_{\text{Dice2}}$ *with elements* S_{Dice1} *and* S_{Dice2} *are Robinson matrices.*
Proof: We consider the proof for $\mathbf{S}_{\text{Dice1}}$ first. Since S_{Dice1} is not symmetric we ignore equations (7.1) and (7.2). We must verify the four directions one may move away from the main diagonal of $\mathbf{S}_{\text{Dice1}}$. We have

$$\frac{a_{jk}}{p_j} \geq \frac{a_{j+1k}}{p_{j+1}} \quad \text{for} \quad 1 \leq k \leq j < m$$

$$\text{and} \quad \frac{a_{jk}}{p_j} \leq \frac{a_{j+1k}}{p_{j+1}} \quad \text{for} \quad 1 \leq j < k \leq m.$$

By Proposition 6.3, both conditions are true if $\mathbf{X}$ is double Petrie. Furthermore, we have

$$\frac{a_{jk}}{p_j} \geq \frac{a_{jk+1}}{p_j} \quad \text{for} \quad 1 \leq k < j \leq m$$

$$\text{and} \quad \frac{a_{jk}}{p_j} \leq \frac{a_{jk+1}}{p_j} \quad \text{for} \quad 1 \leq j \leq k < m.$$

By Proposition 6.2, these conditions are true if $\mathbf{X}$ is double Petrie. This completes the proof for $\mathbf{S}_{\text{Dice1}}$. Because $\mathbf{S}_{\text{Dice2}}$ is the transpose of $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Dice2}}$ is a Robinson matrix if and only if $\mathbf{S}_{\text{Dice1}}$ has the Robinson property. □

Proposition 7.10. *If* $\mathbf{X}$ *is double Petrie, then* $\mathbf{S}_{\text{Kul}}$ *and* $\mathbf{S}_{\text{DK}}$ *with elements*

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \quad \textit{and} \quad S_{\text{DK}} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

are Robinson matrices.
Proof: The property follows from Proposition 7.9 combined with Proposition 7.2 for S_{Kul} and Propositions 7.3 and 7.4 with respect to coefficient S_{DK}. □

7.4 Restricted double Petrie

The two conditions considered in this section are restricted forms of a double Petrie structure. In Proposition 7.11 it is assumed that data table $\mathbf{X}$ satisfies the Guttman scale. Matrix $\mathbf{X}_4$ (Section 6.2) is an example of a Guttman scale. In Proposition 7.12 it is assumed that $\mathbf{X}$ is double Petrie and that $p_j = p_{j+1}$ for $1 \leq j < m$. Matrix $\mathbf{X}_3$ (Section 6.2) is an example of a data table that satisfies the conditions considered in Proposition 7.12. Because the conditions in Propositions 7.11 and 7.12 are quite restrictive, the results have limited applicability and are perhaps of theoretical interest only.

Proposition 7.11. *If the columns of* $\mathbf{X}$ *are ordered such that* (6.12) *and* (6.3) *hold, then* $\mathbf{S}_{\text{SM}}$ *with elements* S_{SM} *and* $\mathbf{S}_{\text{Phi}}$ *with elements* S_{Phi} *are Robinson matrices.*
Proof: Under condition (6.12), the equations of Proposition 7.6 become equivalent to condition (6.3). This completes the proof for coefficient S_{SM}.

Under condition (6.12), S_{Phi} can be written as

$$S_{\text{Phi}} = \begin{cases} \sqrt{\frac{p_k q_j}{p_j q_k}} & \text{for } j < k \\ \sqrt{\frac{p_j q_k}{p_k q_j}} & \text{for } j > k \end{cases} \tag{7.3}$$

and $S_{\text{Phi}} = 1$ if $j = k$.

Using (7.3) in (7.1) and (7.2) we obtain

$$\frac{q_j}{p_j} \leq \frac{q_{j+1}}{p_{j+1}} \quad \text{for } 1 \leq j < k \leq m$$

and

$$\frac{p_j}{q_j} \geq \frac{p_{j+1}}{q_{j+1}} \quad \text{for } 1 \leq k \leq j < m.$$

Both inequalities are true if (6.3) holds. This completes the proof for coefficient S_{Phi}. □

Proposition 7.12. *Let* $\mathbf{X}$ *be double Petrie and let* $p_j = p_{j+1}$ *for* $1 \leq j < m$. *Then* $\mathbf{S}_{\text{SM}}$ *with elements* S_{SM} *and* $\mathbf{S}_{\text{Phi}}$ *with elements* S_{Phi} *are Robinson matrices.*
Proof: If $p_j = p_{j+1}$ *for* $1 \leq j < m$, the equations of Proposition 7.6 become equivalent to the equations in Proposition 6.2. This completes the proof for coefficient S_{SM}. The proof for S_{Phi} is similar. □

7.5 Counterexamples

The Robinson property of S_{RR} established in Theorem 7.1 appears to be unique to S_{RR}. We consider a row Petrie counterexample for the Jaccard coefficient

$$S_{\text{Jac}} = \frac{a_{jk}}{p_j + p_k - a_{jk}}$$

which is a member of family $S_{\text{GL1}}(\theta)$, and the coefficient by Braun-Blanquet (1932)

$$S_{\text{BB}} = \frac{a_{jk}}{\max(p_j, p_k)}.$$

Let the data be in the matrix $\mathbf{X}_1$ from Section 6.2. Using $\mathbf{X}_1$, we may obtain coefficient matrices

$$\mathbf{S}_{\text{Jac}} = \begin{bmatrix} 1 & .33 & 0 & 0 \\ .33 & 1 & .17 & .20 \\ 0 & .17 & 1 & .40 \\ 0 & .20 & .40 & 1 \end{bmatrix} \qquad \mathbf{S}_{\text{BB}} = \begin{bmatrix} 1 & .33 & 0 & 0 \\ .33 & 1 & .25 & .33 \\ 0 & .25 & 1 & .50 \\ 0 & .33 & .50 & 1 \end{bmatrix}$$

and

$$\mathbf{S}_{\text{RR}} = \begin{bmatrix} .14 & .14 & 0 & 0 \\ .14 & .43 & .14 & .14 \\ 0 & .29 & .57 & .29 \\ 0 & .14 & .29 & .43 \end{bmatrix}.$$

The latter matrix is a Robinson matrix, but $\mathbf{S}_{\text{Jac}}$ and $\mathbf{S}_{\text{BB}}$ are not Robinson matrices.

Coefficient matrices corresponding to resemblance measures that include the covariance $(ad - bc)$ or the quantity d in the numerator do not appear to be Robinson matrices if $\mathbf{X}$ is double Petrie. For the simple matching coefficient $S_{\text{SM}} = (a + d)/(a + b + c + d)$ and the Phi coefficient

$$S_{\text{Phi}} = \frac{ad - bc}{\sqrt{p_j p_k q_j q_k}}$$

we consider a counterexample. Let the data be in the matrix $\mathbf{X}_2$ from Section 6.2. Using $\mathbf{X}_2$ we may obtain coefficient matrices

$$\mathbf{S}_{\text{SM}} = \begin{bmatrix} 1 & .5 & 0 & .25 \\ .5 & 1 & .5 & .25 \\ 0 & .5 & 1 & .75 \\ .25 & .25 & 75 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{S}_{\text{Phi}} = \begin{bmatrix} 1 & 0 & -1 & -.58 \\ 0 & 1 & 0 & -.58 \\ -1 & 0 & 1 & -.58 \\ -.58 & -.58 & -.58 & 1 \end{bmatrix}.$$

Both matrices are not Robinson matrices.

7.6 Epilogue

A coefficient matrix is referred to as a Robinson matrix if the highest entries within each row and column are on the main diagonal and moving away from this diagonal, the entries never increase. For a selection of resemblance measures for binary variables we presented sufficient conditions for the corresponding coefficient matrix to exhibit the Robinson property. As sufficient conditions we considered data tables that are referred to as Petrie matrices, that is, matrices of which the columns can be ordered such that the 1s in a row form a consecutive interval.

As it turns out, the sufficient conditions differ with the resemblance measures for (0,1)-data. The occurrence of a Robinson matrix is the interplay between the choice of similarity coefficient and the specific structure in the data at hand.

Some of the sufficient conditions can be ordered from restrictive to most general: Guttman scale $\Rightarrow$ double Petrie $\Rightarrow$ row Petrie. The latter condition is sufficient for the coefficient matrix corresponding to coefficient

$$S_{\mathrm{RR}} = \frac{a}{a+b+c+d} \qquad \text{(Russel and Rao, 1940)}$$

to be a Robinson matrix. Although this result was already presented in Wilkinson (1971), the systematic study presented in this chapter reveals that the Robinson property of S_{RR} is a very general Robinson property compared to the Robinson properties of other resemblance measures for binary variables. Furthermore, the general Robinson property appears to be unique to coefficient S_{RR}. Within the framework of Petrie matrices, we may conclude that the Robinson property is most likely to occur for the coefficient matrix $\mathbf{S}_{\mathrm{RR}}$.

The Guttman scale is also a special case of the Rasch model (see Section 6.1), which in turn is a special case of the model implied by (6.3), (6.4) and (6.6). In Section 7.2 it was shown that the latter model, that corresponds to a probabilistic model with monotonically increasing response functions, is sufficient for the coefficient matrix with elements

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

to be a Robinson matrix.

It should be noted that the results in this chapter are exact. For example, matrix $\mathbf{X}_1$ was used in Section 7.5 to show that the similarity matrix based on S_{Jac} is not a Robinson matrix for all row Petrie data matrices. Nevertheless, it may well as be that matrix $\mathbf{S}_{\mathrm{Jac}}$ is a Robinson matrix for many row Petrie data matrices, and that in many practical cases it has approximately the same properties as $\mathbf{S}_{\mathrm{RR}}$.

CHAPTER 8

Eigenvector properties

The eigendecomposition of matrices is used in various realms of research. In various domains of data analysis, calculating eigenvalues and eigenvectors of certain matrices characterizes various methods and techniques for exploratory data analysis. For example, exploratory methods that are so-called eigenvalue methods, are principal component analysis, homogeneity analysis (Gifi, 1990; Heiser, 1981; Meulman, 1982), classical scaling (Gower, 1966; Torgerson, 1958), or correspondence analysis (Greenacre, 1984; Heiser, 1981).

The topic of study in this chapter are the eigenvectors of similarity matrices corresponding to coefficients for binary data. Various results on the eigenvector elements of coefficient matrices are presented. It is shown that ordinal information can be obtained from eigenvectors corresponding to the largest eigenvalue of various similarity matrices. Using eigenvectors it is therefore possible to uncover correct orderings of various latent variable models. The point to be made here is that the eigendecomposition of some similarity matrices, especially matrices corresponding to asymmetric coefficients, are more interesting compared to the eigendecomposition of other matrices. Many of the results are perhaps of theoretical interest only, since no new insights are developed compared to existing methodology already available for various nonparametric item response theory models.

Homogeneity analysis is a generalization of principal component analysis to categorical data proposed by Guttman (1941). Various authors noted the specific (mathematical) properties of homogeneity analysis when it is applied to binary responses (Guttman, 1950, 1954; Heiser, 1981; Gifi, 1990; Yamada and Nishisato, 1993). If homogeneity analysis is applied to binary data, the category weights for a score 1 or 0 can be obtained as eigenvector elements of two separate matrices. As it turns out, the elements of these matrices have simple formulas. In the last section of this chapter some new insights on the mathematical properties of homogeneity analysis of binary data are presented.

8.1 Ordered eigenvector elements

In this first section the eigenvector corresponding to the largest eigenvalue of various coefficient matrices is studied. It is shown what ordinal information can be obtained from the eigenvector corresponding to the largest eigenvalue of these matrices. The inspiration for the study comes from a result presented in Schriever (1986) who considered the eigenvector corresponding to the first eigenvalue of the coefficient matrices with respective elements

$$S_{\text{Cole1}} = \frac{a_{jk} - p_j p_k}{p_j q_k} \quad \text{and} \quad S_{\text{Cole2}} = \frac{a_{jk} - p_j p_k}{p_k q_j} \quad \text{(Cole, 1949).}$$

Most of the tools used below, come from the proof presented in Schriever (1986). A specific result that will often be used when studying these properties, is the Perron-Frobenius theorem (Gantmacher, 1977, p. 53; Rao, 1973, p. 46). More precisely, only the following weaker version of the Perron-Frobenius theorem will be used.

Theorem 8.1. *If a square matrix* $\mathbf{S}$ *has strictly positive elements, then the eigenvector* $\boldsymbol{y}$ *corresponding to the largest eigenvalue* λ *of* $\mathbf{S}$ *has strictly positive elements.*

We will make use of the following matrices. Let $\mathbf{V}$ denote the $h \times h$ ($h \leq m$) upper triangular matrix with unit elements on and above the diagonal and all other elements zero. Its inverse $\mathbf{V}^{-1}$ is the matrix with unit elements on the diagonal and with elements -1 adjacent and above the diagonal. Furthermore, let $\mathbf{I}$ be the identity matrix of size $(m-h) \times (m-h)$. Denote by $\mathbf{W}$ the diagonal block matrix of order m with diagonal elements $\mathbf{V}$ and $\mathbf{I}$. Examples of $\mathbf{V}$ and $\mathbf{V}^{-1}$ of sizes 3×3 are respectively

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Examples of $\mathbf{W}$ and $\mathbf{W}^{-1}$ of sizes 5×5 are

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Consider the coefficient matrices $\mathbf{S}_{\text{Dice2}}$ and $\mathbf{S}_{\text{RR}}$ with respective elements

$$S_{\text{Dice2}} = \frac{a_{jk}}{p_k} \quad \text{and} \quad S_{\text{RR}} = a_{jk}.$$

Let $\mathbf{y}$ be the eigenvector corresponding to the largest eigenvalue λ of either the matrix $\mathbf{S}_{\text{Dice2}}$ or $\mathbf{S}_{\text{RR}}$. In Proposition 8.1 it is shown that if the columns of the data matrix (or items in item response theory) can be ordered such that $p_j \geq p_{j+1}$ (6.3) and $a_{jk} \geq a_{j+1k}$ (6.4) hold, then this ordering is reflected in $\mathbf{y}$.

Proposition 8.1. *Suppose that h of the m column vectors of the data matrix $\mathbf{X}$, which without loss of generality can be taken as the first h, can be ordered such that* (6.3) *and* (6.4) *hold. Then the elements of $\mathbf{y}$ corresponding to these h items satisfy $y_1 > y_2 > ... > y_h > 0$.*

Proof: We first consider the proof for $\mathbf{S}_{\text{Dice2}}$. Since $\mathbf{W}$ is non-singular, $\mathbf{y}$ is an eigenvector of $\mathbf{S}_{\text{Dice2}}$ corresponding to λ if and only if $\mathbf{z} = \mathbf{W}^{-1}\mathbf{y}$ is an eigenvector of $\mathbf{T} = \mathbf{W}^{-1}\mathbf{S}_{\text{Dice2}}\mathbf{W}$ corresponding to λ. Under the conditions of the theorem, the elements of $\mathbf{T}$ turn out to be positive and the elements of $\mathbf{T}^2$ turn out to be strictly positive. This can be verified as follows.

The matrix $\mathbf{W}^{-1}\mathbf{S}_{\text{Dice2}} = \mathbf{U} = \{u_{jk}\}$ has elements

$$\begin{aligned} u_{jk} &= \frac{a_{jk} - a_{j+1k}}{p_k} && \text{for} \quad 1 \leq j < h \quad \text{and} \quad 1 \leq k \leq m \\ u_{jk} &= \frac{a_{jk}}{p_k} && \text{for} \quad h \leq j \leq m \quad \text{and} \quad 1 \leq k \leq m. \end{aligned}$$

Because $a_{jk} \geq a_{j+1k}$, $\mathbf{U}$ has positive elements except for u_{jj+1}, $j = 1, ..., h-1$. However, since $p_j \geq p_{j+1}$

$$\begin{aligned} u_{jj} + u_{jj+1} &= \frac{p_{j+1}a_{jj} - p_{j+1}a_{jj+1} + p_j a_{jj+1} - p_j a_{j+1j+1}}{p_j p_{j+1}} \\ &= \frac{a_{jj+1}(p_j - p_{j+1})}{p_j p_{j+1}} > 0 \end{aligned}$$

for $j = 1, ..., h-1$. Hence, the matrix $\mathbf{T} = \mathbf{UW}$ has positive elements. Moreover, because the elements in the last row and last column of $\mathbf{T}$ are strictly positive, it follows that the elements of $\mathbf{T}^2$ are strictly positive. Application of Theorem 8.1 yields that the eigenvector $\mathbf{z}$ of $\mathbf{T}$ (or $\mathbf{T}^2$) has strictly positive elements. The fact that $\mathbf{z} = \mathbf{W}^{-1}\mathbf{y}$ completes the proof for $\mathbf{S}_{\text{Dice2}}$.

Next we consider the proof for $\mathbf{S}_{\text{RR}}$, which is similar to the proof $\mathbf{S}_{\text{Dice2}}$. The matrix $\mathbf{W}^{-1}\mathbf{S}_{\text{RR}} = \mathbf{U} = \{u_{jk}\}$ has elements

$$\begin{aligned} u_{jk} &= a_{jk} - a_{j+1k} && \text{for} \quad 1 \leq j < h \quad \text{and} \quad 1 \leq k \leq m \\ u_{jk} &= a_{jk} && \text{for} \quad h \leq j \leq m \quad \text{and} \quad 1 \leq k \leq m. \end{aligned}$$

Because $a_{jk} \geq a_{j+1k}$, $\mathbf{U}$ has positive elements except for u_{jj+1} for $1 \leq j \leq h-1$. Since $p_j \geq p_{j+1}$

$$u_{jj} + u_{jj+1} = a_{jj} - a_{jj+1} + a_{jj+1} - a_{j+1j+1} > 0$$

for $1 \leq j \leq h-1$. This completes the proof for $\mathbf{S}_{\text{RR}}$. □

Consider the similarity matrices $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Cole1}}$ and $\mathbf{S}_{\text{Cole2}}$ with respective elements

$$S_{\text{Dice1}} = \frac{a_{jk}}{p_j}, \quad S_{\text{Cole1}} = \frac{a_{jk} - p_j p_k}{p_j q_k} \quad \text{and} \quad S_{\text{Cole2}} = \frac{a_{jk} - p_j p_k}{p_k q_j}.$$

Let $\mathbf{y}$ be the eigenvector corresponding to the largest eigenvalue λ of one of the three similarity matrices $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Cole1}}$ or $\mathbf{S}_{\text{Cole2}}$. Schriever (1986) showed that if the columns of the data matrix (or items in item response theory) can be ordered such that (6.3) and (6.6)

$$\frac{a_{jk}}{p_j} \leq \frac{a_{j+1k}}{p_{j+1}} \quad \text{for fixed } k\ (\neq j)$$

hold, then this ordering is reflected in $\mathbf{y}$ for $\mathbf{S}_{\text{Cole1}}$ or $\mathbf{S}_{\text{Cole2}}$. Proposition 8.2 is used to demonstrate that the same eigenvector property holds for $\mathbf{S}_{\text{Dice1}}$.

Proposition 8.2. *Suppose that h of the m column vectors of $\mathbf{X}$, which without loss of generality can be taken as the first h, can be ordered such that* (6.3) *and* (6.6) *hold. Then the elements of $\mathbf{y}$ corresponding to these h items satisfy $y_1 > y_2 > \ldots > y_h > 0$.*

Proof: The proof is similar to the proof for $\mathbf{S}_{\text{Dice2}}$ in Proposition 8.1. The matrix $(\mathbf{W}^{-1})^T\mathbf{S}_{\text{Dice1}} = \mathbf{U} = \{u_{jk}\}$ has elements

$$\begin{aligned} u_{jk} &= \frac{p_{j-1}a_{jk} - p_j a_{j-1k}}{p_{j-1}p_j} && \text{for} \quad 2 \leq j \leq h \quad \text{and} \quad 1 \leq k \leq m \\ u_{jk} &= \frac{a_{jk}}{p_j} && \text{for} \quad h < j \leq m \quad \text{and} \quad 1 \leq k \leq m. \end{aligned}$$

Because $p_{j-1}a_{jk} \geq p_j a_{j-1k}$, the matrix $\mathbf{U}$ has positive elements except for u_{jj-1} for $2 \leq j \leq h$. However, since $p_{j-1} \geq p_j$

$$\begin{aligned} u_{jj-1} + u_{jj} &= \frac{p_{j-1}a_{jj-1} - p_j a_{j-1j-1} + p_{j-1}a_{jj} - p_j a_{jj-1}}{p_{j-1}p_j} \\ &= \frac{a_{jj-1}(p_{j-1} - p_j)}{p_{j-1}p_j} > 0 \end{aligned}$$

for $2 \leq j \leq h$. This completes the proof. □

8.2 Related eigenvectors

In the previous section it was shown what ordinal information can be obtained from the eigenvector corresponding to the largest eigenvalue of coefficient matrices $\mathbf{S}_{\text{RR}}$, $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Dice2}}$, $\mathbf{S}_{\text{Cole1}}$ and $\mathbf{S}_{\text{Cole2}}$. In this section it is pointed out what eigendecompositions of various similarity matrices are related.

Let $\mathbf{y}_1^{(t)}$, $\mathbf{y}_0^{(t)}$ and $\mathbf{z}^{(t)}$ denote the eigenvectors of similarity matrices $\mathbf{S}_{\text{Cole1}}$, $\mathbf{S}_{\text{Cole2}}$ and $\mathbf{S}_{\text{Phi}}$ with respective elements

$$S_{\text{Cole1}} = \frac{a_{jk} - p_j p_k}{p_j q_k} \quad \text{and} \quad S_{\text{Cole2}} = \frac{a_{jk} - p_j p_k}{p_k q_j}$$

and

$$S_{\text{Phi}} = \frac{a_{jk} - p_j p_k}{\sqrt{p_j p_k q_j q_k}}.$$

The eigendecomposition of $\mathbf{S}_{\text{Phi}}$ defines principal component analysis for binary data, whereas the decomposition of $\mathbf{S}_{\text{Cole1}}$ and $\mathbf{S}_{\text{Cole2}}$ give the category weights from a homogeneity analysis when applied to binary data (Yamada and Nishisato, 1993; Schriever, 1986; or see Section 8.3). With ordinary principal component analysis there is a single weight $z_j^{(t)}$ for each item j on dimension t. In contrast, in Guttman's categorical principal component analysis there are two weights for each item j on dimension t, one for each response (0 and 1). Let $y_{j0}^{(t)}$ and $y_{j1}^{(t)}$ denotes these weights. The relationships between the eigenvectors of $\mathbf{S}_{\text{Cole1}}$, $\mathbf{S}_{\text{Cole2}}$ and $\mathbf{S}_{\text{Phi}}$ can already be found in Yamada and Nishisato (1993).

Theorem 8.2 [Yamada and Nishisato, 1993]. *The eigenvectors of similarity matrices* $\mathbf{S}_{\text{Cole1}}$, $\mathbf{S}_{\text{Cole2}}$ *and* $\mathbf{S}_{\text{Phi}}$ *are related by*

$$y_{j1}^{(t)} = \sqrt{\frac{q_j}{p_j}}\, z_j^{(t)} \quad \textit{and} \quad y_{j0}^{(t)} = \sqrt{\frac{p_j}{q_j}}\, z_j^{(t)}.$$

Proof: The eigenvectors are related due to the following property. If $\mathbf{T}$ is a nonsingular matrix, then $\mathbf{y}^{(t)}$ is an eigenvector of $\mathbf{S}$ corresponding to the tth eigenvalue λ_t if and only if $\mathbf{z}^{(t)} = \mathbf{T}^{-1}\mathbf{y}^{(t)}$ is an eigenvector of $\mathbf{T}^{-1}\mathbf{S}\mathbf{T}$ corresponding to λ_t. We have

$$S_{\text{Cole1}} = \sqrt{\frac{p_k}{q_k}}\, \frac{a_{jk} - p_j p_k}{\sqrt{p_j p_k q_j q_k}}\, \sqrt{\frac{q_j}{p_j}} = \frac{a_{jk} - p_j p_k}{p_j q_k}. \qquad \square$$

Thus, if we would calculate the matrices $\mathbf{S}_{\text{Cole1}}$, $\mathbf{S}_{\text{Cole2}}$ and $\mathbf{S}_{\text{Phi}}$, these matrices have the same eigenvalues and the various eigenvectors are related by the relations in Theorem 8.2. Note that $\mathbf{S}_{\text{Cole1}}$ and $\mathbf{S}_{\text{Cole2}}$ possess the interesting eigenvector property described in Proposition 8.2, whereas $\mathbf{S}_{\text{Phi}}$ does not.

A similar relation exists between the eigenvectors of the matrices $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Dice2}}$ and $\mathbf{S}_{\text{DK}}$ with respective elements

$$S_{\text{Dice1}} = \frac{a_{jk}}{p_j}, \quad S_{\text{Dice2}} = \frac{a_{jk}}{p_k} \quad \text{and} \quad S_{\text{DK}} = \frac{a_{jk}}{\sqrt{p_j p_k}}.$$

Let $\mathbf{y}_1^{(t)}$, $\mathbf{y}_0^{(t)}$ and $\mathbf{z}^{(t)}$ denote the eigenvectors of similarity matrices $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Dice2}}$ and $\mathbf{S}_{\text{DK}}$. Proposition 8.3 considers the relationships between the eigenvectors of $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Dice2}}$ and $\mathbf{S}_{\text{DK}}$.

Proposition 8.3. *The eigenvectors of similarity matrices* $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Dice2}}$ *and* $\mathbf{S}_{\text{DK}}$ *are related by*

$$y_{j1}^{(t)} = \frac{1}{\sqrt{p_j}}\, z_j^{(t)} \quad \textit{and} \quad y_{j2}^{(t)} = \frac{\sqrt{p_j}}{1}\, z_j^{(t)}.$$

Proof: The proof is similar to the proof of Theorem 8.2. We have

$$S_{\text{Dice1}} = \frac{\sqrt{p_k}}{1}\, \frac{a_{jk}}{\sqrt{p_j p_k}}\, \frac{1}{\sqrt{p_j}} = \frac{a_{jk}}{p_j} \quad \text{and} \quad S_{\text{Dice2}} = \frac{1}{\sqrt{p_j}}\, \frac{a_{jk}}{\sqrt{p_j p_k}}\, \frac{\sqrt{p_j}}{1} = \frac{a_{jk}}{p_k}.$$

□

Again, if we would calculate the eigendecompositions of the matrices $\mathbf{S}_{\text{Dice1}}$, $\mathbf{S}_{\text{Dice2}}$ and $\mathbf{S}_{\text{DK}}$, we would obtain the same eigenvalues for each matrix. The various eigenvectors are related by the relations in Proposition 8.3. Note that $\mathbf{S}_{\text{Dice1}}$ and $\mathbf{S}_{\text{Dice2}}$ possess the eigenvector properties presented in Propositions 8.1 and 8.2.

8.3 Homogeneity analysis

Homogeneity analysis is the generalization of principal component analysis to categorical data proposed by Guttman (1941). In the previous section it was noted that the optimal category weights from a homogeneity analysis are the eigenvectors of the matrices $\mathbf{S}_{\text{Cole1}}$ and $\mathbf{S}_{\text{Cole2}}$ if the data are binary. In this section we consider several other matrices from the homogeneity analysis methodology and present the corresponding formulas for the case that homogeneity analysis is applied to binary data.

Suppose the multivariate data are in a $n \times m$ matrix $\mathbf{X}$ containing the responses of n persons on m categorical items. Let $\mathbf{G}_j$ be an indicator matrix of item j, defined as the order $n \times L_j$ matrix $\mathbf{G}_j = \{g_{il(j)}\}$, where $g_{il(j)}$ is a (0,1) variable. Each column of $\mathbf{G}_j$ refers to the L_j possible responses of item j. If person i responded category l on item j, then $g_{il(j)} = 1$, that is, the cell in the ith row and lth column of $\mathbf{G}_j$ contains a 1, and $g_{il(j)} = 0$ otherwise. The partitioned indicator matrix $\mathbf{G}$ then consists of all $\mathbf{G}_j$ positioned next to each other.

Let $\mathbf{D}$ of size $\sum_j L_j \times \sum_j L_j$ be the diagonal matrix with the diagonal elements of $\mathbf{G}^T\mathbf{G}$ on its main diagonal and 0s elsewhere. The matrix $\mathbf{D}$ reflects the total amount of 1s there are in each column of $\mathbf{G}$. Suppose the category weights of homogeneity analysis are in the vector $\mathbf{y}$ of size $\sum_j L_j \times 1$. The category weights can be obtained from the generalized eigenvalue problem $\mathbf{G}^T\mathbf{G}\mathbf{y} = m\lambda\mathbf{D}\mathbf{y}$. By itself the generalized eigenvalue problem does not tell us which eigenvector to take. The category weights $\mathbf{y}$ are the eigenvectors of the matrix $\mathbf{F} = m^{-1}\mathbf{D}^{-1}\mathbf{G}^T\mathbf{G}$. The eigenvector $\mathbf{y}$ corresponding to the largest eigenvalue λ of $\mathbf{F}$ is considered trivial because it does not correspond to a variance ratio. There are various ways to remove the trivial solution: one way is by setting the matrix $\mathbf{G}$ in deviations from its column means (Gifi, 1990, Section 3.8.2).

It turns out that the matrix $\mathbf{F}$ of size $\sum_j L_j \times \sum_j L_j$ has explicit elements. Note that, for ease of notation, the columns of $\mathbf{G}$ are indexed by j and k in the following.

Proposition 8.4. *The matrix $\mathbf{F} = m^{-1}\mathbf{D}^{-1}\mathbf{G}^T\mathbf{G}$ with $\mathbf{G}$ in deviations from its column means, has elements*

$$\begin{aligned} f_{jk} &= \frac{a_{jk} - p_j p_k}{p_j} && \text{for } j \text{ and } k \text{ from different columns of } \mathbf{X} \\ f_{jk} &= -p_k && \text{for } j \text{ and } k \text{ from the same column of } \mathbf{X} \\ f_{jj} &= 1 - p_j. \end{aligned}$$

Proof: The matrix $\mathbf{G}^T\mathbf{G}$ with $\mathbf{G}$ in deviations from its column means is a covariance matrix corresponding to the columns of binary matrix $\mathbf{G}$, which has elements $a_{jk} - p_j p_k$. Furthermore, the elements of $m^{-1}\mathbf{D}$ equal the p_j. □

The elements of the linear operator $\mathbf{F}$ have even more explicit elements if the data matrix consists of binary scores, that is, when each item has two response categories. The data matrix $\mathbf{X}$ has m columns, whereas the corresponding indicator coding $\mathbf{G}$ then has $2m$ columns. Linear operator $\mathbf{F}$ is then a matrix of size $2m \times 2m$.

Corollary 8.1. *Suppose the data matrix consists of binary items. Then $\mathbf{F}$ has elements*

$$\begin{aligned} f_{jk} &= \frac{a_{jk} - p_j p_k}{p_j} && \text{for } j \text{ and } k \text{ from different items} \\ f_{jk} &= -p_k && \text{for } j \text{ and } k \text{ from the same item} \\ f_{jj} &= q_j. \end{aligned}$$

Proposition 8.5. *Suppose the data matrix consists of binary items. The rows and columns of* $\mathbf{F}$ *can be reordered such that* $\mathbf{F}$ *has block structure*

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 & -\mathbf{F}_1 \\ -\mathbf{F}_2 & \mathbf{F}_2 \end{bmatrix}$$

where $\mathbf{F}_1$ *and* $\mathbf{F}_2$ *are of size* $m \times m$.
Proof: Consider Corollary 8.1. If the column of $\mathbf{G}$ corresponding to category 1 of item l has positive or negative covariance with the jth column of $\mathbf{G}$, then the column of $\mathbf{G}$ corresponding to category 0 of item l has the same covariance with the kth column of $\mathbf{G}$ but with opposite sign. In the case that two columns have zero covariance, the sign may arbitrarily be chosen. Providing that all $2m$ diagonal elements of $\mathbf{D}$ are different, it holds that $\mathbf{F}_1 \neq \mathbf{F}_2$. □

From Proposition 8.4 and 8.5 it follows that $\mathbf{F}$ has explicit elements and, moreover, can be reordered to exhibit simple (block) structure. Proposition 8.5 may be used to derive to the following eigenvector property for the category weights concerning sign. For the next result, let $\mathbf{y}$ be the eigenvector corresponding to the largest eigenvalue of $\mathbf{F}$ of size $2m \times 2m$.

Proposition 8.6. *Suppose the data matrix consists of binary items. The elements in* $\mathbf{y}$ *corresponding to columns of* $\mathbf{G}$ *that have positive covariance, have similar sign.*
Proof: Consider Proposition 8.5. Furthermore, let $\mathbf{I}$ be the identity matrix of size $m \times m$, and let $\mathbf{W}$ be the diagonal block matrix of size $2m \times 2m$ with diagonal elements $\mathbf{I}$ and $-\mathbf{I}$. Since $\mathbf{W}$ is non-singular, it follows that the matrix $\mathbf{U} = \mathbf{W}^{-1}\mathbf{F}\mathbf{W}$ has positive elements. Application of Theorem 8.1 yields that the eigenvector $\mathbf{z}$ corresponding to the largest eigenvalue $\mathbf{U}$ has positive elements. The assertion then follows from $\mathbf{y} = \mathbf{W}^{-1}\mathbf{z}$. □

The linear operator $\mathbf{F}$ considered in Propositions 8.4 to 8.6 is of the similarity type. Heiser (1981) and Meulman (1982) consider the multidimensional scaling approach to homogeneity analysis, which is based on Benzécri or chi-square distances. Meulman (1982) shows how category and persons weights can be obtained from distance matrices using classical scaling (Torgerson, 1958; Gower, 1966).

Let g_{ik} denote the response of person i to the kth column of $\mathbf{G}$ and let d_k denote the number of 1s in the kth column of $\mathbf{G}$. Meulman (1982, p. 48) defines the squared Benzécri distance between person i and l as

$$B_{il}^2 = \frac{1}{m^2} \sum_k \frac{(g_{ik} - g_{lk})^2}{d_k}.$$

If person i and l gave the same response to an item, then this does not contribute to the distance B_{il}^2. If the $n \times m$ data matrix $\mathbf{X}$ consist of m binary items ($1 \leq j \leq m$) then B_{il}^2 can be written as

$$B_{il}^2 = \frac{1}{m^2} \sum_{k=1}^{2m} \frac{(g_{ik} - g_{lk})^2}{d_k} = \frac{1}{m^2} \sum_{j=1}^{m} \frac{(x_{ij} - x_{lj})^2}{d_j} + \frac{1}{m^2} \sum_{j=1}^{m} \frac{(x_{ij} - x_{lj})^2}{n - d_j}$$

where d_j ($n - d_j$) is the number of 1s (0s) in the jth column of $\mathbf{X}$. Suppose that for h items ($1 \leq h \leq m$) person i and l have different responses. Then $m^2 B_{il}^2$ can be written as

$$m^2 B_{il}^2 = \frac{1}{d_1} + \frac{1}{d_2} + ... + \frac{1}{d_h} + \frac{1}{n - d_1} + \frac{1}{n - d_2} + ... + \frac{1}{n - d_h}$$

or B_{il}^2 as

$$B_{il}^2 = \frac{n}{m^2} \sum_{j=1}^{h} \frac{1}{d_j(n - d_j)}.$$

Squared distance B_{il}^2 may be interpreted as a weighted symmetric set difference. Meulman (1982, p. 37) defines the squared Benzécri distance between category j and k as

$$B_{jk}^2 = \sum_{i=1}^{n} \left[\frac{g_{ij}}{d_j} - \frac{g_{ik}}{d_k} \right]^2.$$

In general, not just with binary data, four types of persons can be distinguished. We define the three quantities

$$\begin{aligned} a &= \text{number of times } g_{ij} = 1 \text{ and } g_{ik} = 1; \\ b &= \text{number of times } g_{ij} = 1 \text{ and } g_{ik} = 0; \\ c &= \text{number of times } g_{ij} = 0 \text{ and } g_{ik} = 1. \end{aligned}$$

Note that $d_j = a + b$ and $d_k = a + c$. The Benzécri distance B_{jk}^2 then equals

$$B_{jk}^2 = a \left[\frac{1}{d_j} - \frac{1}{d_k} \right]^2 + b \left[\frac{1}{d_j} \right]^2 + c \left[\frac{1}{d_k} \right]^2 = \frac{1}{d_j} + \frac{1}{d_k} - \frac{2a}{d_j d_k} = \frac{d_j + d_k - 2a}{d_j d_k}.$$

When category j and k are two categories of the same item, $a = 0$ and therefore $B_{jk}^2 = d_j^{-1} + d_k^{-1}$.

8.4 Epilogue

For several coefficient matrices we studied in this chapter the eigenvector elements corresponding to the largest eigenvalue. It was shown that ordinal information on model probabilities is reflected in the eigenvector elements. It is thus possible to uncover correct orderings of various latent variable models presented in Chapter 6 using eigenvectors of coefficient matrices. For coefficients

$$S_{\text{Dice2}} = \frac{a_{jk}}{p_k} \quad \text{and} \quad S_{\text{RR}} = a_{jk}$$

it was demonstrated by Proposition 8.1 that if a set of items can be ordered such that double monotonicity model holds, then this ordering is reflected in the elements of the eigenvector corresponding to the largest eigenvalue of the similarity matrices. The conventional method of discovering this order is by inspecting the proportion item correct (p_j). A similar, although less general, eigenvector property holds for coefficients

$$S_{\text{Cole1}} = \frac{a_{jk} - p_j p_k}{p_j p_k}, \quad S_{\text{Cole2}} = \frac{a_{jk} - p_j p_k}{p_k q_j} \quad \text{and} \quad S_{\text{Dice1}} = \frac{a_{jk}}{p_j}.$$

In Proposition 8.2 it was shown that if a set of items can be ordered such that the double monotonicity model holds and, moreover, the response functions satisfy total positivity of order 2, then this ordering is reflected in the elements of the eigenvector corresponding to the largest eigenvalue of the coefficient matrices.

In addition to the eigenvector properties of several asymmetric matrices, various matrix methodology of homogeneity analysis was studied. Homogeneity analysis is a versatile technique and it can be studied from various points of view. It was shown that several of the different matrices corresponding to this form of categorical principal component analysis have often explicit elements. If the data matrix contains binary data, then the category weights corresponding to categories with positive covariance have the same sign.

Heiser (1981) and Meulman (1982) consider the multidimensional scaling approach to homogeneity analysis, which is based on dissimilarities or distances. The distances called Benzécri distances in Meulman (1982) are nowadays referred to as chi-square distances. The chi-square distance between two persons is a form of the extended matching coefficient weighted inversely by the response frequencies.

CHAPTER 9

Homogeneity analysis and the 2-parameter IRT model

Guttman (1941) presented a method that can be used to obtain a representation of the structure of multivariate categorical data. The technique was briefly mentioned in Sections 8.2 and 8.3. The method gives a multidimensional decomposition of the data with the most informative structural dimension extracted first, then the second most informative dimension, and so on, until the information in the data is exhaustively extracted. The method is typically used for the construction of geometrical representations of the dependencies in the data in low-dimensional Euclidean space, often two-dimensional, from the extracted dimensions. Given that the data are in a person by item table, each dimension consists of weights for the item categories (known as optimal weights) and scores for the persons. The discovery or rediscovery of Guttman's method by many authors has led to the fact that the method is known under many different names, for example, dual scaling (Nishisato, 1980), multiple correspondence analysis (Greenacre, 1984), Fisher's method of optimal scores (Gower, 1990), or homogeneity analysis (Gifi, 1990).

[0]Parts of this chapter appeared in Warrens, M.J., De Gruijter, D.N.M. and Heiser, W.J. (2007), A systematic comparison between classical optimal scaling and the two-parameter IRT model, *Applied Psychological Measurement, 31 (2),* 106–120.

Warrens, Heiser and De Gruijter (2006), Warrens and Heiser (2006) and Warrens, De Gruijter and Heiser (2007) showed that homogeneity analysis is useful for analyzing binary data. Gifi (1990, p. 425-440) and Cheung and Mooi (1994) showed that homogeneity analysis is useful for analyzing Likert data. In addition, the latter authors compared the homogeneity scaling findings to an item response theory analysis using the rating scale model (Andrich, 1988). They evaluated both the similarities and differences and concluded that there is great similarity between the two contrasting approaches. A systematic comparison of homogeneity analysis and the item response theory approach is lacking however. The present chapter is therefore used to systematically explore the relationship between a one-dimensional homogeneity analysis and the logistic 2-parameter model.

9.1 Classical item analysis

Let ω denote a latent variable and let δ_j and β_j be respectively a discrimination and location parameter of the logistic 2-parameter model (Section 6.1). The probability of a response 1 on item j under the logistic 2-parameter model is given by

$$p_j(\omega) = \frac{\exp[\delta_j(\omega - \beta_j)]}{1 + \exp[\delta_j(\omega - \beta_j)]}. \tag{9.1}$$

On pages 377 and 378 of their by now classic book, Lord and Novick (1968) show how the item parameters of the normal ogive 2-parameter model are related to the indices used in classical item analysis. Two conditions are assumed:

1) the latent variable is normally distributed with zero mean and unit variance;
2) the appropriate model is the 2-parameter normal ogive.

Under these conditions the mean of ω, conditional on a score 1 on item j, equals

$$\mu_{j1} = \frac{\phi(\gamma_j)\,\rho'_j}{p_j}$$

where $p_j = \Phi(-\gamma_j)$ is the item proportion correct, where Φ denotes the cumulative normal distribution function and $\gamma_j = \beta_j\,\rho'_j$. Furthermore, $\phi(\gamma_j)$ is the ordinate of the standard normal distribution, and

$$\rho'_j = \frac{\delta_j}{\sqrt{1 + \delta_j^2}}$$

is the biserial correlation between item j and the latent variable.

Due to the fact that the logistic formulation of the 2-parameter model is more tractable then the normal ogive, the former is sometimes preferred in item response theory work. Let us derive how the above relations on the basis of the normal ogive hold under the logistic approximation. The logistic 2-parameter model and its approximate relation with the normal ogive 2-parameter model are given by

$$p_j(\omega) = \Psi\left[\delta_j(\omega - \beta_j)\right] \approx \Phi\left[D^{-1}\delta_j(\omega - \beta_j)\right]$$

where Ψ denotes the logistic function, and $D = 1.7$ is a constant. Under the logistic approximation the mean of ω, conditional on a response 1 on item j, equals

$$\mu_{j1} \approx \frac{\phi(\gamma_j^*)\,\rho_j^*}{\Psi(-D\gamma_j^*)} \tag{9.2}$$

where

$$\Psi(-D\gamma_j^*) \approx p_j \tag{9.3}$$

$\gamma_j^* = \beta_j\,\rho_j^*$, and

$$\rho_j^* = \frac{\delta_j}{D\sqrt{1 + D^{-2}\delta_j^2}}.$$

Furthermore, under the logistic approximation

$$\phi(\gamma_j^*) \approx D\Psi(D\gamma_j^*)\left[1 - \Psi(D\gamma_j^*)\right] \;=\; D\Psi(-D\gamma_j^*)\left[1 - \Psi(-D\gamma_j^*)\right]$$

and (9.2) can be rewritten as

$$\mu_{j1} \approx (1 - p_j)D\rho_j^*.$$

9.2 Person parameter

With binary responses, the 2-parameter item response model uses two item parameters whereas a one-dimensional homogeneity analysis produces two category weights. Furthermore, both approaches use one parameter for locating persons. Let us show how the item response theory person parameter estimate, denoted by ω_i, and the optimal person score, denoted by x_i, are related. This relationship is used in the remaining sections of this chapter, where it is assumed that the optimal person score is a reasonable approximation of the latent variable, that is $x_i \approx \omega_i$. In the following we will show that this approximation is a reasonable one.

Two data sets were generated from both the logistic 2-parameter model and the Rasch model under the following conditions. The data sets consisted of the responses of 1000 persons on 50 items; for each data set the location parameters β_j's were sampled from a standard normal distribution; the discrimination parameters for the 2-parameter model were sampled from a uniform distribution on the range [1,2], for the Rasch (1960) model these were set to unity; the latent variable was sampled from a standard normal distribution.

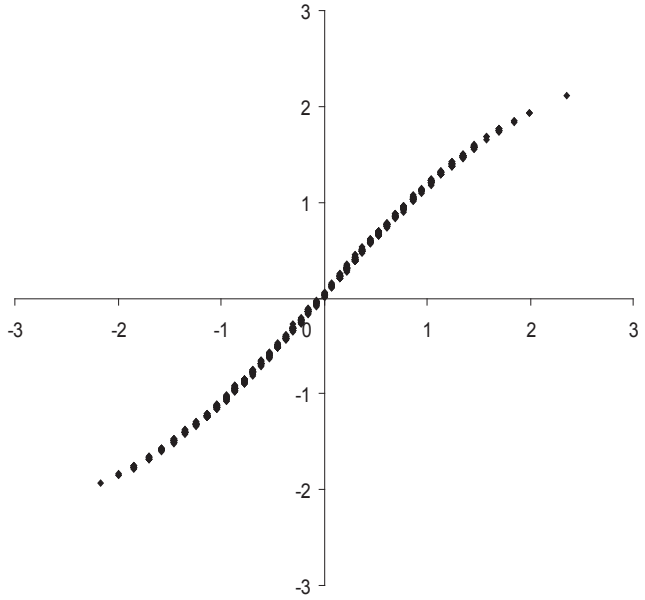

Figure 9.1: *Plot of maximum a posteriori person estimates (horizontal) versus homogeneity person scores (vertical) for the Rasch data set.*

For both data sets the optimal scaling and item response theory person estimates were obtained. The item response theory analysis was performed using the Multilog software program (Thissen, Chen and Bock, 2003) to obtain maximum a posteriori estimates. The person estimates of both approaches are plotted in Figures 9.1 and 9.2 for respectively the Rasch model and the logistic 2-parameter model. The correlations between the two sets of estimates are in both figures $> .99$. The root mean squared errors are $< .2$, which concurs with the slight nonlinearity that can be observed upon close inspection. Apart from the nonlinearity, the optimal person score seems a reasonable approximation of the latent variable, that is, $\omega_i \approx x_i$ under the 2-parameter model.

9.3 Discrimination parameter

Lord (1958) showed that the optimal category weights on the first dimension maximize coefficient alpha (Cronbach, 1951), an important lower bound to reliability, a concept used in classical test theory (De Gruijter and Van der Kamp, 2008). An application of Guttman's method in which this property is explicitly used, can be found in Serlin and Kaiser (1978). The second, third and subsequent dimensions of the technique may be considered sets of weights corresponding to local maximums of alpha. If the data are binary, there are only two category weights for each item j. For this special case it is possible to construct a single index for each item that reflects all information for maximizing coefficient alpha. This can be done by translating the two optimal homogeneity weights $y_{j0}^{(t)}$ and $y_{j1}^{(t)}$ into new weights $v_{j0}^{(t)}$ and

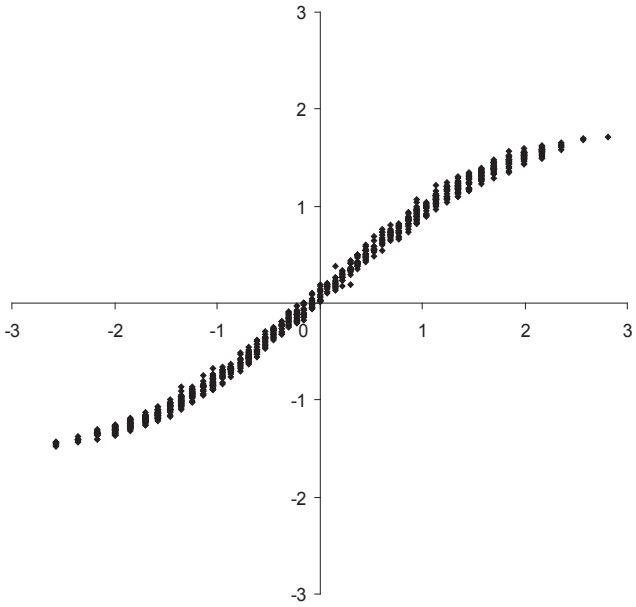

Figure 9.2: *Plot of maximum a posteriori person estimates (horizontal) versus homogeneity person scores (vertical) for the logistic 2-parameter model data set.*

$v_{j1}^{(t)}$ (where t denotes the dimension). With the translations

$$\begin{aligned} v_{j0}^{(t)} &= y_{j0}^{(t)} - y_{j0}^{(t)} = 0 \\ \text{and} \quad v_{j1}^{(t)} &= y_{j1}^{(t)} - y_{j0}^{(t)} \end{aligned}$$

the category weight $y_{j0}^{(t)}$ is set to zero and all information of item j on maximizing coefficient alpha is reflected in $v_{j1}^{(t)}$. The latter weight is therefore denoted by $\max(\alpha)_j^{(t)} = v_{j1}^{(t)}$ in the following.

Let $\mathbf{z}_j^{(t)}$ be the eigenvector corresponding to the tth eigenvalue of the matrix $\mathbf{S}_{\text{Phi}}$ with elements

$$S_{\text{Phi}} = \frac{a_{jk} - p_j p_k}{\sqrt{p_j p_k q_j q_k}}.$$

Proposition 9.1. *The weight* $\max(\alpha)_j^{(t)}$ *is related to the principal component weight* $z_j^{(t)}$ *by*

$$\max(\alpha)_j^{(t)} = z_j^{(t)} \frac{1}{[p_j(1-p_j)]^{1/2}}.$$

Proof: The relationship follows from using the equations in Theorem 8.2 in

$$\max(\alpha)_j^{(t)} = y_{j1}^{(t)} - y_{j0}^{(t)}. \quad \square$$

Proposition 9.2. *The weights* $max(\alpha)_j^{(t)}$ *are elements of the eigenvector corresponding to the* t*th eigenvalue of the matrix* $\mathbf{S}_{\text{MA}}$ *with elements*

$$S_{\text{MA}} = \frac{a_{jk} - p_j p_k}{p_j(1-p_j)}.$$

Proof: The proof is similar to the proof of Theorem 8.2 and Proposition 8.3. Using the formulas in Proposition 8.1, we have

$$\begin{aligned} S_{\text{MA}} &= \left[\frac{p_k(1-p_k)}{1}\right]^{1/2} \frac{a_{jk} - p_j p_k}{[p_j(1-p_j)p_k(1-p_k)]^{1/2}} \left[\frac{1}{p_j(1-p_j)}\right]^{1/2} \\ &= \frac{a_{jk} - p_j p_k}{p_j(1-p_j)}. \quad \square \end{aligned}$$

From this point on, let $\max(\alpha)_j$ be short for $\max(\alpha)_j^{(1)} = y_{j1}^{(1)} - y_{j0}^{(1)}$, and let y_{j1} and y_{j0} be short for $y_{j1}^{(1)}$ and $y_{j0}^{(1)}$. The definition of $\max(\alpha)_j$ reveals that the item weight becomes greater as the mean values of all persons who responded 1 to item j and those who responded 0 become further apart. Hence, $\max(\alpha)_j$ has a clear interpretation as an index of discrimination.

An often used normalization in homogeneity analysis when applied to binary data, is $p_j y_{j1} + (1-p_j)y_{j0} = 0$, which can be written as

$$y_{j0} = -\frac{p_j y_{j1}}{1-p_j}. \tag{9.4}$$

With the help of (9.4), $\max(\alpha)_j$ can be written as

$$\max(\alpha)_j = \frac{y_{j1}}{1-p_j}. \tag{9.5}$$

In the following it is assumed that $x_i \approx \omega_i$ (Section 9.2). In addition it is assumed that

1) the latent variable is normally distributed with zero mean and unit variance;
2) the appropriate model is the 2-parameter model.

Under these assumptions the work of Lord and Novick (1968) on the relationship between the item response theory item parameters and some indices from classical item analysis becomes available. Under the above three assumptions it follows from Section 9.1 that

$$\max(\alpha)_j \approx D\rho_j^* = \frac{\delta_j}{\sqrt{1 + D^{-2}\delta_j^2}}. \tag{9.6}$$

The functional relationship in (9.6) was derived in a different way by De Gruijter (1984). Since, ρ_j^* has a maximum of unity, the quantity in (9.6) has a maximum value of $D = 1.7$. Since, the $\max(\alpha)_j$ weight is a function of δ_j only, δ_j can be expressed as a function of $\max(\alpha)_j$. The resulting function gives an estimate of the discrimination parameter of the logistic 2-parameter model given by

$$\hat{\delta}_j = \frac{D \max(\alpha)_j}{\sqrt{D^2 - [\max(\alpha)_j]^2}} \qquad \text{for } |\max(\alpha)_j| \leq D \tag{9.7}$$

which is a function of $\max(\alpha)_j$ only.

9.4 More discrimination parameters

A third measure of discrimination for item j, next to δ_j and $\max(\alpha)_j^{(t)}$, is described in Gifi (1990, Section 3.8.4). With binary data the measure is given by

$$\left[\eta_j^{(t)}\right]^2 = p_j \left[y_{j1}^{(t)}\right]^2 + (1 - p_j) \left[y_{j0}^{(t)}\right]^2. \tag{9.8}$$

Theorem 9.1 [Yamada and Nishisato, 1993, p. 60]. *The weight* $max(\alpha)_j^{(t)}$ *is related to* $\left[\eta_j^{(t)}\right]^2$ *by*

$$\max(\alpha)_j^{(t)} = \frac{\eta_j^{(t)}}{[p_j(1 - p_j)]^{1/2}}.$$

Proof: Equation (9.8) can be re-expressed in terms of $y_{j1}^{(t)}$ and $y_{j0}^{(t)}$ with the help of (9.4), which gives

$$y_{j1}^{(t)} = \eta_j^{(t)} \left[\frac{1 - p_j}{p_j}\right]^{1/2}$$
$$-y_{j0}^{(t)} = \eta_j^{(t)} \left[\frac{p_j}{1 - p_j}\right]^{1/2}.$$

Hence, we obtain

$$\max(\alpha)_j^{(t)} = y_{j1}^{(t)} - y_{j0}^{(t)} = \frac{\eta_j^{(t)}}{[p_j(1-p_j)]^{1/2}}$$

or

$$\left[\eta_j^{(t)}\right]^2 = p_j(1-p_j)\left[\max(\alpha)_j^{(t)}\right]^2.$$

In words, $\left[\eta_j^{(t)}\right]^2$ is the squared $\max(\alpha)_j^{(t)}$ of item j on dimension t, times the variance of item j. □

A fourth measure of discrimination is described in McDonald (1983). In a more general context than the one considered in the present chapter, McDonald argued not to interpret the category weights themselves, but the regression weights of each category on the person score x_i. With McDonald's formulation there is not one discrimination measure for each item j on dimension t, but one for each category. When each item has two categories, the measures are given by $\text{reg}_{j1}^{(t)} = p_j y_{j1}^{(t)}$ and $\text{reg}_{j0}^{(t)} = 1 - p_j y_{j0}^{(t)}$. Equation (9.4) can be written as

$$p_j y_{j1}^{(t)} = (p_j - 1) y_{j0}^{(t)} \quad \Leftrightarrow \quad \text{reg}_{j1}^{(t)} = -\text{reg}_{j0}^{(t)}.$$

Since, with binary data, the two regression weights contain the same information, it suffices to look at $\text{reg}_{j1}^{(t)}$, assumed to be positive, only.

Proposition 9.3. *The weight* $\max(\alpha)_j^{(t)}$ *is related to* $\text{reg}_{j1}^{(t)}$ *by*

$$\text{reg}_{j1}^{(t)} = p_j(1-p_j)\max(\alpha)_j^{(t)}.$$

Proof: Equation (9.5) can be written as

$$y_{j1}^{(t)} = (1-p_j)\max(\alpha)_j^{(t)}. \tag{9.9}$$

Multiplication of both sides of (9.9) by p_j gives the desired result. □

9.5 Location parameter and category weights

Now that the functional relationship between the discrimination indices has been established we turn our intention to the remaining information in the weights y_{j1} and y_{j0} (short for $y_{j1}^{(1)}$ and $y_{j0}^{(1)}$). Since $\max(\alpha)_j$ is given by the difference between y_{j1} and y_{j0}, the remaining information in the weights can be summarized in

$$\text{sum}_j = y_{j1} + y_{j0}.$$

With the help of (9.4), sum_j can be written as

$$\text{sum}_j = \frac{1 - 2p_j}{1 - p_j} y_{j1}.$$

Under the same three assumptions as used in Section 9.3, it follows that

$$\text{sum}_j \approx \; D\rho_j^* \left(1 - 2\Psi[-\beta_j \, D\rho_j^*]\right). \tag{9.10}$$

Suppose now that ρ_j^* in (9.10) is constant for all j. For this limited case it holds that if β_j increases, then sum_j also increases. Since, β_j and sum_j are monotonically related under this restriction, sum_j can be interpreted as a location parameter for a model of which the discrimination parameters are equal for all j, that is, the Rasch (1960) model.

From (9.10) an estimate for the location parameter β_j of the logistic 2-PM can be obtained. This estimate can be simplified. In addition to $\max(\alpha)_j$ only p_j is needed. Let Ψ denote the logistic function. Then, from (9.3) it follows that

$$p_j \approx \Psi[-\beta_j \max(\alpha)_j]. \tag{9.11}$$

If one takes the inverse of the logistic function on both sides of (9.11) and rewrites the resulting equation in terms of β_j, one obtains an estimate of location for item j given by

$$\hat{\beta}_j = -\frac{\ln\left(\frac{p_j}{1-p_j}\right)}{\max(\alpha)_j}. \tag{9.12}$$

The estimate derived in (9.12) is related to the estimate proposed by Cohen (1979) for the Rasch (1960) model.

9.6 Epilogue

Homogeneity analysis or multiple correspondence analysis is a method that can be used to obtain a representation of the structure of multivariate categorical data. If the data are binary, there are only two category weights for each item j of a homogeneity analysis, namely, y_{j1} and y_{j0}. Category weights y_{j1} (y_{j0}) are the elements of the eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{S}_{\text{Cole1}}$ ($\mathbf{S}_{\text{Cole2}}$) with elements

$$S_{\text{Cole1}} = \frac{a_{jk} - p_j p_k}{p_j q_k} \qquad \left(S_{\text{Cole2}} = \frac{a_{jk} - p_j p_k}{p_k q_j} \right).$$

In this chapter the relationship between a one-dimensional homogeneity analysis and the logistic 2-parameter model was systematically explored. It was first studied how the item response theory person parameter estimate and the optimal person score are related. It was shown that that the optimal person score is a reasonable approximation of the latent variable. Next, the homogeneity category weights of the first dimension were related to the parameters of the 2-parameter model, using some results on the relationship between item response theory and classical item analysis from Lord and Novick (1968, p. 377-378).

At this point the question arises, what is the point of knowing the functional relationship between a one-dimensional homogeneity analysis and item response theory? First of all, it is useful in general to study equivalences or functional relationships between different methods of data analysis, primarily because this often gives new insight into the methods themselves. More precisely, approximate estimates for the item parameters of the logistic 2-parameters were derived which are based on the conditional means. The estimates were not meant as possible replacement of the current item response theory estimates. One might be tempted to ask if these estimates may be used to obtain perhaps less biased parameter estimates (maximum likelihood estimation is already most efficient). In non-reported simulation experiments it turns out that the estimates based on homogeneity analysis do not give less biased estimates nor smaller standard errors. On the other hand, the closeness of the optimal person score to the latent variable under a variety of item response theory models shows that homogeneity analysis is a useful multi-purpose data analysis method. Even without specifying a model one cannot be far off.

The findings in this chapter do give several new insights into the application of homogeneity analysis. A typical use of homogeneity analysis and other optimal scaling methods, is the construction of geometrical representations of the dependencies in the data in low-dimensional Euclidean space, often two-dimensional, from the extracted dimensions. The use of two-dimensional (sometimes three-dimensional) plots is embedded so strongly in the optimal scaling community that it is often regarded as impossible that all relevant information is in the first dimension only.

CHAPTER 10

Metric properties of two-way coefficients

Various methods of data analysis use the facility of fitting distances to a table of coefficients, where the coefficients are summary measures of the data. An example is metric multidimensional scaling, and a popular distance measure is the Euclidean distance. In this chapter a review is presented on metric properties of various coefficients for binary data. Metric properties of various similarity coefficients can be found in Gower (1986), Fichet (1986) and the exposé by Gower and Legendre (1986). The foremost requirement that must be satisfied by a coefficient, before it is said to be a metric, is the triangle inequality. The other metric axioms are more easily verified. The proofs of the metric properties for two-way similarity coefficients reviewed here, are essential blueprints and tools for the proofs of metric properties of multi-way coefficients discussed later on in the thesis (Chapter 18).

The present chapter focuses solely on metric properties and not on the closely related Euclidean property, which is satisfied if the functions can be embedded in an Euclidean space. Since an Euclidean distance is also a metric, the former is a stronger requirement. The dissimilarity coefficients corresponding to similarity coefficients

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

are not Euclidean using the transformation $D = 1 - S$, but they are Euclidean after transformation $D = \sqrt{1-S}$ (Gower and Legendre, 1986, p. 23).

The transformation $D = 1 - S$, D is the complement of S, can easily be applied to the case of multi-way similarities considered in Part IV. It is however unclear how the transformation $D = \sqrt{1 - S}$ generalizes to multi-way dissimilarities. The transformation is therefore not considered in this chapter.

A property that is often studied in close relation to metric and Euclidean properties, is the concept of positive semidefiniteness. A similarity matrix $\mathbf{S}$ is called positive semidefinite if all eigenvalues are nonnegative, in which case $\mathbf{S}$ is sometimes called a Gramian matrix. This property is not reviewed in this chapter, because no attempt is made to generalize these properties to the multi-way case. Various results on positive semidefinite coefficient matrices with respect to resemblance measures for binary data can be found in Janson and Vegelius (1981), Zegers (1986) and Gower and Legendre (1986).

10.1 Dissimilarity coefficients

In Section 1.2 requirements or axioms for similarities as well as dissimilarities were considered. Let x_1 and x_2 be two variables or objects. A two-way or bivariate function $D(x_1, x_2)$ is referred to as a dissimilarity if it satisfies

$$\begin{aligned} & D(x_1, x_2) \geq 0 && \text{(nonnegativity)} \\ & D(x_1, x_2) = D(x_2, x_1) && \text{(symmetry)} \\ \text{and} \quad & D(x_1, x_1) = 0 && \text{(minimality).} \end{aligned}$$

A straightforward way to transform a similarity coefficient S into a dissimilarity coefficient D is by taking the complement $D = 1 - S$. This require that $S(x_1, x_1) = 1$, otherwise $D(x_1, x_1) \neq 0$. For several coefficients, the transformation $D = 1 - S$ gives simple formulas. For example,

$$\begin{aligned} D_{\text{Jac}} &= 1 - S_{\text{Jac}} = \frac{b + c}{a + b + c} \\ D_{\text{Gleas}} &= 1 - S_{\text{Gleas}} = \frac{b + c}{2a + b + c} = \frac{b + c}{p_1 + p_2} \\ D_{\text{SM}} &= 1 - S_{\text{SM}} = \frac{b + c}{a + b + c + d} = b + c \\ D_{\text{Kul}} &= 1 - S_{\text{Kul}} = \frac{bp_2 + cp_1}{2p_1p_2} \\ D_{\text{Sim}} &= 1 - S_{\text{Sim}} = \frac{\min(b, c)}{\min(p_1, p_2)} \\ D_{\text{BB}} &= 1 - S_{\text{BB}} = \frac{\max(b, c)}{\max(p_1, p_2)}. \end{aligned}$$

In order for coefficient $D_{\text{RR}} = 1 - S_{\text{RR}}$ to satisfy minimality, D_{RR} must be defined as

$$D_{\text{RR}} = \begin{cases} 0 & \text{if } x_1 = x_2 \\ 1 - a & \text{otherwise.} \end{cases}$$

For D to be a metric, it must satisfy the metric axioms definiteness, given by

$$D(x_1, x_2) = 0 \quad \text{if and only if } x_1 = x_2$$

and foremost, the triangle inequality, which is given by

$$D(x_1, x_2) \leq D(x_1, x_3) + D(x_2, x_3). \tag{10.1}$$

10.2 Main results

Inequality (10.1) is the main topic of this chapter. The other metric axioms are less difficult to verify. Since (10.1) describes the relation between three variables or objects instead of just two, some additional notation is required. Let

$$p^{111} = P\left(\overset{1}{x_1}, \overset{1}{x_2}, \overset{1}{x_3}\right)$$

denote the proportion of 1s shared by variables x_1, x_2 and x_3 in the same positions, and let

$$p^{110} = P\left(\overset{1}{x_1}, \overset{1}{x_2}, \overset{0}{x_3}\right)$$

denote the proportion of 1s shared by variables x_1 and x_2, and 0s by variable h_3 in the same positions. With this notation we have that $a = p_{12}^{11} = p^{111} + p^{110}$. For convenience, notation p^{111} will be used instead of $P\left(\overset{1}{x_1}, \overset{1}{x_2}, \overset{1}{x_3}\right)$. The quantities a, b, c, and d have subscripts

$$\begin{aligned} a_{12} &= a(x_1, x_2) \\ b_{12} &= b(x_1, x_2) \\ c_{12} &= c(x_1, x_2) \\ d_{12} &= d(x_1, x_2) \end{aligned}$$

when comparing variables or objects x_1 and x_2. Furthermore, let D_{12} be short for $D(x_1, x_2)$. The subscripts are dropped whenever possible.

Theorem 10.1 covers the metric property for the relatively simple functions given by

$$D_{\text{RR}} = 1 - a \quad \text{and} \quad D_{\text{SM}} = b + c.$$

Theorem 10.1. *Functions D_{RR}, D_{SM} and $D = 1 - d$ satisfy the triangle inequality* (10.1).

Proof: Using D_{RR} in (10.1) we obtain

$$\begin{aligned} 1 - a_{12} &\leq 1 - a_{13} + 1 - a_{23} \\ 2 - 2p^{111} - p^{101} - p^{011} &\geq 1 - p^{111} - p^{110} \\ 1 + p^{110} &\geq p^{111} + p^{101} + p^{011}. \end{aligned} \tag{10.2}$$

Using $D = 1 - d$ and D_{SM} in (10.1) we obtain respectively

$$1 + p^{001} \geq p^{000} + p^{100} + p^{010} \tag{10.3}$$

and

$$1 + p^{110} + p^{001} \geq p^{111} + p^{101} + p^{011} + p^{100} + p^{010} + p^{000}. \tag{10.4}$$

(Interestingly, it does not suffice that for (10.4) to hold, both (10.2) and (10.3) are true). Inequalities (10.2), (10.3) and (10.4) are true because

$$1 = p^{111} + p^{110} + p^{101} + p^{011} + p^{100} + p^{010} + p^{001} + p^{000}. \tag{10.5}$$

□

The proof of the metric property of D_{Jac} is less straightforward compared the proof for coefficients considered in Theorem 10.1. The tool used is not adopted from Gower and Legendre (1986). Instead, the idea comes from Heiser and Bennani (1997), where it is used for three-way dissimilarities. The application below describes the tool for the simpler (two-way) case. In Chapter 18 a generalization of the proof of Theorem 10.2 is used. The next result shows that both

$$D_{\text{Jac}} = \frac{b+c}{a+b+c} \quad \text{and} \quad D = \frac{b+c}{1-a} = \frac{b+c}{b+c+d}$$

satisfy the triangle inequality.

Theorem 10.2. *The functions D_{Jac} and*

$$D = \frac{b+c}{b+c+d}$$

satisfy (10.1).

Proof: We consider the proof for D_{Jac} first. Adding p^{001} to both sides and p^{110} to the left side of (10.5), we obtain

$$1 + p^{110} + p^{001} \geq p^{111} + p^{110} + p^{101} + p^{011} + p^{100} + p^{010} + 2p^{001} + p^{000}$$

which equals

$$(b_{13} + c_{13}) + (b_{23} + c_{23}) - (b_{12} + c_{12}) \geq p^{001}. \tag{10.6}$$

$D_{\text{SM}} = 1 - S_{\text{SM}}$ and D_{Jac} are related by

$$D_{\text{SM}} = (1 - d_{12})\frac{b_{12} + c_{12}}{1 - d_{12}} = (1 - p^{000} - p^{001})D_{\text{Jac}}. \tag{10.7}$$

Using (10.7) in (10.6) we obtain

$$(1 - p^{000})\left[\frac{b_{13} + c_{13}}{1 - d_{13}} + \frac{b_{23} + c_{23}}{1 - d_{23}} - \frac{b_{12} + c_{12}}{1 - d_{12}}\right] \geq$$
$$p^{010}\left[\frac{b_{13} + c_{13}}{1 - d_{13}}\right] + p^{100}\left[\frac{b_{23} + c_{23}}{1 - d_{23}}\right] + p^{001}\left[1 - \frac{b_{12} + c_{12}}{1 - d_{12}}\right].$$

Since $(1 - p^{000}) \geq 0$ and $D_{\text{Jac}} \leq 1$, we conclude that D_{Jac} satisfies (10.1).

Next, we consider the proof for D. Adding p^{110} to both sides and p^{001} to the left side of (10.5), we obtain

$$(b_{13}+c_{13})+(b_{23}+c_{23})-(b_{12}+c_{12}) \geq p^{110} \tag{10.8}$$

instead of (10.6). D_{SM} and D are related by

$$D_{\mathrm{SM}} = (1-a_{12})\frac{b_{12}+c_{12}}{1-a_{12}} = (1-p^{110}-p^{111})D. \tag{10.9}$$

Using (10.9) in (10.8) we obtain

$$(1-p^{111})\left[\frac{b_{13}+c_{13}}{1-a_{13}}+\frac{b_{23}+c_{23}}{1-a_{23}}-\frac{b_{12}+c_{12}}{1-a_{12}}\right] \geq$$
$$p^{101}\left[\frac{b_{13}+c_{13}}{1-a_{13}}\right]+p^{011}\left[\frac{b_{23}+c_{23}}{1-a_{23}}\right]+p^{110}\left[1-\frac{b_{12}+c_{12}}{1-a_{12}}\right].$$

Since $(1-p^{111}) \geq 0$ and $D \leq 1$, we conclude that D satisfies (10.1).
This completes the proof. □

Before studying any other coefficient, we note the following well-known result (see, for example, Gower and Legendre, 1986).

Theorem 10.3. *Let e be a positive constant. If D satisfies* (10.1), *then $D/(e+D)$ satisfies* (10.1).
Proof: We have

$$\frac{D_{12}}{e+D_{12}}+\frac{D_{13}}{e+D_{13}} \geq \frac{D_{23}}{e+D_{23}}$$

if and only if

$$e^2(D_{12}+D_{13}-D_{23})+2eD_{12}D_{13}+D_{12}D_{13}D_{23} \geq 0. \quad \square$$

Combining Theorem 10.3 with Theorem 10.1 or 10.2, various new results can be obtained. Consider the dissimilarities

$$D_{\mathrm{SS1}} = 1-S_{\mathrm{SS1}} = \frac{2(b+c)}{a+2(b+c)} = \frac{2D_{\mathrm{Jac}}}{1+D_{\mathrm{Jac}}}$$
$$\frac{2(b+c)}{2(b+c)+d} = \frac{2D}{1+D} \quad \text{where} \quad D = \frac{b+c}{b+c+d}$$
$$D_{\mathrm{RT}} = 1-S_{\mathrm{RT}} = \frac{2(b+c)}{a+2(b+c)+d} = \frac{2D_{\mathrm{SM}}}{1+D_{\mathrm{SM}}}.$$

Since D_{Jac} and D_{SM} satisfy (10.1), application of Theorem 10.3 leads to the next result.

Proposition 10.1. *The functions* D_{SS3}, D_{RT} *and*

$$D = \frac{2(b+c)}{2(b+c)+d} \quad \textit{satisfy (10.1).}$$

Next, it is shown what other members of

$$D_{\mathrm{GL1}}(\theta) = 1 - S_{\mathrm{GL1}}(\theta) = 1 - \frac{a}{(1-\theta)a + \theta(1-d)}, \tag{10.10}$$

apart from D_{Jac} and D_{SS1}, satisfy the triangle inequality.

Theorem 10.4. *The function* $D_{\mathrm{GL1}}(\theta)$ *satisfies* (10.1) *for* $0 < \theta \leq 1$.
Proof: By Theorem 10.2 $D_{\mathrm{GL1}}(\theta = 1) = D_{\mathrm{Jac}}$ satisfies (10.1). For $0 < \theta < 1$, let $\theta = (e+1)/e$, where e is a positive real number. Then (10.10) can be written as

$$D_{\mathrm{GL1}}(\theta) = \frac{\theta D_{\mathrm{SM}}}{a + \theta D_{\mathrm{SM}}} = \frac{(e+1)D_{\mathrm{SM}}}{ea + (e+1)D_{\mathrm{SM}}}. \tag{10.11}$$

Dividing both numerator and denominator of (10.11) by $1-d$ we obtain

$$D_{\mathrm{GL1}}(\theta) = \frac{(e+1)D_{\mathrm{Jac}}}{eS_{\mathrm{Jac}} + (e+1)D_{\mathrm{Jac}}} = \frac{(e+1)D_{\mathrm{Jac}}}{e + D_{\mathrm{Jac}}}. \tag{10.12}$$

The right part of (10.12) satisfies (10.1) if and only if $D_{\mathrm{Jac}}/(e+D_{\mathrm{Jac}})$ satisfies (10.1). The result then follows from application of the Theorem 10.3. □

10.3 Counterexamples

We finish the chapter with coefficients that do not satisfy the triangle inequality. For each coefficient, it suffices to present a counterexample (see also Gower and Legendre, 1986, Appendix II). Consider the three binary vectors

$$\begin{bmatrix}1\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}1\\1\end{bmatrix}.$$

We have

$$\begin{aligned}
D_{\mathrm{SS2}} &= 1 - \frac{2(a+d)}{2a+b+c+2d} && \rightarrow D_{12} = 1 \text{ and } D_{13} = D_{23} = \frac{1}{3}\\
D_{\mathrm{Gleas}} &= 1 - \frac{2a}{p_1+p_2} && \rightarrow D_{12} = 1 \text{ and } D_{13} = D_{23} = \frac{1}{3}\\
D_{\mathrm{DK}} &= 1 - \frac{a}{\sqrt{p_1 p_2}} && \rightarrow D_{12} = 1 \text{ and } D_{13} = D_{23} = 1 - \frac{1}{\sqrt{2}} < \frac{1}{3}\\
D_{\mathrm{Kul}} &= 1 - \frac{a(p_1+p_2)}{2p_1p_2} && \rightarrow D_{12} = 1 \text{ and } D_{13} = D_{23} = \frac{1}{4}\\
D_{\mathrm{Sim}} &= 1 - \frac{a}{\min(p_1,p_2)} && \rightarrow D_{12} = 1 \text{ and } D_{13} = D_{23} = 0.
\end{aligned}$$

The dissimilarities do not satisfy the triangle inequality.

Consider the three binary vectors

$$\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}1\\1\\0\\0\end{bmatrix}.$$

We have

$$\begin{aligned}
D_{\text{Cohen}} &= 1 - \frac{2(ad-bc)}{p_1q_2 + p_2q_1} && \rightarrow D_{12} = \frac{4}{3} \text{ and } D_{13} = D_{23} = \frac{1}{2} \\
D_{\text{Phi}} &= 1 - \frac{ad-bc}{\sqrt{p_1p_2q_1q_2}} && \rightarrow D_{12} = \frac{4}{3} \text{ and } D_{13} = D_{23} = 1 - \frac{1}{\sqrt{3}} < \frac{1}{2} \\
D_{\text{Loe}} &= 1 - \frac{ad-bc}{\min(p_1q_2, p_2q_1)} && \rightarrow D_{12} = \frac{4}{3} \text{ and } D_{13} = D_{23} = \frac{1}{3}.
\end{aligned}$$

The dissimilarities do not satisfy the triangle inequality.

10.4 Epilogue

Only a few dissimilarities obtained with transformation $D = 1 - S$ turn out to be metric, that is, satisfy the triangle inequality. The key coefficients here are

$$D_{\text{RR}} = 1 - a = b + c + d \quad \text{and} \quad D_{\text{SM}} = 1 - a - d = b + c$$

and

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

Counterexamples were presented for various other coefficients. Since these two-way dissimilarities do not satisfy the triangle inequality, their multi-way formulations presented in Chapters 16 and 17 do not satisfy the generalizations of the triangle inequality considered in Part III of the thesis. Therefore, no metric properties of these coefficients are considered in Chapter 18.

Similarly to Chapters 7 and 8, it may be investigated if one of the functions that do not satisfy the triangle inequality in general, do satisfy the triangle inequality if the data matrix exhibits certain patterns or contains some form of structure. For example, if the data are Guttman vectors, the function

$$D_{\text{Dice}} = 1 - \frac{2a}{p_1 + p_2} \tag{10.13}$$

does satisfy inequality (10.1).

Proposition 10.2. *Suppose that* $a_{12} = \min(p_1, p_2)$. *Then* D_{Dice} *satisfies* (10.1).
Proof: First, let $p_1 \geq p_2 \geq p_3$. Using (10.13) in (10.1), we obtain

$$1 + \frac{2p_2}{p_1 + p_2} \geq \frac{2p_3}{p_1 + p_3} + \frac{2p_3}{p_2 + p_3}. \tag{10.14}$$

Equation (10.14) is true if

$$\begin{aligned}(p_1 + p_2)(p_1 + p_3)(p_2 + p_3) + 2p_2(p_1 + p_3)(p_2 + p_3) \;\geq \\ 2p_3(p_1 + p_2)(p_2 + p_3) + 2p_3(p_1 + p_2)(p_1 + p_3)\end{aligned}$$

if and only if

$$p_1^2(p_2 - p_3) + 3p_1(p_2^2 - p_3^2) + p_2p_3(p_2 - p_3) \geq 0 \tag{10.15}$$

holds. Since $p_2 \geq p_3$, (10.15) is true.

Alternatively, let $p_3 \geq p_2 \geq p_1$. Using (10.13) in (10.1), we obtain

$$1 + \frac{2p_1}{p_1 + p_2} \geq \frac{2p_1}{p_1 + p_3} + \frac{2p_2}{p_2 + p_3}. \tag{10.16}$$

Equation (10.16) is true if

$$\begin{aligned}(p_1 + p_2)(p_1 + p_3)(p_2 + p_2) + 2p_1(p_1 + p_3)(p_2 + p_3) \;\geq \\ 2p_1(p_1 + p_2)(p_2 + p_3) + 2p_2(p_1 + p_2)(p_1 + p_3)\end{aligned}$$

if and only if

$$p_1^2(p_3 - p_2) + 3p_1(p_3^2 - p_2^2) + p_2p_3(p_3 - p_2) \geq 0 \tag{10.17}$$

holds. Since $p_3 \geq p_2$, (10.17) is true. This completes the proof. □

Metric properties given a certain data structure may be investigated for other similarity coefficients as well. The applications of these coefficients would be very limited with respect to the general results for other coefficients in Section 10.2. Such results would be of theoretical interest only.

Part III

Multi-way metrics

CHAPTER 11

Axiom systems for two-way, three-way and multi-way dissimilarities

Dissimilarities are functions that are used with various multivariate data analysis techniques. Well-known examples are multidimensional scaling and cluster analysis. A function is called a dissimilarity if it satisfies certain axioms, that is, it is nonnegative and symmetric, and it satisfies the axiom of minimality. In addition, a dissimilarity may satisfy axioms like the triangle inequality or the ultrametric inequality. Dependencies between certain axioms have been noted by various authors (see, for example, Gower and Legendre (1986), Van Cutsem (1994) or Batagelj and Bren (1995) for the two-way case, and Joly and Le Calvé (1995), Bennani-Dosse (1993) and Heiser and Bennani (1997) for the three-way case).

Although many authors (including the above-mentioned) point out that the used set of axioms do not form a system with a minimum number of axioms (due to dependencies between axioms), it remains (sometimes) unclear what this minimum set looks like. An axiom system can be a minimum set of axioms if it forms an independent system of axioms. Within an axiom system an axiom is called independent if it cannot be derived from the other axioms in the system. Another (perhaps more) important property of an axiom system is consistency. An axiom system is consistent if it lacks contradiction, that is, the ability to derive both a statement and its negation from a set of axioms.

In this chapter the axiom systems for two-way and three-way dissimilarities are studied. Some axioms for two-way dissimilarities were briefly considered in Section 1.2 and Section 10.1. To obtain axiom systems with a minimum number of axioms, the (known) dependencies between various axioms are reviewed. Next, consistency and independence of several axiom systems are established by means of simple models. The remainder of the chapter is used to explore how basic axioms for multi-way dissimilarities, like nonnegativity, minimality and symmetry, may be defined. Generalizations of the two-way metric and the three-way metrics are further studied in Chapter 12. Multi-way extensions of the three-way ultrametric inequalities are investigated in Chapter 13. Using the tools for the axioms for three-way dissimilarities, independence and consistency may be established for the multi-way case.

11.1 Two-way dissimilarities

Let the function $d(x_1, x_2) : E \times E \to \mathbb{R}$ assign a real number to each pair (x_1, x_2), elements of the nonempty set E. The function $d(x_1, x_2)$ is called a two-way dissimilarity between objects x_1 and x_2 if it satisfies the axioms

$$\begin{array}{lll} (A1) & d(x_1, x_2) \geq 0 & \text{(nonnegativity)} \\ (A2) & d(x_1, x_1) = 0 & \text{(minimality)} \\ (A3) & d(x_1, x_2) = d(x_2, x_1) & \text{(symmetry).} \end{array}$$

In the French literature, a dissimilarity $d(x_1, x_2)$ is called respectively semi-proper and proper if it satisfies

$$\begin{array}{lll} (A4) & d(x_1, x_2) = 0 \;\Rightarrow\; d(x_1, x_3) = d(x_2, x_3) & \text{(evenness)} \\ (A5) & d(x_1, x_2) = 0 \;\Rightarrow\; x_1 = x_2 & \text{(definiteness).} \end{array}$$

Let

$$p_{123}^{111} = P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{1}{x}_3\right)$$

denote the proportion of 1s shared by variables x_1, x_2 and x_3 in the same positions, let

$$p_{123}^{110} = P\left(\overset{1}{x}_1, \overset{1}{x}_2, \overset{0}{x}_3\right)$$

denote the proportion of 1s shared by variables x_1 and x_2, and 0s by variable x_3 in the same positions, and let

$$p_1^1 = P\left(\overset{1}{x}_1\right)$$

denote the proportion of 1s in variable x_1. For example, it holds that

$$p_1^1 = p_{12}^{10} + p_{12}^{11} \quad \text{and} \quad p_{12}^{10} = p_{123}^{100} + p_{123}^{101}.$$

Proposition 11.1. $(A1)$, $(A2)$, $(A3)$ and $(A4)$ *form a consistent and independent system of axioms.* $(A1)$, $(A2)$, $(A3)$ and $(A5)$ *form a consistent and independent system of axioms.*
Proof: First, note that $(A5) \Rightarrow (A4)$. Consistency of the two axiom systems is established by the first example of $d(x_1, x_2)$ in the table below. The independence of $(A1)$, $(A2)$ and $(A3)$ with respect to the remaining four axioms is established with the bottom three examples of $d(x_1, x_2)$ in the table below.

	Is the axiom valid?				
$d(x_1, x_2)$	$(A1)$	$(A2)$	$(A3)$	$(A4)$	$(A5)$
$p_1^1 + p_2^1 - 2p_{12}^{11}$	Yes	Yes	Yes	Yes	Yes
$2p_{12}^{11} - p_1^1 - p_2^1$	No	Yes	Yes	Yes	Yes
$p_1^1 + p_2^1 - p_{12}^{11}$	Yes	No	Yes	Yes	Yes
$2p_1^1 + p_2^1 - 3p_{12}^{11}$	Yes	Yes	No	Yes	Yes

Next, consider the function $d(x_1, x_2) = \min(p_1^1, p_2^1) - p_{12}^{11}$. It is readily verified that $d(x_1, x_2)$ satisfies $(A1)$, $(A2)$ and $(A3)$. However, $(A4)$ and $(A5)$ are not valid if there is a pair (x_1, x_2) for which $p_{12}^{11} = \min(p_1^1, p_2^1)$. □

A two-way dissimilarity $d(x_1, x_2)$ is called a distance if it satisfies definiteness and

$$(A6) \quad d(x_1, x_2) \leq d(x_1, x_3) + d(x_2, x_3) \quad \text{(triangle inequality)}.$$

A dissimilarity may also satisfy one of two axioms that define properties of trees, that is, an inequality by Buneman (1974)

$$(A7) \quad d(x_1, x_2) + d(x_3, x_4) \leq \max[d(x_1, x_3) + d(x_2, x_4), d(x_1, x_4) + d(x_2, x_3)]$$

(additive tree) or

$$(A8) \quad d(x_1, x_2) \leq \max[d(x_1, x_3), d(x_2, x_3)] \quad \text{(ultrametric inequality)}.$$

Proposition 11.2.

$$\begin{aligned}
&(i) \quad (A6) \textit{ together with } (A2) \Rightarrow (A1), (A3) \textit{ and } (A4) \\
&(ii) \quad (A7) \textit{ together with } (A2) \Rightarrow (A1), (A3), (A4) \textit{ and } (A6) \\
&(iii) \quad (A8) \textit{ together with } (A2) \Rightarrow (A1), (A3), (A4) \textit{ and } (A6).
\end{aligned}$$

Proof: The proof of (i) can be found in Gower and Legendre (1986, p. 6). For (ii) setting x_3 equal to x_4 in $(A7)$ and applying $(A2)$, we obtain $(A6)$. For (iii), for triplet (x_1, x_1, x_2) we obtain $d(x_1, x_2) \geq 0$, that is (A1). Moreover, $(A8)$ together with $(A1) \Rightarrow (A6)$. □

Proposition 11.3. *(A2), (A5) and (A6) (or (A7) or (A8)) form a consistent and independent system of axioms.*
Proof: Consider the assertion with respect to (A6) first. An example for consistency is the function given by

$$d(x_1, x_2) = 1 - p_{12}^{11} - p_{12}^{00}.$$

Validity of (A2) and (A5) is readily verified. Using $d(x_1, x_2)$ in (A6) we obtain

$$1 + p_{12}^{11} + p_{12}^{00} \geq p_{13}^{11} + p_{13}^{00} + p_{23}^{11} + p_{23}^{00} \qquad \text{if and only if} \qquad 2p_{123}^{110} + 2p_{123}^{001} \geq 0.$$

With respect to independence, consider the function $d(x_1, x_2) = 1 - p_{12}^{11}$. Using $d(x_1, x_2)$ in (A6) we obtain

$$1 + p_{12}^{11} \geq p_{13}^{11} + p_{23}^{11} \qquad \text{if and only if} \qquad p_{123}^{000} + p_{123}^{100} + p_{123}^{010} + p_{123}^{001} + 2p_{123}^{110} \geq 0.$$

Hence, $d(x_1, x_2)$ satisfies (A6). Moreover, axiom (A5) is not violated. However, as long as $p_1^1 \neq 1$, $d(x_1, x_2)$ does not satisfy (A2). Hence, (A2) is independent from (A5) and (A6).

Second, consider the function $d(x_1, x_2) = \min(p_1^1, p_2^1) - p_{12}^{11}$. Axiom (A2) is valid. Assuming $p_1^1 \geq p_2^1 \geq p_3^1$ and Using $d(x_1, x_2)$ in (A6), we obtain

$$2p_3^1 + p_{12}^{11} \geq p_1^1 + p_{13}^{11} + p_{23}^{11} \qquad \text{if and only if} \qquad 2p_{123}^{001} + p_{123}^{101} \geq p_{123}^{010}.$$

Furthermore, (A5) is not valid if $p_{12}^{11} = \min(p_1^1, p_2^1) = p_2^1$ if and only if p_{12}^{01} equals 0. Thus, (A2) and (A6) may be valid, while (A5) is not.

Third, consider the function $d(x_1, x_2) = 2p_{12}^{11} - p_1^1 - p_2^1$. It is readily verified that for this function (A2) and (A5) are valid. However, (A6) is only valid if $p_{123}^{110} + p_{123}^{001} \leq 0$ if and only if $p_{123}^{110} = p_{123}^{001} = 0$, since p_{123}^{110} and p_{123}^{001} are nonnegative quantities.

The proofs of the assertion with respect to (A7) and (A8) are very similar to that of (A6). Furthermore, suppose $d(x_1, x_2)$ satisfies (A8). Then for the three two-way dissimilarities defined on the same three objects, the largest two are equal. This property is unrelated to the value of $d(x_1, x_2)$. □

11.2 Three-way dissimilarities

Axioms for three-way dissimilarities and distances can be found in Bennani-Dosse (1993), Heiser and Bennani (1997) and Chepoi and Fichet (2007). In addition, three-way distances are considered in Joly and Le Calvé (1995). Let $d_3(x_1, x_2, x_3) : E \times E \times E \to \mathbb{R}$ be a function that assigns a real number to each triplet (x_1, x_2, x_3). Heiser and Bennani (1997, p. 191) call $d_3(x_1, x_2, x_3)$ a three-way dissimilarity if it satisfies the axioms

$$\begin{aligned}
(B1a) \quad & d_3(x_1, x_2, x_3) \geq 0 && \text{(nonnegativity)} \\
(B2a) \quad & d_3(x_1, x_1, x_1) = 0 && \text{(minimality)} \\
(B3) \quad & d_3(x_1, x_2, x_3) = d_3(x_1, x_3, x_2) = d_3(x_2, x_1, x_3) = \\
& d_3(x_2, x_3, x_1) = d_3(x_3, x_1, x_2) = d_3(x_3, x_2, x_1) && \text{(symmetry),}
\end{aligned}$$

the three-way generalizations of $(A1)$, $(A2)$ and $(A3)$, and in addition

$$d_3(x_1, x_1, x_2) = d_3(x_1, x_2, x_2). \tag{11.1}$$

Equality (11.1) is referred to as the diagonal-plane equality by Heiser and Bennani (1997), and is also proposed in Joly and Le Calvé (1995).

Equality (11.1) is an answer to a complication that arises with three-way dissimilarities, not encountered with two-way dissimilarities, when one of three variables or entities is identical to one of the others. For this reason, Chepoi and Fichet (2007) studied explicitly the case of three-way dissimilarities for which all entities are different. The lack of resemblance between the two nonidentical entities should, according to Heiser and Bennani (1997), remain invariant regardless of which two entities are the same:

$$\begin{aligned} &d_3(x_1, x_1, x_2) = d_3(x_1, x_2, x_2) = d_3(x_1, x_2, x_1) = \\ &d_3(x_2, x_1, x_1) = d_3(x_2, x_1, x_2) = d_3(x_2, x_2, x_1). \end{aligned}$$

Equality (11.1) is referred to as the diagonal-plane equality in Heiser and Bennani (1997), because it requires equality of the three matrices

$$\{d_3(x_1, x_1, x_2)\}, \ \{d_3(x_1, x_2, x_2)\} \text{ and } \{d_3(x_1, x_2, x_1)\}$$

which are formed by cutting the three-way cube or block diagonally, starting at one of the three edges joining at the node or corner $d(1, 1, 1)$. This seems to be a misnomer, since equality (11.1) only requires equality of the first two matrices. Equality (11.1) together with three-way symmetry $(B3)$ implies the stronger equality

$$(B4) \quad d_3(x_1, x_1, x_2) = d_3(x_1, x_2, x_2) = d_3(x_1, x_2, x_1).$$

Proposition 11.4. *$(B1a)$, $(B2a)$, $(B3)$ and $(B4)$ form a consistent and independent system of axioms.*

Proof: Consistency of the axiom system is shown with the first example of $d_3(x_1, x_2, x_3)$ in the table below.

	Is the axiom valid?			
$d_3(x_1, x_2, x_3)$	$(B1a)$	$(B2a)$	$(B3)$	$(B4)$
$1 - p_{123}^{111} - p_{123}^{000}$	Yes	Yes	Yes	Yes
$p_{123}^{111} + p_{123}^{000} - 1$	No	Yes	Yes	Yes
$1 - p_{123}^{111}$	Yes	No	Yes	Yes
$p_1^1 - p_{123}^{111}$	Yes	Yes	No	Yes
$p_1^1 + p_2^1 + p_3^1 - 3p_{123}^{111}$	Yes	Yes	Yes	No

Independence is established with the bottom four examples of $d_3(x_1, x_2, x_3)$ in the table. Each function satisfies three out of four axioms. □

At this point it should be noted that there exists mathematical literature on multi-way concepts, including distances and metrics, that is older that the above mentioned literature. Some of the references from this literature may be found in Deza and Rosenberg (2000, 2005). Characteristic of this literature are the extensions of axioms $(A1)$ and $(A2)$ given by

$$\begin{aligned} &(B1b) \qquad x_1 \neq x_2 \Rightarrow d_3(x_1, x_2, x_3) > 0 \text{ for some } x_3 \in E \\ &(B2b) \qquad d_3(x_1, x_1, x_2) = 0 \end{aligned}$$

and axiom $(B6c)$ presented below. Axiom $(B2b)$ makes perfect sense in geometry where $d_3(x_1, x_1, x_2)$ is, for example, the area of the triangle with vertices x_1, x_2, and x_3. Deza and Rosenberg (2000, 2005) find axioms $(B1b)$ and $(B2b)$ too restrictive and drop them. The two axioms are also ignored in this chapter.

A three-way dissimilarity $d_3(x_1, x_2, x_3)$ is called a three-way distance in Heiser and Bennani (1997, p. 191) if it satisfies

$$(B5) \qquad d_3(x_1, x_2, x_3) = 0 \quad \Rightarrow \quad x_1 = x_2 = x_3 \qquad \text{(definiteness)}$$

and the so-called tetrahedral inequality

$$(B6a) \qquad 2d_3(x_1, x_2, x_3) \leq d_3(x_2, x_3, x_4) + d_3(x_1, x_3, x_4) + d_3(x_1, x_2, x_4).$$

Alternatively, Joly and Le Calvé (1995) call $d(x_1, x_2, x_3)$ a three-way distance if it satisfies

$$\begin{aligned} &(B6b) \qquad d_3(x_1, x_2, x_3) \leq d_3(x_2, x_3, x_4) + d_3(x_1, x_3, x_4) \\ &(B7) \qquad d_3(x_1, x_2, x_3) \geq d_3(x_1, x_1, x_3) \end{aligned}$$

and a proper three-way distance if it, in addition, satisfies $(B5)$. Axioms $(B6a)$ and $(B6b)$ are called respectively strong and weak metrics in Chepoi and Fichet (2007). Deza and Rosenberg (2000, 2005) present yet another extension of the triangle inequality. The so-called tetrahedron inequality is given by

$$(B6c) \quad d_3(x_1, x_2, x_3) \leq d_3(x_2, x_3, x_4) + d_3(x_1, x_3, x_4) + d_3(x_1, x_2, x_4).$$

Axiom $(B6c)$ is not studied further in this chapter (but see Chapter 12).

Three-way generalizations of two-way ultrametric inequality $(A8)$ are considered in Joly and Le Calvé (1995, p. 195) and Bennani-Dosse (1993, p. 99-110):

$$\begin{aligned} &(B8a) \qquad d_3(x_1, x_2, x_3) \leq \max\left[d_3(x_2, x_3, x_4), d_3(x_1, x_3, x_4)\right] \\ &(B8b) \qquad d_3(x_1, x_2, x_3) \leq \max\left[d_3(x_2, x_3, x_4), d_3(x_1, x_3, x_4), d_3(x_1, x_2, x_4)\right]. \end{aligned}$$

Axioms $(B8a)$ and $(B8a)$ are called respectively strong and weak ultrametrics in Chepoi and Fichet (2007).

As noted in Bennani-Dosse (1993, p. 20), the dependencies between $(B1)$ to $(B8)$ are not as straightforward as the dependencies between $(A1)$ to $(A8)$ given in Proposition 11.2.

Proposition 11.5.

$$\begin{array}{ll} & (B6b) \text{ together with } (B7) \text{ and } (B2a) \Rightarrow (B1a) \\ (i) & (B6b) \text{ together with } (B3) \Rightarrow (B1a) \\ & (B6a) \text{ together with } (B3) \Rightarrow (B1a) \text{ and } (B6b) \\ & (B7) \text{ together with } (B3) \Rightarrow (B4) \\ (ii) & (B8a) \Rightarrow (B6a), (B7) \text{ and } (B8b). \end{array}$$

The proofs for (i) and (ii) are presented below. The proofs of the other assertions can be found in Joly and Le Calvé (1995, p. 193) and Heiser and Bennani (1997, p. 192).
Proof: For (i), adding the two variants of $(B6b)$

$$\begin{array}{ll} & d_3(x_1, x_2, x_3) \leq d_3(x_2, x_3, x_4) + d_3(x_1, x_3, x_4) \\ \text{and} & d_3(x_2, x_3, x_4) \leq d_3(x_1, x_2, x_3) + d_3(x_1, x_3, x_4) \end{array}$$

we obtain $2d_3(x_1, x_3, x_4) \geq 0$. With respect to (ii), note that, if $d(x_1, x_2, x_3)$ satisfies $(B8a)$, then for any four three-way dissimilarities the largest three are equal. □

The dependencies in Proposition 11.5 suggest the independence of various axiom systems. First, we consider a system of structural, that is, non-metric axioms.

Proposition 11.6. $(B1a)$, $(B2a)$, $(B3)$, $(B5)$ *and* $(B7)$ *form a consistent and independent system of axioms.*
Proof: An example of consistency of the axiom system is the function $d_3(x_1, x_2, x_3) = 1 - p_{123}^{111} - p_{123}^{000}$. It is readily verified that $(B1a)$, $(B2a)$, $(B3)$ and $(B5)$ are valid. Using $d_3(x_1, x_2, x_3)$ in $(B7)$ we obtain

$$p_{13}^{11} + p_{13}^{00} \geq p_{123}^{111} + p_{123}^{000} \quad \text{if and only if} \quad p_{123}^{101} + p_{123}^{010} \geq 0.$$

With respect to independence, consider the function $d_3(x_1, x_2, x_3) = 3p_{123}^{111} - p_1^1 - p_2^1 - p_3^1$. Axioms $(B2a)$, $(B3)$ and $(B5)$ are valid, but $(B1a)$ is not. Using the function in $(B7)$ we obtain

$$\begin{array}{c} 3p_{123}^{111} + p_1^1 \geq 3p_{13}^{11} + p_3^1 \\ p_{123}^{100} + p_{123}^{110} \geq 3p_{123}^{101} + p_{123}^{001} + p_{123}^{011} \\ p_{13}^{10} \geq 3p_{123}^{101} + p_{13}^{01}. \end{array}$$

Thus, $(B1a)$ is independent from $(B2a)$, $(B3)$, $(B5)$ and $(B7)$.

Second, consider the function $d_3(x_1, x_2, x_3) = p_1^1 + p_2^1 + p_3^1 - 2p_{123}^{111}$. Axioms $(B1a)$, $(B3)$ and $(B5)$ are valid, but $(B2a)$ is not. The function satisfies $(B7)$ if and only if $p_{12}^{01} + 2p_{123}^{101} \geq p_{12}^{10}$. Thus, axiom $(B2a)$ is independent from $(B1a)$, $(B3)$, $(B5)$ and $(B7)$.

Third, consider the function $d_3(x_1, x_2, x_3) = 2p_1^1 + p_2^1 + p_3^1 - 4p_{123}^{111}$. Axioms $(B1a)$, $(B2a)$ and $(B5)$ are valid, but $(B3)$ is not. The function satisfies $(B7)$ if and only if $p_{12}^{01} + 4p_{123}^{101} \geq p_{12}^{10}$, which shows that $(B3)$ is independent from the remaining four axioms.

Next, consider the function

$$d_3(x_1, x_2, x_3) = \min(p_{12}^{11}, p_{13}^{11}, p_{23}^{11}) - p_{123}^{111}.$$

It is readily verified that $(B1a)$, $(B2a)$, $(B3)$ and $(B7)$ are valid. However, if there is a triple (x_1, x_2, x_3) for which $p_{123}^{111} = \min(p_{12}^{11}, p_{13}^{11}, p_{23}^{11})$, then $(B5)$ does not hold.

Finally, consider the function $d_3(x_1, x_2, x_3) = p_1^1 + p_2^1 + p_3^1 - 3p_{123}^{111}$. It is readily verified that $(B1a)$, $(B2a)$, $(B3)$ and $(B5)$ are valid. Furthermore, we have $d_3(x_1, x_2, x_3) \leq d_3(x_1, x_1, x_2)$ if and only if $p_{12}^{01} + 3p_{123}^{101} \leq p_{12}^{10}$, which show the independence of $(B7)$ with respect to the remaining four axioms. □

Finally, we consider an axiom system with a minimum number of axioms.

Proposition 11.7. *$(B2a)$, $(B3)$, $(B5)$, $(B6a)$ and $(B7)$ form a consistent and independent system of axioms.*

Proof: An example for the consistency of the axiom system is the function $d_3(x_1, x_2, x_3) = 1 - p_{123}^{111} - p_{123}^{000}$. It is readily verified that $(B2a)$, $(B3)$, $(B5)$ and $(B7)$ are valid. Using $d_3(x_1, x_2, x_3)$ in $(B6a)$ we obtain

$$1 - (p_{234}^{111} + p_{134}^{111} + p_{124}^{111} + p_{234}^{000} + p_{134}^{000} + p_{124}^{000}) + 2p_{123}^{111} + 2p_{123}^{000} \geq 0. \tag{11.2}$$

Since the quantity in between brackets in (11.2) is smaller than unity, $(B6a)$ is valid.

With respect to independence, consider the function $d_3(x_1, x_2, x_3) = p_1^1 + p_2^1 + p_3^1 - 2p_{123}^{111}$. Axioms $(B3)$ and $(B5)$ are valid, and $(B2a)$ is not. Using the function in $(B6a)$ we obtain

$$3p_4^1 + 4p_{123}^{111} \geq p_{234}^{111} + p_{134}^{111} + p_{124}^{111}$$

which holds if and only if

$$3p_{1234}^{0001} + 3p_{1234}^{1001} + 3p_{1234}^{0101} + 3p_{1234}^{0011} + p_{1234}^{1101} + p_{1234}^{1011} + p_{1234}^{0111} + p_{1234}^{1111} + 4p_{1234}^{1110} \geq 0.$$

Furthermore, axiom $(B7)$ is valid if and only if

$$p_2^1 + 2p_{12}^{11} \geq p_1^1 + 2p_{123}^{111} \quad \text{if and only if} \quad p_{12}^{01} + 2p_{123}^{110} \geq p_{12}^{10}.$$

Thus, $(B2a)$ is independent from the remaining four axioms.

Second, consider the function $d_3(x_1, x_2, x_3) = 2p^1_1 + p^1_2 + p^1_3 - 4p^{111}_{123}$. Axioms $(B2a)$, $(B5)$ and $(B7)$ are valid, but $(B3)$ is not. Using the function in $(B6a)$, we obtain the inequality

$$p^1_2 + 3p^1_4 + 8p^{111}_{123} \geq 4p^{111}_{234} + 4p^{111}_{134} + 4p^{111}_{124}$$

which holds if and only

$$p^{0100}_{1234} + p^{1100}_{1234} + p^{0110}_{1234} + 4p^{0101}_{1234} + 8p^{1110}_{1234} + 3p^{0001}_{1234} + 3p^{1001}_{1234} + 3p^{0011}_{1234} \geq p^{1011}_{1234}$$

which shows that $(B3)$ is independent from the remaining four axioms.

Third, consider the function

$$d(x_1, x_2, x_3) = \min(p^{11}_{12}, p^{11}_{13}, p^{11}_{23}) - p^{111}_{123}$$

Axioms $(B2a)$, $(B3)$ and $(B7)$ are valid. Assuming $p^{11}_{12} \geq p^{11}_{13} \geq p^{11}_{14} \geq p^{11}_{23} \geq p^{11}_{24} \geq p^{11}_{34}$ and Using $d(x_1, x_2, x_3)$ in $(B6a)$, we obtain

$$2p^{11}_{34} + p^{11}_{24} + 2p^{111}_{123} \geq 2p^{11}_{23} + p^{111}_{234} + p^{111}_{134} + p^{111}_{124}$$

if and only if

$$2p^{0011}_{1234} + p^{1011}_{1234} + p^{0101}_{1234} \geq 2p^{0110}_{1234}.$$

Note that axiom $(B5)$ is not valid if $p^{111}_{123} = \min(p^{11}_{12}, p^{11}_{13}, p^{11}_{23}) = p^{11}_{23}$ if and only if $p^{011}_{123} = 0$. The latter implies that $p^{0110}_{1234} = 0$, from which it follows that $(B6a)$ holds. Thus, $(B5)$ is independent from the remaining four axioms.

Next, consider the function $d_3(x_1, x_2, x_3) = 3p^{111}_{123} - p^1_1 - p^1_2 - p^1_3$. Axioms $(B2a)$, $(B3)$ and $(B5)$ are valid for both $d_3(x_1, x_2, x_3)$ and $-d_3(x_1, x_2, x_3)$. Axiom $(B6a)$ is valid for $-d_3(x_1, x_2, x_3)$, since filling in $-d_3(x_1, x_2, x_3)$ in $(B6a)$ gives

$$p^1_4 + 2p^{111}_{123} \geq p^{111}_{234} + p^{111}_{134} + p^{111}_{124}$$

if and only if

$$2p^{1110}_{1234} + p^{0001}_{1234} + p^{1001}_{1234} + p^{0101}_{1234} + p^{0011}_{1234} \geq 0.$$

Using similar arguments it is clear that $(B6a)$ is not valid for $d_3(x_1, x_2, x_3)$. Finally, $(A7)$ is valid for $d_3(x_1, x_2, x_3)$ not valid for $-d_3(x_1, x_2, x_3)$ if and only if $p^{01}_{12} + 2p^{101}_{123} \leq p^{100}_{123}$. Hence, $(B6a)$ and $(B7)$ are independent from the remaining four axioms. □

11.3 Multi-way dissimilarities

In this final section it is explored how basic axioms for multi-way dissimilarities, like nonnegativity, minimality and symmetry, may be defined. However, axioms for the four-way and five-way case are considered first. Generalizations of the two-way metric and the three-way metrics to k-way metrics are further studied in the next chapter (Chapter 12). Multi-way formulations of the three-way ultrametrics are explored in Chapter 13. Independence and consistency of axioms for multi-way dissimilarities may be established using the tools from the previous section.

As it turns out, definitions of some axioms are considerably more complicated in the four-way case compared to the three-way case. Let

$$d_4(x_1, x_2, x_3, x_4) : E^4 \to \mathbb{R} \quad \text{or} \quad d_{1234} : E^4 \to \mathbb{R}$$

be a function that assigns a real number to each quadruplet (x_1, x_2, x_3, x_4). Formulations of nonnegativity and minimality are straightforward:

$$\begin{array}{lll} (C1) & d_4(x_1, x_2, x_3, x_4) \geq 0 & \text{(nonnegativity)} \\ (C2) & d_4(x_1, x_1, x_1, x_1) = 0 & \text{(minimality).} \end{array}$$

The definition of four-way symmetry is somewhat more involved. Four-way symmetry is given by

$$\begin{aligned} d_{1234} = d_{1243} = d_{1324} = d_{1342} = d_{1423} = d_{1432} = \\ d_{2134} = d_{2143} = d_{2314} = d_{2341} = d_{2413} = d_{2431} = \\ d_{3124} = d_{3142} = d_{3214} = d_{3241} = d_{3412} = d_{3421} = \\ d_{4123} = d_{4132} = d_{4213} = d_{4231} = d_{4312} = d_{4321}. \end{aligned}$$

If $d_4(x_1, x_2, x_3, x_4)$ is four-way symmetric, then for all $x_1, x_2, x_3, x_4 \in E$ and every permutation π of $\{1, 2, 3, 4\}$

$$(C3) \qquad d_4(x_{\pi(1)}, x_{\pi(2)}, x_{\pi(3)}, x_{\pi(4)}) = d_4(x_1, x_2, x_3, x_4).$$

Similar to the three-way case, the four-way function can be defined on a quadruplet or four-tuple of which some entities are identical. Following the reasoning in Heiser and Bennani (1997), it seems reasonable to require that when one of four variables or entities is identical to one of the others, then the lack of resemblance between the three nonidentical entities should remain invariant regardless of which two entities are the same. A generalization of equality (11.1) is given by

$$d_4(x_1, x_1, x_2, x_3) = d_4(x_1, x_2, x_2, x_3) = d_4(x_1, x_2, x_3, x_3) \tag{11.3}$$

or $d_{1123} = d_{1223} = d_{1233}$. Equality (11.3) together with four-way symmetry, implies

$$\begin{aligned} d_{1123} = d_{1132} = d_{1213} = d_{1312} = d_{1231} = d_{1321} = \\ d_{2113} = d_{3112} = d_{2131} = d_{3121} = d_{2311} = d_{3211} = \\ d_{2213} = d_{2231} = d_{2123} = d_{2321} = d_{2132} = d_{2312} = \\ d_{1223} = d_{3221} = d_{1232} = d_{3212} = d_{1322} = d_{3122} = \\ d_{3312} = d_{3321} = d_{3132} = d_{3231} = d_{3123} = d_{3213} = \\ d_{1332} = d_{2331} = d_{1323} = d_{2313} = d_{1233} = d_{2133}. \end{aligned}$$

The latter equality is the mathematical formulation of the requirement that, when one of four vectors or entities is identical to one of the others, then the lack of similarity between the three nonidentical entities should remain invariant regardless of which two entities are the same.

Apart from the possibility that two entities are identical, up to two additional possibilities may be encountered in the four-way case. First of all, the four-way function may be defined on a quadruplet of which three entities are identical. Secondly, the four-way function may be defined on two pairs of identical entities. Following the above reasoning, we require that if the resemblance between two groups of identical entities is measured, then the lack of resemblance between the two nonidentical groups should remain invariant regardless of the group sizes. The requirement may be formalized with the definition of equality

$$d_4(x_1, x_1, x_1, x_2) = d_4(x_1, x_1, x_2, x_2) = d_4(x_1, x_2, x_2, x_2) \tag{11.4}$$

or $d_{1112} = d_{1122} = d_{1222}$. Equality (11.4), together with four-way symmetry, implies

$$\begin{aligned} & d_{1112} = d_{1121} = d_{1211} = d_{2111} \\ =& d_{1122} = d_{1212} = d_{1221} = d_{2112} = d_{2121} = d_{2211} \\ =& d_{1222} = d_{2122} = d_{2212} = d_{2221}. \end{aligned}$$

The definitions of axioms for five-way dissimilarities are now straightforward. Let

$$d_5(x_1, x_2, x_3, x_4, x_5) : E^5 \to \mathbb{R} \quad \text{or} \quad d_{12345} : E^5 \to \mathbb{R}$$

be a function that assigns a real number to each tuple $(x_1, x_2, x_3, x_4, x_5)$. The basic axioms for the five-way case are

$(D1)$	$d_5(x_1, x_2, x_3, x_4, x_5) \geq 0$	(nonnegativity)
$(D2)$	$d_5(x_1, x_1, x_1, x_1, x_1) = 0$	(minimality)
$(D3)$	$d_5(x_{\pi(1)}, x_{\pi(2)}, ..., x_{\pi(5)}) = d_5(x_1, x_2, ..., x_5)$	(symmetry).

In the case that two out of five entities are identical, the first additional requirement is given by

$$d_{11234} = d_{12234} = d_{12334} = d_{12344}.$$

If there are three sets of identical entities (size of the set unspecified), the second additional requirement is given by

$$d_{11123} = d_{12223} = d_{12333} = d_{11223} = d_{11233} = d_{11233}.$$

When there are two sets of identical entities (size of the set unspecified), the third additional requirement is given by

$$d_{11112} = d_{11122} = d_{11222} = d_{12222}.$$

Thus, for the k-way case up to $(k-2)$ additional requirements must be specified to cover all the cases of identical entities or objects.

For the definition of the axioms for general multi-way dissimilarities the following notation is used. Let $x_{1,k} = \{x_1, x_2, ..., x_k\}$ be a k-tuple and let

$$d_k(x_{1,k}) : E^k \rightarrow \mathbb{R}$$

denote the multi-way dissimilarity for k objects or variables. The basic axioms for the measure $d_k(x_{1,k})$ are given by

$(K1)$	$d_k(x_{1,k}) \geq 0$	(nonnegativity)
$(K2)$	$d_k(\mathbf{x}_1) = 0$	(minimality)
$(K3)$	$d_k(x_{1,k}) = d_k(x_{\pi(1)}, x_{\pi(2)}, ..., x_{\pi(k)})$	(symmetry)

where $\mathbf{x}_1$ is a k-tuple with elements x_1.

11.4 Epilogue

The topic of this chapter was axioms, like nonnegativity, minimality and symmetry, for two-way, three-way and general multi-way dissimilarities. Generalizations of the triangle inequality are studied in the next chapter, Chapter 12. For the axioms of two-way and three-way dissimilarities several axiom systems were studied. Using simple models, the consistency and independence of these axiom systems were established.

In the final section of the chapter axioms of multi-way dissimilarities were considered. Multi-way axioms are already quite complicated for the four-way and five-way case. Multi-way definitions of nonnegativity, minimality and symmetry are straightforward. If $x_{1,k}$ is a k-tuple, then $d(x_{1,k}) = 0$ if all elements in $x_{1,k}$ are identical. However, for $k \geq 3$ it may occur that not all but some elements in $x_{1,k}$ are identical. Additional axioms are required to deal with these new possibilities. For the three-way case Heiser and Bennani (1997) required that when one of three variables is identical to one of the others, then the lack of resemblance between the two non-identical entities should remain invariant regardless of which two entities are the same. Following this line of reasoning, additional axioms may be formulated for the four-way case, the five-way case, and the general multi-way case.

CHAPTER 12

Multi-way metrics

Measures of resemblance play an important role in many domains of data analysis. However, similarity coefficients often only allow pairwise or two-way comparison of objects or entities. An alternative to two-way resemblance measures is to formulate multi-way coefficients (see, for example, Diatta, 2006, 2007). Several authors have studied three-way dissimilarities and generalized various concepts defined for the two-way case to the three-way case (see, for example, Bennani-Dosse, 1993; Joly and Le Calvé, 1995; Heiser and Bennani, 1997). Axioms for two-way and three-way dissimilarities were reviewed in the previous chapter. Chapter 11 was also used to investigate and formulate basic axioms, like nonnegativity, minimality and symmetry for multi-way dissimilarities. In the present chapter extensions of the two-way metric and the three-way metric axioms are explored. Chapter 13 is concerned with extensions of the two three-way ultrametric axioms.

In mathematics, a metric space is a set where a notion of distance between elements of the set is defined. A two-way dissimilarity is called a metric if it is nonnegative, symmetric, satisfies minimality, and (most importantly) if it satisfies the triangle inequality. Both Joly and Le Calvé (1995) and Heiser and Bennani (1997) have considered three-way generalizations of the triangle inequality, defined for the two-way case. The two different metrics are called weak and strong in Chepoi and Fichet (2007). In this chapter the ideas on three-way metrics presented in Joly and Le Calvé (1995) and Heiser and Bennani (1997) are adopted and extended to multi-way metrics.

The inspiration for this chapter on multi-way metricity comes from the paper by Heiser and Bennani (1997). Various ideas on, and properties of, the three-way tetrahedral inequality presented in their paper, are extended in this chapter for a broad class of inequalities that generalize the triangle inequality. An important topic is how the k-way inequalities are related to the $(k-1)$-way inequalities.

12.1 Definitions

In this chapter we study a family of k-way metrics that generalize the two-way metric. Let $x_{1,k}$ denote the k-tuple $(x_1, x_2, ..., x_k)$ and let $x_{1,k}^{-i}$ denote the $(k-1)$-tuple $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_k)$ where the minus in the superscript of $x_{1,k}^{-i}$ is used to indicate that element x_i drops out. In the following the elements of tuple $x_{1,k}$ will be referred to as objects.

A dissimilarity $d_k : E^k \to \mathbb{R}_+$ is totally symmetric if for all $x_1, x_2, ..., x_k \in E$ and every permutation π of $\{1, 2, ..., k\}$

$$d_k(x_{\pi(1)}, ..., x_{\pi(k)}) = d_k(x_1, ..., x_k).$$

As a generalization of minimality we define $d_k(x_1, ..., x_1) = 0$. It is assumed throughout the chapter that the equations hold for all objects in E that are involved in a definition.

Both Joly and Le Calvé (1995) and Heiser and Bennani (1997) introduced three-way generalizations of the triangle inequality. The two inequalities are given by respectively

$$d_3(x_{1,3}) \leq d_3(x_{2,4}) + d_3(x_{1,4}^{-2}) \tag{12.1}$$

$$2d_3(x_{1,3}) \leq d_3(x_{2,4}) + d_3(x_{1,4}^{-2}) + d_3(x_{1,4}^{-3}). \tag{12.2}$$

Inequalities (12.1) and (12.2) are called respectively weak and strong metrics in Chepoi and Fichet (2007). Deza and Rosenberg (2000, 2005) generalize (12.1) to

$$d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}). \tag{12.3}$$

De Rooij (2001, p. 128) noted that inequality (12.2) can be generalized to

$$(k-1) \times d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}) \quad \text{(the polyhedral inequality)}. \tag{12.4}$$

We may generalize (12.3) and (12.4) to

$$u \times d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}) \tag{12.5}$$

where u is a positive real number. We can further generalize (12.5) to

$$u \times d_k(x_{1,k}) \leq \sum_{i=1}^{v} d_k(x_{1,n+1}^{-i}) \tag{12.6}$$

where v is a positive integer bounded by $2 \leq v \leq k$. Note that the number of linear terms on the right-hand side of (12.5) is determined by k, whereas the number of linear terms on the right-hand side of (12.6) is determined by v.

If u^* is a positive integer and $u \geq u^*$, then (12.6) implies

$$u^* \times d_k(x_{1,k}) \leq \sum_{i=1}^{v} d_k(x_{1,k+1}^{-i}).$$

Furthermore, if $v \leq v^*$, then (12.6) implies

$$u \times d_k(x_{1,k}) \leq \sum_{i=1}^{v^*} d_k(x_{1,k+1}^{-i}).$$

Moreover, for $u = 1$ and $k = 1$, adding the two inequalities

$$\begin{aligned} & d_k(x_{1,k}) \leq d_k(x_{2,k+1}) + d_k(x_{1,k+1}^{-2}) \\ \text{and} \quad & d_k(x_{2,k+1}) \leq d_k(x_{1,k}) + d_k(x_{1,k+1}^{-2}) \end{aligned}$$

shows that dissimilarity $d_k(x_{1,k}) \geq 0$. In addition, we have the following property.

Proposition 12.1. *For $u > 1$, (12.6) implies*

$$(u-1) \times d_k(x_{1,k}) \leq \sum_{i=2}^{v} d_k(x_{1,k+1}^{-i}). \tag{12.7}$$

Proof: Interchanging the roles of x_1 and x_{k+1} in (12.6) and dividing the result by u, we obtain

$$d_k(x_{2,k+1}) \leq \frac{1}{u}\, d_k(x_{1,k}) + \frac{1}{u} \sum_{i=2}^{v} d_k(x_{1,k+1}^{-i}). \tag{12.8}$$

Adding (12.8) to (12.6) we obtain

$$\frac{u^2-1}{u} \times d_k(x_{1,k}) \leq \frac{u+1}{u} \sum_{i=2}^{v} d_k(x_{1,k+1}^{-i}). \tag{12.9}$$

Using $u^2 - 1 = (u+1)(u-1)$, multiplication of (12.9) by $u/(u+1)$ yields (12.7). □

12.2 Two identical objects

In the remainder of the chapter we are interested in how dissimilarity d_k is related to d_{k-1}. In Section 12.3 we consider lower and upper bounds of d_k in terms of d_{k-1}. Furthermore, in Section 12.4 we study what $(k-1)$-way metrics are implied by (12.6). Apart from minimality, symmetry and (12.6), we discuss below several additional requirements that specify how d_k and d_{k-1} are related when two objects of d_k are identical.

A first requirement is the following condition. Following Heiser and Bennani (1997) for the three-way case and Deza and Rosenberg (2000, 2005) for the k-way case, we require that, if two objects are identical then d_k should remain invariant regardless which two objects are the same, that is,

$$d_k(x_1, x_{1,k-1}) = d_k(x_{1,2}, x_{2,k-1}) = ... = d_k(x_{1,k-1}, x_{k-1}). \tag{12.10}$$

In view of the total symmetry, (12.10) implies that $d_k(x_1, ..., x_k)$ only depends on the h-element set $\{x_{i_1}, ..., x_{i_h}\}$ such that $\{x_1, ..., x_k\} = \{x_{i_1}, ..., x_{i_h}\}$ where $1 \leq i_1 \leq i_h \leq k$. We consider the following example that satisfies (12.10).

Deza and Rosenberg (2000, p. 803) introduced the k-way extension of the three-way star distance discussed in Joly and Le Calvé (1995). Let $|\{x_1, ..., x_n\}|$ denote the cardinality of set $\{x_1, ..., x_k\}$. Let $\alpha : E \to \mathbb{R}_+$ and $k \geq 3$. The star k-distance $d_k^\alpha : E^k \to \mathbb{R}_+$ is defined as follows. Let $x_1, ..., x_k \in E$ and let $0 \leq i_1 \leq ... \leq i_h \leq k$ be such that $|\{x_1, ..., x_k\}| = |\{x_{i_1}, ..., x_{i_h}\}| = h$. Set

$$d_k^\alpha(x_{1,k}) = \begin{cases} \sum_{j=1}^h \alpha(x_{i_j}) & \text{if } h > 1, \\ 0 \quad \text{if } h = 1. \end{cases}$$

Deza and Rosenberg (2000, p. 803) showed that the star k-distance d_k^α satisfies (12.10).

Condition (12.10) is perhaps not an intuitive requirement, since it may not hold for certain functions. For example, the perimeter distance gives a geometrical interpretation of the concept "average distance" between objects. Heiser and Bennani (1997) and De Rooij and Gower (2003) study the three-way perimeter distance function

$$d_3^p(x_{1,3}) = d(x_1, x_2) + d(x_1, x_3) + d(x_2, x_3). \tag{12.11}$$

A possible k-way extension of (12.11) is

$$d_k^p(x_{1,k}) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(x_i, x_j).$$

Perimeter distance d_k^p is the sum of all pairwise distances between the objects involved. It may be verified that d_k^p does not satisfy (12.10) for $k \geq 4$.

In the remainder of this chapter it is assumed that $d_k(x_{1,k})$ satisfies (12.10). To relate a k-way dissimilarity d_k to a $(k-1)$-way dissimilarity d_{k-1}, we study two additional restrictions. Let p be a real positive value. Suppose that, if two objects of the k-way dissimilarity are identical, d_k and d_{k-1} are equal up to multiplication by a factor p, that is,

$$d_{k-1}(x_{1,k-1}) = \frac{1}{p}\, d_k(x_1, x_{1,k-1}). \tag{12.12}$$

The value of p in (12.12) may depend on the particular distance model or function that is used. For example, Joly and Le Calvé (1995) introduce the three-way semi-perimeter distance

$$d_3^{sp}(x_{1,3}) = \frac{d(x_1, x_2) + d(x_1, x_3) + d(x_2, x_3)}{2}. \tag{12.13}$$

Applying (12.11) with tuple (x_1, x_1, x_2) we obtain $d_3^p(x_1, x_1, x_2) = 2d(x_1, x_2)$. However, applying (12.13) with tuple (x_1, x_1, x_2) we obtain $d_3^{sp}(x_1, x_1, x_2) = d(x_1, x_2)$.

For generality we let p in (12.12) be a positive real number. Of course, it may be argued that $p \geq 1$. The bounds studied in the Section 12.3 depend on the value of p. The bounds of d_k in terms of the d_{k-1} therefore depend on the distance function that is used to relate the k-way dissimilarity and $(k-1)$-way dissimilarity. The results in Section 12.4 however, do not depend on the value of p.

The final requirement we discuss in this section is given by

$$d_k(x_1, x_{1,k-1}) \leq d_k(x_{1,k}). \tag{12.14}$$

In (12.14), the k-way dissimilarity without identical objects is equal to or greater than the k-way dissimilarity with two identical objects. Condition (12.14) seems to be a natural requirement for a multi-way dissimilarity. Combining (12.12) and (12.14) we obtain

$$p\, d_{k-1}(x_{1,k-1}) \leq d_k(x_{1,k}). \tag{12.15}$$

12.3 Bounds

In this section we study the lower and upper bounds of dissimilarity d_k in terms of the d_{k-1}. We first turn our attention to the lower bound of k-way dissimilarity $d_k(x_{1,k})$ that satisfies minimality, total symmetry, and (12.10).

Proposition 12.2. *If* (12.12) *and* (12.14) *hold, then for k-way dissimilarity $d_k(x_{1,k})$ we have*

$$\frac{p}{k}\sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \leq d_k(x_{1,k}). \tag{12.16}$$

Proof: For given k, there are k variants of $d_{k-1}(x_{1,k-1})$, which are given by $d_{k-1}(x_{1,k}^{-i})$ for $i = 1, 2, ..., k$. We obtain k variants of (12.15) by substituting $d_{k-1}(x_{1,k-1})$ on the left-hand side of (12.15) by one of its variants. Adding up all k variants of (12.15), that is, adding inequalities

$$\begin{aligned}
p\, d_{k-1}(x_{1,k}^{-k}) &\leq d_k(x_{1,k}) \\
p\, d_{k-1}(x_{1,k}^{-(k-1)}) &\leq d_k(x_{1,k}) \\
&\vdots \\
p\, d_{k-1}(x_{1,k}^{-3}) &\leq d_k(x_{1,k}) \\
p\, d_{k-1}(x_{1,k}^{-2}) &\leq d_k(x_{1,k}) \\
p\, d_{k-1}(x_{2,k}) &\leq d_k(x_{1,k})
\end{aligned}$$

followed by division by k, we obtain (12.16). □

For $p = 1$, lower bound (12.16) is equivalent to the arithmetic mean of the $(k-1)$-way dissimilarities $d_{k-1}(x_{1,k}^{-i})$.

For the case $(u - v + 2) > 0$, we have the following lower bound for a k-way distance (that is, $d_k(x_{1,n})$ satisfies minimality, total symmetry, (12.6) and (12.10)). In contrast to Proposition 12.2, we only require validity of (12.12), not (12.14), for this lower bound.

Proposition 12.3. *Suppose* (12.12) *holds and* $(u - v + 2) > 0$. *Then for* k*-way distance* $d_k(x_{1,k})$ *we have*

$$\frac{p(u-v+2)}{2k} \sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \leq d_k(x_{1,k}). \tag{12.17}$$

Proof: Applying (12.6) with $(k+1)$-tuple $(x_1, x_1, x_3, ..., x_{k+1})$, and replacing x_{k+1} by x_2 in the result, we obtain

$$p\,u \times d_{k-1}(x_{1,k}^{-2}) \leq 2d_k(x_{1,k}) + p \sum_{i=3}^{v} d_{k-1}(x_1, x_2, x_{3,k}^{-i}) \quad \text{for} \quad v \geq 3 \tag{12.18}$$

$$p\,u \times d_{k-1}(x_{1,k}^{-2}) \leq 2d_k(x_{1,k}) \quad \text{for} \quad v = 2. \tag{12.19}$$

We have k variants of d_{k-1} for given k, for example $d_{k-1}(x_{1,k}^{-2})$ in left-hand side of (12.19). We may obtain k variants of (12.19) by replacing $d_{k-1}(x_{1,k}^{-2})$ by one of the other $(k-1)$ variants. Adding up all k variants of (12.19), followed by division by $2k$, we obtain

$$\frac{p\,u}{2k} \sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \leq d_k(x_{1,k})$$

which is the inequality that is obtained by using $v = 2$ in (12.17).

We may obtain k variants of (12.18) by replacing $d_{k-1}(x_{1,k}^{-2})$ in the left-hand side of (12.18) by one of the other $(k-1)$ variants. Considering all k variants of (12.18), the k variants of d_{k-1} on the right-hand side each occur a total of $(v-2)$ times. Adding up all k variants of (12.18), followed by division by $2k$, we obtain (12.17). □

If (12.12) and (12.4) hold, then $d_k(x_{1,k})$ has a lower bound

$$\frac{p}{2k} \sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \leq d_k(x_{1,k}). \tag{12.20}$$

We obtain (12.20) by using $u = k - 1$ and $v = k$ in (12.17). For $p = 2$ the lower bound of $d_k(x_{1,k})$ is equivalent to the arithmetic mean of the $(k-1)$-way dissimilarities $d_{k-1}(x_{1,k}^{-i})$. If not only (12.12) but also (12.14) is valid, then (12.16) is the lower bound of $d_k(x_{1,k})$. Note that (12.16) is sharper than (12.20).

Next, we focus on the upper bound of k-way distance $d_k(x_{1,k})$.

Proposition 12.4. *If* (12.12) *holds, then for* k*-way distance* $d_k(x_{1,k})$ *we have*

$$d_k(x_{1,k}) \leq \frac{vp}{ku} \sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \quad \textit{for} \quad 2 \leq v \leq k-1 \tag{12.21}$$

$$d_k(x_{1,k}) \leq \frac{(k-1)p}{k(u-1)} \sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \quad \textit{for} \quad v = k. \tag{12.22}$$

Proof: Applying (12.6) with $(k+1)$-tuple $(x_1, ..., x_k, x_k)$ we obtain

$$u \times d_k(x_{1,k}) \leq p \sum_{i=1}^{v} d_{k-1}(x_{1,k}^{-i}) \quad \text{for} \quad 2 \leq v \leq k-1 \tag{12.23}$$

$$(u-1) \times d_k(x_{1,k}) \leq p \sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}) \quad \text{for} \quad v = k. \tag{12.24}$$

We have k variants of $d_{k-1}(x_{1,k}^{-i})$ in (12.23) and (12.24). Considering all k variants of (12.23) and (12.24), each $d_{k-1}(x_{1,k}^{-i})$ occurs a total of v times. Adding up all k variants of (12.23) and (12.24), followed by division by ku, respectively $k(u-1)$, we obtain (12.21) and (12.22). □

Using $u = k$ and $v = k$ in (12.6) yields

$$k \times d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}). \tag{12.25}$$

If (12.12) and (12.25) hold, then the k-way distance $d_k(x_{1,k})$ is bounded from above by

$$d_k(x_{1,k}) \leq \frac{p}{k} \sum_{i=1}^{k} d_k(x_{1,k}^{-i}). \tag{12.26}$$

We obtain (12.26) by using $u = k$ in (12.22). For $p = 1$ the upper bound of $d_k(x_{1,k})$ is equivalent to the arithmetic mean of the $(k-1)$-way distances $d_{k-1}(x_{1,k}^{-i})$.

12.4 $(k-1)$-Way metrics implied by k-way metrics

In this section we study what $(k-1)$-way metrics are implied by the family of k-way metrics defined in (12.6). Again k-way dissimilarity $d_k(x_{1,k})$ satisfies minimality, total symmetry, and (12.10). It is interesting to note that, although we use condition (12.12) throughout this section, the results do not depend on the value of p in (12.12). Unless stated otherwise we assume $k \geq 3$ throughout this section.

Proposition 12.5. *If* (12.12) *and* (12.14) *hold, then* (12.6) *implies*

$$u \times d_{k-1}(x_{1,k-1}) \leq \sum_{i=1}^{v} d_{k-1}(x_{1,k}^{-i}) \quad \textit{for} \quad 2 \leq v \leq k-1 \tag{12.27}$$

$$(u-1) \times d_{k-1}(x_{1,k-1}) \leq \sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}) \quad \textit{for} \quad v = k,\ k > 1. \tag{12.28}$$

Proof: Inequalities (12.27) and (12.28) are obtained from combining (12.15) with (12.23), respectively (12.24). □

As it turns out, condition (12.14) is not required to obtain (12.27). We first show that if (12.12) holds, then (12.6) implies (12.27) for $k \geq 4$ and $2 \leq v \leq k-2$.

Proposition 12.6. *If* (12.12) *holds, then* (12.6) *implies* (12.27) *for* $k \geq 4$ *and* $2 \leq v \leq k-2$.
Proof: Applying (12.6) with $(k+1)$-tuple $(x_1, ..., x_{k-1}, x_{k-1}, x_{k+1})$ and replacing x_{k+1} by x_k in the result, we obtain (12.27). □

Using $v = k-1$ in (12.6) we obtain

$$u \times d_k(x_{1,k}) \leq \sum_{i=1}^{k-1} d_k(x_{1,k+1}^{-i}). \tag{12.29}$$

Using $v = k-1$ in (12.27) we obtain

$$u \times d_{k-1}(x_{1,k-1}) \leq \sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}). \tag{12.30}$$

Next, we show that if (12.12) holds, then (12.29) implies (12.30) for $u \geq 1$.

Proposition 12.7. *If* (12.12) *holds, then for* $u \geq 1$, (12.29) *implies* (12.30).
Proof: Applying (12.29) with $(k+1)$-tuple $(x_1, ..., x_{k-1}, x_{k-1}, x_{k+1})$ and replacing x_{k+1} by x_k in the result, we obtain

$$p\,u \times d_{k-1}(x_{1,k-1}) \leq p \sum_{i=1}^{k-2} d_{k-1}(x_{1,k}^{-i}) + d_k(x_{1,k}). \tag{12.31}$$

Using $v = k - 1$ in (12.23) we obtain

$$u \times d_k(x_{1,k}) \leq p \sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}). \tag{12.32}$$

Adding (12.33) to $u \times$ (12.31) yields

$$u^2 \times d_{k-1}(x_{1,k-1}) \leq u \sum_{i=1}^{k-2} d_{k-1}(x_{1,k}^{-i}) + d_{k-1}(x_{1,k}^{-(k-1)}). \tag{12.33}$$

Apart from variant $d_{k-1}(x_{1,k-1})$ on the left-hand side of (12.33), there are $(k-1)$ variants of d_{k-1}, for example, variant $d_{k-1}(x_{1,k}^{-(k-1)})$, on the right-hand side of (12.33). We have $(k-1)$ variants of (12.33) by varying all $(k-1)$ variants of d_{k-1} on the right-hand side of (12.33). Adding up all $(k-1)$ variants of (12.33), followed by division by $(k-1)u$, yields

$$u \times d_{k-1}(x_{1,k-1}) \leq \left[\frac{(k-2)u+1}{(k-1)u} \right] \sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}). \tag{12.34}$$

To complete the proof, it must be shown that parametrized inequality (12.34) is stronger than (12.30). We have

$$\frac{(k-2)u+1}{(k-1)u} \leq 1$$

if and only if $u \geq 1$. The latter requirement is true under the conditions of the theorem. This completes the proof. □

Using $v = k$ in (12.6) we obtain (12.5). From Proposition 12.5 we know that if both (12.12) and (12.14) hold, then (12.5) implies (12.28). If only (12.12) is valid, (12.5) implies the parametrized inequality

$$(u-1) \times d_{k-1}(x_{1,k-1}) \leq \left[1 + \frac{k-u}{(k-1)u} \right] \sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}). \tag{12.35}$$

Proposition 12.8. *If* (12.12) *holds, then for* $u > 1$, (12.5) *implies* (12.35).
Proof: Applying (12.5) with $(k+1)$-tuple $(x_1, ..., x_{k-1}, x_{k-1}, x_{k+1})$ and replacing x_{k+1} by x_k in the result, we obtain

$$p\, u \times d_{k-1}(x_{1,k-1}) \leq p \sum_{i=1}^{k-2} d_{k-1}(x_{1,k}^{-i}) + 2d_k(x_{1,k}). \tag{12.36}$$

Adding $2 \times$ (12.24) to $(u-1) \times$ (12.36) we obtain

$$u(u-1) \times d_{k-1}(x_{1,k-1}) \leq (u+1) \sum_{i=1}^{k-2} d_{k-1}(x_{1,k}^{-i}) + 2d_{k-1}(x_{1,k}^{-(k-1)}). \tag{12.37}$$

Apart from variant $d_{k-1}(x_{1,k-1})$ on the left-hand side of (12.37), there are $(k-1)$ variants of d_{k-1} on the right-hand side of (12.37). We have $(k-1)$ variants of (12.37) by varying all $(k-1)$ variants of d_{k-1} on the right-hand side of (12.37). Adding up these $(k-1)$ variants of (12.37), followed by division by $(k-1)u$, yields (12.35). □

The parametrized inequality (12.35) is weaker than (12.28) for $k > u$, and stronger than (12.28) for $3 \leq k < u$. With respect to quantity

$$1 + \frac{k-u}{(k-1)u} \tag{12.38}$$

in (12.35) we have limits

$$\lim_{k\to\infty}\left[1+\frac{k-u}{(k-1)u}\right] = 1+\frac{1}{u}, \qquad \lim_{u\to\infty}\left[1+\frac{k-u}{(k-1)u}\right] = 1-\frac{1}{k}$$

and

$$\lim_{k,u\to\infty}\left[1+\frac{k-u}{(k-1)u}\right] = 1.$$

Because of these limits it may be argued that (12.38) and (12.35) are only interesting for small k and u. Furthermore, if $k = u$, then (12.39) $= 1$, and (12.35) is equivalent to (12.28).

Using $u = k-1$ in (12.5) we obtain the polyhedral inequality (12.4). If (12.12) holds, then for $k \geq 3$ the polyhedral inequality (12.4) implies

$$(u-2) \times d_{k-1}(x_{1,k-1}) \leq \left[1+\frac{1}{(k-1)^2}\right]\sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}). \tag{12.39}$$

We obtain (12.39) by using $u = k-1$ in (12.35) and noting that $k^2-2k+2 = (k-1)^2+1$. The quantity

$$1+\frac{1}{(k-1)^2} \quad \text{in (12.39) with limit} \quad \lim_{k\to\infty}\left[1+\frac{1}{(k-1)^2}\right] = 1$$

approximates 1 rapidly as k increases. As shown in Heiser and Bennani (1997, p. 192), if (12.12) holds then the tetrahedral inequality (12.2) does not imply the triangle inequality, but the weaker parametrized triangle inequality

$$d(x_1,x_2) \leq \frac{5}{4}\left[d(x_2,x_3)+d(x_1,x_3)\right].$$

Furthermore, if (12.12) holds, then

$$3d_4(x_{1,4}) \leq d_4(x_{2,5}) + d_4(x_{1,5}^{-2}) + d_4(x_{1,5}^{-3}) + d_4(x_{1,5}^{-4})$$

does not imply the tetrahedral inequality (12.2), but the weaker parametrized inequality

$$2d_3(x_{1,3}) \leq \frac{10}{9}\left[d_3(x_{2,4}) + d_3(x_{1,4}^{-2}) + d_3(x_{1,4}^{-3})\right].$$

12.5 Epilogue

In this chapter a family of k-way metrics that extend the usual two-way metric was studied. The three-way metrics introduced by Joly and Le Calvé (1995) and Heiser and Bennani (1997) and the k-way metrics studied in Deza and Rosenberg (2000) are in the family. The family gives an indication of the many possible extensions for introducing k-way metricity. It was shown how k-way metrics and k-way dissimilarities are related to their $(k-1)$-way counterparts under different set of axioms.

Validity of a metric axiom for $k \geq 3$ appears not to be important for methods used in applied multi-way data analysis, such as multi-way principal component and factor analysis (Kroonenberg, 2008), or multi-way dimensional scaling (Gower and De Rooij, 2003; Heiser and Bennani, 1997). For example, the three-way multidimensional scaling done in Gower and De Rooij (2003) merely required that the underlying two-way coefficients satisfied the triangle inequality, since the three-way dissimilarities are linear transformations of the two-way information. The multi-way procedure based on the gradient method used in Cox, Cox and Branco (1991) and the three-way least squares procedure used in Heiser and Bennani (1997) do not require that the dissimilarities satisfy stronger conditions. At this point the formulations and properties presented in this chapter appear to be of theoretical interest only. From a theoretical point of view it is unfortunate that no well-established basic multi-way metric structure emerged from the study.

CHAPTER 13

Multi-way ultrametrics

Multi-way dissimilarities are natural generalizations of pairwise dissimilarities, that allow global comparison of more than two objects or variables. Various authors have studied three-way dissimilarities and generalized various concepts defined for the two-way case to the three-way case (see, for example, Bennani-Dosse, 1993; Joly and Le Calvé, 1995; Heiser and Bennani, 1997). One of these topics is ultrametric dissimilarities (Diatta and Fichet, 1998; Murtagh, 2004; Diatta, 2007). A two-way dissimilarity $d(x_1, x_2)$ is called a two-way ultrametric if it satisfies the ultrametric inequality, which is given by

$$d(x_1, x_2) \leq \max[d(x_1, x_3), d(x_2, x_3)].$$

The two-way ultrametric inequality implies that the triangle formed by the three points x_1, x_2 and x_3 is isosceles, that is, at least the largest two sides are of equal length. A recent review on where ultrametricity may be encountered is given by Murtagh (2004). Diatta and Fichet (1998) and Diatta (2006, 2007) consider a class of multi-way quasi-ultrametrics that extend the fundamental bijection in classification between ultrametric dissimilarities and indexed hierarchies.

Joly and Le Calvé (1995) and Bennani-Dosse (1993) describe three-way generalizations of the ultrametric inequality, defined for the two-way case. The two different ultrametrics are called weak and strong in Chepoi and Fichet (2007). In this chapter the ideas on three-way ultrametrics presented in Joly and Le Calvé (1995) and Bennani-Dosse (1993) are adopted and extended to multi-way ultrametrics. For the two-way case we have the ultrametric inequality; for the three-way case two equalities have been proposed; for the four-way case three inequalities are presented; and for the multi-way case $(k-1)$ inequalities may be defined. The inspiration for this chapter comes from the thesis by Bennani-Dosse (1993). Some ideas on the three-way ultrametrics presented in that thesis, are explored in this chapter for multi-way dissimilarities.

13.1 Definitions

Let $x_{1,k} = \{x_1, x_2, ..., x_k\}$ be a k-tuple and let $x_{1,k}^{-i}$ be a $(k-1)$-tuple with elements x_1 to x_k where the minus in $x_{1,k}^{-i}$ is used to indicate that element x_i drops out. Both Bennani-Dosse (1993) and Chepoi and Fichet (2007, p. 5) consider two three-way generalizations of the ultrametric inequality, namely

$$\begin{aligned} d(x_{1,3}) &\leq \max\left[d(x_{2,4}), d(x_{1,4}^{-2}), d(x_{1,4}^{-3})\right] \\ d(x_{1,3}) &\leq \max\left[d(x_{2,4}), d(x_{1,4}^{-2})\right]. \end{aligned}$$

These inequalities are called respectively weak and strong ultrametrics in Chepoi and Fichet (2007). For groups of size $k = 4$ it is possible to formulate three ultrametric inequalities. From weak to strong, the three ultrametrics are given by

$$\begin{aligned} d(x_{1,4}) &\leq \max\left[d(x_{2,5}), d(x_{1,5}^{-2}), d(x_{1,5}^{-3}), d(x_{1,5}^{-4})\right] \\ d(x_{1,4}) &\leq \max\left[d(x_{2,5}), d(x_{1,5}^{-2}), d(x_{1,5}^{-3})\right] \\ d(x_{1,4}) &\leq \max\left[d(x_{2,5}), d(x_{1,5}^{-2})\right]. \end{aligned}$$

We may thus formulate $(k-1)$ ultrametrics for a group of k objects.

For the properties in this chapter it is more convenient to define an ultrametric on the number of dissimilarities involved. For example, the inequality $d_3 \leq \max(d_1, d_2)$ represents all metrics of which the definition involves three multi-way dissimilarities, that is,

$$\begin{aligned} d(x_{1,2}) &\leq \max\left[d(x_{2,3}), d(x_{1,3}^{-2})\right] \\ d(x_{1,3}) &\leq \max\left[d(x_{2,4}), d(x_{1,4}^{-2})\right] \\ d(x_{1,4}) &\leq \max\left[d(x_{2,5}), d(x_{1,5}^{-2})\right] \\ d(x_{1,5}) &\leq \max\left[d(x_{2,6}), d(x_{1,6}^{-2})\right] \\ d(x_{1,6}) &\leq \max\left[d(x_{2,7}), d(x_{1,7}^{-2})\right] \\ &\text{etc. ...} \end{aligned}$$

The inequality

$$d_3 \leq \max(d_1, d_2) \tag{13.1}$$

defines the strongest class of ultrametrics, whereas

$$d_4 \leq \max(d_1, d_2, d_3) \tag{13.2}$$

defines the second strongest class. To see that inequality (13.1) defines a stronger ultrametric compared to inequality (13.2), suppose the multi-way dissimilarities are given by

$$d_1 = d_2 = 5 \qquad d_3 = 3 \qquad \text{and} \qquad d_4 = 2.$$

These multi-way dissimilarities satisfy (13.2), since $5 \leq \max(2, 3, 5)$, but not (13.1), because $5 \neq \max(2, 3)$. As a second example, the multi-way dissimilarities given by

$$d_1 = d_2 = 5 \qquad d_3 = 3 \qquad d_4 = 4 \qquad \text{and} \qquad d_5 = 2$$

do not satisfy either (13.1) or (13.2). However, these multi-way dissimilarities do satisfy the weaker ultrametric inequality

$$d_5 \leq \max(d_1, d_2, d_3, d_4) \qquad (\text{for example}, 5 \leq \max(2, 3, 4, 5)).$$

Following this line of reasoning we may conclude that a multi-way ultrametric implies all (possible) weaker ultrametrics.

Proposition 13.1. *Let $d_1, d_2, ..., d_n$ be n multi-way dissimilarities. Then*

$$d_{n-1} \leq \max(d_1, d_2, ..., d_{n-2}) \qquad \Rightarrow \qquad d_n \leq \max(d_1, d_2, ..., d_{n-1}).$$

Let $d_{1,k} = \{d_1, d_2, ..., d_k\}$ be a k-tuple. Then

$$d_{k+1} \leq \max(d_{1,k})$$

defines the weakest class of ultrametrics.

13.2 Strong ultrametrics

The strongest class of ultrametrics is characterized by inequality (13.1). It turns out that, if n multi-way dissimilarities satisfy inequality (13.1), then the $(n-1)$ largest dissimilarities are equal. The sufficiency of this statement is clear from the definition of the class of ultrametrics in inequality (13.1). The proof of necessity goes as follows. We first consider the proof for $n = 3, 4, 5$. The proof for $n = 4$ was already presented in Bennani-Dosse (1993). Furthermore, for $n = 4, 5$ alternative proofs are presented, where the fact is used that the assertion is true for $n - 1$. Finally, the proof is completed by means of induction.

Proposition 13.2. *Let $d_1, d_2, ..., d_n$ be n multi-way dissimilarities. If the n dissimilarities satisfy inequality* (13.1)*, then the largest $n-1$ dissimilarities are equal.*
Proof for $n = 3$: Assume $d_2 \leq d_1$. From $d_3 \leq \max(d_1, d_2)$ we obtain $d_3 \leq d_1$. Then

$$\begin{aligned} d_3 \leq d_2 \text{ and } d_1 \leq \max(d_2, d_3) &\Rightarrow d_1 \leq d_2 \Rightarrow d_3 \leq d_1 = d_2 \\ d_2 \leq d_3 \text{ and } d_1 \leq \max(d_2, d_3) &\Rightarrow d_1 \leq d_3 \Rightarrow d_2 \leq d_1 = d_3. \end{aligned}$$

Proof for $n = 4$: Assume $d_2 \leq d_1$. From $d_3 \leq \max(d_1, d_2)$ we obtain $d_3 \leq d_1$.
First, if $d_3 \leq d_2$

$$\begin{aligned} \text{then } d_1 \leq \max(d_2, d_3) &\Rightarrow d_1 \leq d_2 \Rightarrow d_3 \leq d_1 = d_2 \\ \text{and } d_4 \leq \max(d_2, d_3) &\Rightarrow d_4 \leq d_2. \end{aligned}$$

Then

$$\begin{aligned} d_4 \leq d_3 \text{ and } d_2 \leq \max(d_3, d_4) &\Rightarrow d_2 \leq d_3 \Rightarrow d_4 \leq d_1 = d_2 = d_3 \\ d_3 \leq d_4 \text{ and } d_2 \leq \max(d_3, d_4) &\Rightarrow d_2 \leq d_4 \Rightarrow d_3 \leq d_1 = d_2 = d_4. \end{aligned}$$

Alternatively, if $d_2 \leq d_3$

$$\begin{aligned} \text{then } d_1 \leq \max(d_2, d_3) &\Rightarrow d_1 \leq d_3 \Rightarrow d_2 \leq d_1 = d_3 \\ \text{and } d_4 \leq \max(d_2, d_3) &\Rightarrow d_4 \leq d_3. \end{aligned}$$

Then

$$\begin{aligned} d_4 \leq d_2 \text{ and } d_3 \leq \max(d_2, d_4) &\Rightarrow d_3 \leq d_2 \Rightarrow d_4 \leq d_1 = d_2 = d_3 \\ d_2 \leq d_4 \text{ and } d_3 \leq \max(d_2, d_4) &\Rightarrow d_3 \leq d_4 \Rightarrow d_2 \leq d_1 = d_3 = d_4. \end{aligned}$$

This completes the proof for $n = 4$.
Alternative proof for $n = 4$: Assume that the assertion is true for $n = 3$. If $d_3 \leq d_2 \leq d_1$, then $d_3 \leq d_1 = d_2$ and $d_4 \leq d_2$. Then

$$\begin{aligned} d_4 \leq d_3 \text{ and } d_2 \leq \max(d_3, d_4) &\Rightarrow d_2 \leq d_3 \Rightarrow d_4 \leq d_1 = d_2 = d_3 \\ d_3 \leq d_4 \text{ and } d_2 \leq \max(d_3, d_4) &\Rightarrow d_2 \leq d_4 \Rightarrow d_3 \leq d_1 = d_3 = d_4. \end{aligned}$$

This completes the alternative proof for $n = 4$.

Proof for $n = 5$: Assume $d_2 \leq d_1$. From $d_3 \leq \max(d_1, d_2)$ we obtain $d_3 \leq d_1$.
First, if $d_3 \leq d_2$

$$\begin{aligned} \text{then } d_1 \leq \max(d_2, d_3) &\Rightarrow d_1 \leq d_2 \Rightarrow d_3 \leq d_1 = d_2 \\ \text{and } d_4 \leq \max(d_2, d_3) &\Rightarrow d_4 \leq d_2. \end{aligned}$$

Furthermore, if $d_4 \leq d_3$

$$\begin{aligned} \text{then } d_2 \leq \max(d_3, d_4) &\Rightarrow d_2 \leq d_3 \Rightarrow d_4 \leq d_1 = d_2 = d_3 \\ \text{and } d_5 \leq \max(d_3, d_4) &\Rightarrow d_5 \leq d_3. \end{aligned}$$

Then

$$d_5 \le d_4 \text{ and } d_3 \le \max(d_4, d_5) \;\Rightarrow\; d_3 \le d_4 \;\Rightarrow\; d_5 \le d_1 = d_2 = d_3 = d_4$$
$$d_4 \le d_5 \text{ and } d_3 \le \max(d_4, d_5) \;\Rightarrow\; d_3 \le d_5 \;\Rightarrow\; d_4 \le d_1 = d_2 = d_3 = d_5.$$

Alternatively, if $d_3 \le d_4$

$$\text{then } d_2 \le \max(d_3, d_4) \;\Rightarrow\; d_2 \le d_4 \;\Rightarrow\; d_3 \le d_1 = d_2 = d_4$$
$$\text{and } d_5 \le \max(d_3, d_4) \;\Rightarrow\; d_5 \le d_4.$$

Then

$$d_5 \le d_3 \text{ and } d_4 \le \max(d_3, d_5) \;\Rightarrow\; d_4 \le d_3 \;\Rightarrow\; d_5 \le d_1 = d_2 = d_3 = d_4$$
$$d_3 \le d_5 \text{ and } d_4 \le \max(d_3, d_5) \;\Rightarrow\; d_4 \le d_5 \;\Rightarrow\; d_3 \le d_1 = d_2 = d_4 = d_5.$$

Second, if $d_2 \le d_3$

$$\text{then } d_1 \le \max(d_2, d_3) \;\Rightarrow\; d_1 \le d_3 \;\Rightarrow\; d_2 \le d_1 = d_3$$
$$\text{and } d_4 \le \max(d_2, d_3) \;\Rightarrow\; d_4 \le d_3.$$

Furthermore, if $d_4 \le d_2$

$$\text{then } d_3 \le \max(d_2, d_4) \;\Rightarrow\; d_3 \le d_2 \;\Rightarrow\; d_4 \le d_1 = d_2 = d_3$$
$$\text{and } d_5 \le \max(d_2, d_4) \;\Rightarrow\; d_5 \le d_2.$$

Then

$$d_5 \le d_4 \text{ and } d_2 \le \max(d_4, d_5) \;\Rightarrow\; d_2 \le d_4 \;\Rightarrow\; d_5 \le d_1 = d_2 = d_3 = d_4$$
$$d_4 \le d_5 \text{ and } d_2 \le \max(d_4, d_5) \;\Rightarrow\; d_2 \le d_5 \;\Rightarrow\; d_4 \le d_1 = d_2 = d_3 = d_5.$$

Alternatively, if $d_2 \le d_4$

$$\text{then } d_3 \le \max(d_2, d_4) \;\Rightarrow\; d_3 \le d_4 \;\Rightarrow\; d_2 \le d_1 = d_3 = d_4$$
$$\text{and } d_5 \le \max(d_2, d_4) \;\Rightarrow\; d_5 \le d_4.$$

Then

$$d_5 \le d_2 \text{ and } d_4 \le \max(d_2, d_5) \;\Rightarrow\; d_4 \le d_2 \;\Rightarrow\; d_5 \le d_1 = d_2 = d_3 = d_4$$
$$d_2 \le d_5 \text{ and } d_4 \le \max(d_2, d_5) \;\Rightarrow\; d_4 \le d_5 \;\Rightarrow\; d_2 \le d_1 = d_3 = d_4 = d_5.$$

This completes the proof for $n = 5$.
Alternative proof for $n = 5$: Assume that the assertion is true for $n = 4$. If $d_4 \le d_3 \le d_2 \le d_1$, then $d_4 \le d_1 = d_2 = d_3$ and $d_5 \le d_3$. Then

$$d_5 \le d_4 \text{ and } d_3 \le \max(d_4, d_5) \;\Rightarrow\; d_3 \le d_4 \;\Rightarrow\; d_5 \le d_1 = d_2 = d_3 = d_4$$
$$d_4 \le d_5 \text{ and } d_3 \le \max(d_4, d_5) \;\Rightarrow\; d_3 \le d_5 \;\Rightarrow\; d_4 \le d_1 = d_2 = d_3 = d_5.$$

This completes the alternative proof for $n = 5$.
General proof: Assume that the assertion is true for $n = m$. If $d_m \leq d_{m-1} \leq ... \leq d_2 \leq d_1$, then $d_m \leq d_1 = d_2 = ... = d_{m-2} = d_{m-1}$ and $d_{m+1} \leq d_{m-1}$. Then $d_{m+1} \leq d_m$ and $d_{m-1} \leq \max(d_m, d_{m+1})$ lead to

$$d_{m-1} \leq d_m \;\Rightarrow\; d_{m+1} \leq d_1 = d_2 = ... = d_{m-1} = d_m$$

and $d_m \leq d_{m+1}$ and $d_{m-1} \leq \max(d_m, d_{m+1})$ lead to

$$d_{m-1} \leq d_{m+1} \;\Rightarrow\; d_m \leq d_1 = d_2 = ... = d_{m-1} = d_{m+1}.$$

Hence, the assertion is true for $n = m + 1$. □

13.3 More strong ultrametrics

The second strongest class of ultrametrics is characterized by inequality (13.2). As it turns out, if n multi-way dissimilarities satisfy inequality (13.2), then the $(n-2)$ largest dissimilarities are equal. Similar to Proposition 13.2, sufficiency follows from the definition of ultrametric inequality (13.2). The proof of necessity is slightly more involved compared to the proof of Proposition 13.2. We only consider the proof for $n = 4$ of the assertion, and therefore refer to it as a conjecture.

Conjecture 13.1. *Let $d_1, d_2, ..., d_n$ be n multi-way dissimilarities. If* (13.2) *holds, then the largest $n-2$ dissimilarities are equal.*
Proof for $n = 4$: Assume $d_3 \leq d_4$.
First, if $d_2 \leq d_3$, then from $d_1 \leq \max(d_2, d_3, d_4)$ we obtain $d_1 \leq d_4$. Then

$$d_1 \leq d_3 \text{ and } d_4 \leq \max(d_1, d_2, d_3) \quad \Rightarrow\; d_4 \leq d_3 \;\Rightarrow\; \begin{cases} d_1 \leq d_3 = d_4 \\ d_2 \leq d_3 = d_4 \end{cases}$$

$$d_3 \leq d_1 \text{ and } d_4 \leq \max(d_1, d_2, d_3) \quad \Rightarrow\; d_4 \leq d_1 \;\Rightarrow\; d_2 \leq d_3 \leq d_1 = d_4.$$

Second, assume $d_3 \leq d_2$. If $d_2 \leq d_4$, then from $d_1 \leq \max(d_2, d_3, d_4)$ we obtain $d_1 \leq d_4$. Then

$$d_1 \leq d_3 \text{ and } d_4 \leq \max(d_1, d_2, d_3) \;\Rightarrow\; d_4 \leq d_2 \;\Rightarrow\; d_1 \leq d_3 \leq d_2 = d_4.$$

Alternatively, if $d_3 \leq d_1$, then

$$d_1 \leq d_2 \text{ and } d_4 \leq \max(d_1, d_2, d_3) \quad \Rightarrow\; d_4 \leq d_2 \;\Rightarrow\; d_3 \leq d_1 \leq d_2 = d_4$$

$$d_2 \leq d_1 \text{ and } d_4 \leq \max(d_1, d_2, d_3) \quad \Rightarrow\; d_4 \leq d_1 \;\Rightarrow\; d_3 \leq d_2 \leq d_1 = d_4.$$

Next, if $d_4 \leq d_2$, then

$$d_1 \leq d_3 \text{ and } d_2 \leq \max(d_1, d_3, d_4) \;\Rightarrow\; d_2 \leq d_4 \;\Rightarrow\; d_1 \leq d_3 \leq d_2 = d_4.$$

Alternatively, if $d_3 \leq d_1$, then from $d_1 \leq \max(d_2, d_3, d_4)$ we obtain $d_1 \leq d_2$. Then

$$d_1 \leq d_4 \text{ and } d_2 \leq \max(d_1, d_3, d_4) \quad \Rightarrow\; d_2 \leq d_4 \;\Rightarrow\; d_3 \leq d_1 \leq d_2 = d_4$$

$$d_4 \leq d_1 \text{ and } d_2 \leq \max(d_1, d_3, d_4) \quad \Rightarrow\; d_2 \leq d_1 \;\Rightarrow\; d_3 \leq d_4 \leq d_1 = d_2.$$

This completes the proof for $n = 4$.

13.4 Metrics implied by ultrametrics

In this section we apply the notation used in the first sections of this chapter to multi-way metrics, which were studied in Chapter 12. We are only concerned with the number of dissimilarities involved. For example, the inequality $d_3 \leq d_1 + d_2$ represents all metrics of which the definition involves three multi-way dissimilarities, that is,

$$\begin{aligned} d(x_{1,2}) &\leq d(x_{2,3}) + d(x_{1,3}^{-2}) \\ d(x_{1,3}) &\leq d(x_{2,4}) + d(x_{1,4}^{-2}) \\ d(x_{1,4}) &\leq d(x_{2,5}) + d(x_{1,5}^{-2}) \\ &\text{etc. ...} \end{aligned}$$

Three metric inequalities and two ultrametric inequalities for three-way dissimilarities were considered in Chapter 11. The strong metric $2d_1 \leq d_2 + d_3 + d_4$ introduced by Heiser and Bennani (1997) implies the metric $d_1 \leq d_2 + d_3$, introduced in Joly and Le Calvé (1995). The latter inequality in turn implies the weak metric $d_1 \leq d_2 + d_3 + d_4$. This metric is not considered by the above authors, nor is it considered a metric in Chepoi and Fichet (2007). Furthermore, the strong ultrametric $d_1 \leq \max(d_2, d_3)$ implies the weak ultrametric $d_1 \leq \max(d_2, d_3, d_4)$. The five inequalities are related as follows.

$$\begin{array}{ccc} d_1 \leq \max(d_2, d_3) & \Rightarrow & 2d_1 \leq d_2 + d_3 + d_4 \\ & & \Downarrow \\ \Downarrow & & d_1 \leq d_2 + d_3 \\ & & \Downarrow \\ d_1 \leq \max(d_2, d_3, d_4) & \Rightarrow & d_1 \leq d_2 + d_3 + d_4 \end{array}$$

For the four-way case we may formulate eight inequalities. The inequalities are related as follows.

$$\begin{array}{ccc} d_1 \leq \max(d_2, d_3) & \Rightarrow & 3d_1 \leq d_2 + d_3 + d_4 + d_5 \\ & & \Downarrow \\ & & 2d_1 \leq d_2 + d_3 + d_4 \\ \Downarrow & & \Downarrow \\ & & d_1 \leq d_2 + d_3 \\ & & \Downarrow \\ d_1 \leq \max(d_2, d_3, d_4) & \Rightarrow & d_1 \leq d_2 + d_3 + d_4 \\ \Downarrow & & \Downarrow \\ d_1 \leq \max(d_2, d_3, d_4, d_5) & \Rightarrow & d_1 \leq d_2 + d_3 + d_4 + d_5 \end{array}$$

A variety of properties can immediately be deduced from the above definitions of multi-way ultrametrics and metrics. First of all, the strongest ultrametric inequality for k-way dissimilarities implies the strongest metric inequality for k-way dissimilarities. Remember that, if the strongest k-way ultrametric inequality holds, then the k largest of the $(k+1)$ dissimilarities are equal. With respect to Proposition 13.3 and 13.4, let $d_{1,k} = \{d_1, d_2, ..., d_k\}$ be a k-tuple.

Proposition 13.3. *Let $d_1, d_2, ..., d_k, d_{k+1}$ be $(k+1)$ k-way dissimilarities. Then*

$$d_1 \leq \max(d_{2,k+1}) \qquad \Rightarrow \qquad (k-1)d_1 \leq \sum_{i=2}^{k+1} d_i.$$

Let $d_1, d_2, ..., d_n$ be n k-way dissimilarities ($n \leq k$). All other multi-way ultrametric inequalities, other than the strongest, imply a metric inequality of the form

$$d_1 \leq \sum_{i=2}^{n} d_i.$$

Proposition 13.4. *Let $d_1, d_2, ..., d_n$ be n k-way dissimilarities ($n \leq k$). Then*

$$d_1 \leq \max(d_{2,n}) \qquad \Rightarrow \qquad d_1 \leq \sum_{i=2}^{n} d_i.$$

13.5 Epilogue

Multi-way ultrametrics and some of their properties were the topic of investigation of this chapter. The tetrahedral inequality introduced in Heiser and Bennani (1997) is implied by the strong ultrametric inequality. Suppose we define "interesting" in the sense that a metric inequality is interesting if it is the strongest metric implied by an ultrametric inequality. Then we may say that the tetrahedral inequality (and its multi-way generalization) is more interesting compared to the three-way metric inequality introduced in Joly and Le Calvé (1995).

Some of the ultrametrics and corresponding properties discussed here may find their way into a procedure or algorithm. It is well known that a distance is an ultrametric if and only if it can be represented by a hierarchical tree. Joly and Le Calvé (1995) line out how a hierarchical algorithm may be adopted to the three-way case. First the triple corresponding to the smallest distance is aggregated and the new distances are computed involving this triple as defined in the specific algorithm. The resulting dendrogram has approximately the same properties as in the ordinary two-way case. The only difference is that there will be many levels with three clusters instead of two in the hierarchical tree representation. Applications of three-way ultrametrics and hierarchical trees can be found in Joly and Le Calvé (1995) and Bennani-Dosse (1993).

CHAPTER 14

Perimeter models

Dissimilarities are important tools in many domains of data analysis. Most dissimilarity analysis has however been limited to the two-way case. Multi-way dissimilarities may be used to evaluate complex relationships between three or more objects (see, for example, Diatta, 2006, 2007).

Perimeter models are linear functions that can be used to relate k-way dissimilarities of different degrees k. Their linear form makes perimeter functions simple models with a straightforward interpretation. For example, the three-way perimeter distance is equivalent to the sum of the three two-way distances formed between the three objects. This distance is equivalent to the sum of the three sides of the triangle formed by the three objects. The perimeter distance gives a geometrical interpretation of the concept "average distance" between objects.

The present chapter explores two extensions of the three-way perimeter model. Decompositions and metric properties of both generalizations are investigated. As an extra, the three-way maximum function, together with its multi-way extension and a metric property of the generalization, is studied in the last section.

14.1 Definitions

Let $x_{1,k}$ denote the k-tuple $(x_1, x_2, ..., x_k)$ and let $x_{1,k}^{-i}$ denote the $(k-1)$-tuple $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_k)$ where the minus in the superscript of $x_{1,k}^{-i}$ is used to indicate that element x_i drops out. Let E be a nonempty set of n objects. A dissimilarity $d_k : E^k \to \mathbb{R}_+$ is totally symmetric if for all $x_1, x_2, ..., x_k \in E$ and every permutation π of $\{1, 2, ..., k\}$

$$d_k(x_{\pi(1)}, ..., x_{\pi(k)}) = d_k(x_1, ..., x_k).$$

Furthermore, as a generalization of minimality we define $d_k(x_1, ..., x_1) = 0$.

We define two types of k-way perimeter models. For $k \geq 3$ we define

$$d_k(x_{1,k}) = \frac{1}{p} \sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \tag{14.1}$$

and

$$d_k(x_{1,k}) = \frac{1}{p} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(x_i, x_j) \tag{14.2}$$

where p is a positive real number. Dissimilarity $d_k(x_{1,k})$ in (14.1) is equivalent to the sum of the k dissimilarities $d_{k-1}(x_{1,k}^{-i})$ divided by a factor p. Distance measure $d_n(x_{1,n})$ in (14.2) may be interpreted as the sum of the sides of the polyhedron formed by the k objects in $\{x_1, x_2, ..., x_k\}$, rescaled by a factor p.

Using $k = 3$ in either (14.1) or (14.2) we obtain

$$d_3(x_{1,3}) = \frac{d(x_1, x_2) + d(x_1, x_3) + d(x_2, x_3)}{p}. \tag{14.3}$$

Using $p = 1$ in (14.3) we obtain the three-way perimeter model considered in Heiser and Bennani (1997), De Rooij and Gower (2003), and Chepoi and Fichet (2007). Using $p = 2$ in (14.3) we obtain the three-way semi-perimeter model which is studied in Bennani-Dosse (1993) and Joly and Le Calvé (1995).

Instead of the notation used in (14.3) we will use a shorter, more convenient notation in the next section on decompositions of perimeter models. We write (14.3) as

$$d_{ijl}^{(3)} = \frac{d_{ij} + d_{il} + d_{jl}}{p}. \tag{14.4}$$

Using $k = 4$ in (14.1) and (14.2) we obtain respectively

$$d_{ijlh}^{(4)} = \frac{d_{ijl}^{(3)} + d_{ijh}^{(3)} + d_{ilh}^{(3)} + d_{jlh}^{(3)}}{p} \tag{14.5}$$

and

$$d_{ijlh}^{(4)} = \frac{d_{ij} + d_{il} + d_{ih} + d_{jl} + d_{jh} + d_{lh}}{p}. \tag{14.6}$$

Note that we have expressed (14.4) and (14.5) in the same notation as (14.3).

14.2 Decompositions

The following theorem generalizes a result in Joly and Le Calvé (1995, p. 196), derived for the semi-perimeter model. As it turns out, their result holds for (14.4) and does not depend on the value of p.

Proposition 14.1. *Function* $d^{(3)}_{ijl}$ *satisfies* (14.4) *if and only if*

$$d^{(3)}_{ijl} = \left[d^{(3)}_{ij.} + d^{(3)}_{i.l} + d^{(3)}_{.jl}\right] - \left[d^{(3)}_{i..} + d^{(3)}_{.j.} + d^{(3)}_{..l}\right] + d^{(3)}_{...} \tag{14.7}$$

where

$$\begin{aligned} d^{(3)}_{ij.} &= n^{-1}\sum_{l} d^{(3)}_{ijl} \\ d^{(3)}_{i..} &= n^{-1}\sum_{j} d^{(3)}_{ij.} \\ \text{and} \qquad d^{(3)}_{...} &= n^{-1}\sum_{i} d^{(3)}_{i..}. \end{aligned}$$

Proof: Averaging over l, j, and i in (14.4) we obtain

$$\begin{aligned} pd^{(3)}_{ij.} &= d_{ij} + d_{i.} + d_{j.} \\ pd^{(3)}_{i..} &= 2d_{i.} + d_{..} \\ pd^{(3)}_{...} &= 3d_{..}. \end{aligned}$$

Expressing d_{ij} in terms of $d^{(3)}_{ij.}$, $d^{(3)}_{i..}$, and $d^{(3)}_{...}$, we obtain

$$d_{ij} = pd^{(3)}_{ij.} - \frac{p\left[d^{(3)}_{i..} + d^{(3)}_{.j.}\right]}{2} - \frac{pd^{(3)}_{...}}{3}. \tag{14.8}$$

Using (14.8) in (14.4) we obtain (14.7), which does not depend on p. □

Condition (14.7) for $d^{(3)}_{ijl}$ in (14.4) generalizes naturally to condition (14.9) for $d^{(4)}_{ijlh}$ in (14.5).

Proposition 14.2. *Function* $d^{(4)}_{ijlh}$ *satisfies* (14.5) *if and only if*

$$\begin{aligned} d^{(4)}_{ijlh} = &\left[d^{(4)}_{ijl.} + d^{(4)}_{ij.h} + d^{(4)}_{i.lh} + d^{(4)}_{.jlh}\right] - \left[d^{(4)}_{ij..} + d^{(4)}_{i.l.} + d^{(4)}_{i..h} + d^{(4)}_{.jl.} + d^{(4)}_{.j.h} + d^{(4)}_{..lh}\right] \\ &+ \left[d^{(4)}_{i...} + d^{(4)}_{.j..} + d^{(4)}_{..l.} + d^{(4)}_{...h}\right] - d^{(4)}_{....} \end{aligned} \tag{14.9}$$

where

$$d^{(4)}_{ijl.} = n^{-1}\sum_{l} d^{(4)}_{ijlh}$$
$$d^{(4)}_{ij..} = n^{-1}\sum_{k} d^{(4)}_{ijl.}$$
$$d^{(4)}_{i...} = n^{-1}\sum_{j} d^{(4)}_{ij..}$$
$$\text{and} \qquad d^{(4)}_{....} = n^{-1}\sum_{i} d^{(4)}_{i...}.$$

Proof: Averaging over h, l, j, and i in (14.5) we obtain

$$pd^{(4)}_{ijl.} = d^{(3)}_{ijl} + d^{(3)}_{ij.} + d^{(3)}_{il.} + d^{(3)}_{jl.}$$
$$pd^{(4)}_{ij..} = 2d^{(3)}_{ij.} + d^{(3)}_{i..} + d^{(3)}_{j..}$$
$$pd^{(4)}_{i...} = 3d^{(3)}_{i..} + d^{(3)}_{...}$$
$$pd^{(4)}_{....} = 4d^{(3)}_{...}.$$

Expressing $d^{(3)}_{ijl}$ in terms of $d^{(4)}_{ijl.}$, $d^{(4)}_{ij..}$, $d^{(4)}_{i...}$, and $d^{(4)}_{....}$, we obtain

$$d^{(3)}_{ijl} = pd^{(4)}_{ijl.} - \frac{p\left[d^{(4)}_{ij..} + d^{(4)}_{i.l.} + d^{(4)}_{.jl.}\right]}{2} + \frac{p\left[d^{(4)}_{i...} + d^{(4)}_{.j..} + d^{(4)}_{..l.}\right]}{3} - \frac{pd^{(4)}_{....}}{4}. \tag{14.10}$$

Using (14.10) in (14.5) we obtain (14.9). □

We obtain a different generalization of (14.7) if $d^{(4)}_{ijlh}$ satisfies (14.6).

Proposition 14.3. *Function $d^{(4)}_{ijlh}$ satisfies* (14.6) *if and only if*

$$d^{(4)}_{ijlh} = \left[d^{(4)}_{ij..} + d^{(4)}_{i.l.} + d^{(4)}_{i..h} + d^{(4)}_{.jl.} + d^{(4)}_{.j.h} + d^{(4)}_{..lh}\right] - 2\left[d^{(4)}_{i...} + d^{(4)}_{.j..} + d^{(4)}_{..l.} + d^{(4)}_{...h}\right] + 3d^{(4)}_{....}. \tag{14.11}$$

Proof: Averaging over h, l, j, and i in (14.6) we obtain

$$pd^{(4)}_{ijl.} = d_{ij} + d_{il} + d_{jl} + d_{i.} + d_{j.} + d_{l.}$$
$$pd^{(4)}_{ij..} = d_{ij} + 2d_{i.} + 2d_{j.} + d_{..}$$
$$pd^{(4)}_{i...} = 3d_{i.} + 3d_{..}$$
$$pd^{(4)}_{....} = 6d_{..}.$$

Expressing d_{ij} in terms of $d^{(4)}_{ij..}$, $d^{(4)}_{i...}$, and $d^{(4)}_{....}$, we obtain

$$d_{ij} = pd^{(4)}_{ij..} - \frac{2p\left[d^{(4)}_{i...} + d^{(4)}_{.j..}\right]}{3} + \frac{pd^{(4)}_{....}}{2}. \tag{14.12}$$

Using (14.12) in (14.6) yields (14.11). □

14.3 Metric properties

In this section we study metric properties of perimeter models (14.1) and (14.2). Consider metric inequalities

$$(k-1) \times d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}). \tag{14.13}$$

and

$$(k-2) \times d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}). \tag{14.14}$$

Inequality (14.13) implies inequality (14.14).

Proposition 14.4. *(i) Dissimilarity* $d_n(x_{1,n})$ *in* (14.2) *satisfies* (14.14). *(ii) Dissimilarity* $d_n(x_{1,n})$ *in* (14.2) *satisfies* (14.13) *if and only if* $d(x_i, x_j)$ *satisfies the triangle inequality.*
Proof (i): Using (14.2) in (14.14) we obtain

$$0 \leq (k-1) \sum_{i=1}^{k} d(x_i, x_{k+1})$$

which is true.
Proof (ii): Using (14.2) in (14.13) we obtain

$$\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(x_i, x_j) \leq (k-1) \sum_{i=1}^{k} d(x_i, x_{k+1}). \tag{14.15}$$

Applying (14.15) with the $(k+1)$-tuple $(x_1, x_2, x_3, ..., x_3)$ we obtain $d(x_1, x_2) \leq d(x_2, x_3) + d(x_1, x_3)$.

Conversely, inequality (14.15) follows from adding the k triangle inequalities formed by all pairs in the set $\{x_1, x_2, ..., x_k\}$ and x_{k+1}, for example, $d(x_1, x_2) \leq d(x_2, x_{k+1}) + d(x_1, x_{k+1})$. □

Consider metric inequalities

$$d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}) \tag{14.16}$$

and

$$u \times d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i}) \tag{14.17}$$

where u is a positive real number. Note that inequality (14.16) is implied by (14.13), (14.14) and (14.17).

Proposition 14.5. (*i*) *Dissimilarity* $d_k(x_{1,k})$ *in* (14.1) *satisfies* (14.16). (*ii*) *Dissimilarity* $d_k(x_{1,k})$ *in* (14.1) *satisfies* (14.17) *for* $u > 1$ *if* $d_{k-1}(x_{1,k-1})$ *satisfies*

$$(u-1) \times d_{k-1}(x_{1,k-1}) \leq \sum_{i=1}^{k-1} d_{k-1}(x_{1,k}^{-i}). \tag{14.18}$$

Proof (i): Using (14.1) in (14.16) we obtain

$$0 \leq (k-1) \sum_{i=1}^{k} d(x_i, x_{k+1})$$

which is true.
Proof (ii): Using (14.1) in (14.17) we obtain

$$(u-1) \sum_{i=1}^{k} d_{k-1}(x_{1,k}^{-i}) \leq 2\mathcal{S} \tag{14.19}$$

where $\mathcal{S}$ is the sum of the d_{k-1} dissimilarities that can be formed by all $(k-2)$-tuples in the set $\{x_1, x_2, ..., x_k\}$ and x_{k+1}. Inequality (14.19) follows from adding the k variants of (14.18) that can be formed by using each $(u-1) \times d_{k-1}(x_{1,k}^{-i})$ on the left-hand side of (14.19), on the left-hand side of each polyhedral inequality, and by summing the corresponding k dissimilarities from $\mathcal{S}$ on the right-hand side of the polyhedral inequality. □

14.4 Maximum distance

In the final section of this chapter on perimeter models we explore the multi-way extensions and properties of a somewhat different three-way function. For the three-way case, the maximum distance function is defined as

$$d_3(x_{1,3}) = \max\left[d(x_1, x_2), d(x_1, x_3), d(x_2, x_3)\right] \tag{14.20}$$

by both Heiser and Bennani (1997) and De Rooij and Gower (2003). Function (14.20) has two straightforward four-way generalizations, which are given by

$$d_4(x_{1,4}) = \max\left[d(x_1, x_2), d(x_1, x_3), d(x_1, x_4), d(x_2, x_3), d(x_2, x_4), d(x_3, x_4)\right] \tag{14.21}$$

and

$$d_4(x_{1,4}) = \max\left[d_3(x_{2,4}), d_3(x_{1,4}^{-2}), d_3(x_{1,4}^{-3}), d_3(x_{1,3})\right]$$

where $d_3(x_{1,3})$ is defined as in (14.20). Fortunately, the two formulations are equivalent.

The k-way formulation of (14.21) is given by

$$d_k(x_{1,k}) = \max\left[d(x_1, x_2), d(x_1, x_3), ..., d(x_{k-2}, x_k), d(x_{k-1}, x_k)\right]. \tag{14.22}$$

On the right-hand side of (14.22) we have the maximum dissimilarity that can be constructed from all pairs in the set $\{x_1, x_2, ..., x_k\}$. The multi-way function in (14.22) satisfies inequality (14.13) due to the following result.

Proposition 14.6. *Let $d(x_i, x_j) = 0$ if and only if $x_i = x_j$. Then $d_k(x_{1,k})$ in (14.22) satisfies (14.13) if $d(x_i, x_j)$ satisfies the triangle inequality.*
Proof for $k = 3$: It must be shown that

$$\begin{aligned} 2\max\left[d(x_1,x_2),d(x_1,x_3),d(x_2,x_3)\right] \leq \\ \max\left[d(x_2,x_3),d(x_2,x_4),d(x_3,x_4)\right] + \max\left[d(x_1,x_3),d(x_1,x_4),d(x_3,x_4)\right] + \\ \max\left[d(x_1,x_2),d(x_1,x_4),d(x_2,x_4)\right] \end{aligned} \tag{14.23}$$

holds. The proof is immediate if the maximum of the six dissimilarities is $d(x_i, x_4)$ for $i = 1, 2, 3$. For instance, if $d(x_1, x_4)$ is the largest, then (14.23) becomes

$$\begin{aligned} 2\max\left[d(x_1,x_2),d(x_1,x_3),d(x_2,x_3)\right] \leq & 2d(x_1,x_4) + \\ & \max\left[d(x_2,x_3),d(x_2,x_4),d(x_3,x_4)\right] \end{aligned}$$

which is true, since $d(x_1, x_4) \geq \max\left[d(x_1, x_2), d(x_1, x_3), d(x_2, x_3)\right]$. Furthermore, suppose $d(x_1, x_2)$ is the maximum of the six values. Then (14.23) can be written as

$$\begin{aligned} d(x_1,x_2) \leq & \max\left[d(x_1,x_3),d(x_1,x_4),d(x_3,x_4)\right] + \\ & \max\left[d(x_2,x_3),d(x_2,x_4),d(x_3,x_4)\right]. \end{aligned} \tag{14.24}$$

Inequality (14.24) is true if the triangle inequality holds, which completes the proof for $k = 3$.
Proof for $k = 4$: It must be verified that

$$\begin{aligned} 3\max\left[d(x_1,x_2),d(x_1,x_3),d(x_1,x_4),d(x_2,x_3),d(x_2,x_4),d(x_3,x_4)\right] \leq \\ \max\left[d(x_1,x_2),d(x_1,x_3),d(x_1,x_5),d(x_2,x_3),d(x_2,x_5),d(x_3,x_5)\right] + \\ \max\left[d(x_1,x_2),d(x_1,x_4),d(x_1,x_5),d(x_2,x_4),d(x_2,x_5),d(x_4,x_5)\right] + \\ \max\left[d(x_1,x_3),d(x_1,x_4),d(x_1,x_5),d(x_3,x_4),d(x_3,x_5),d(x_4,x_5)\right] + \\ \max\left[d(x_2,x_3),d(x_2,x_4),d(x_2,x_5),d(x_3,x_4),d(x_3,x_5),d(x_4,x_5)\right]. \end{aligned} \tag{14.25}$$

Again, the proof is immediate if the largest of the ten dissimilarities is $d(x_i, x_5)$ for $i = 1, ..., 4$. Suppose $d(x_1, x_2)$ is the maximum of the ten values. Then (14.25) can be written as

$$\begin{aligned} & d(x_1,x_2) \leq \\ & \max\left[d(x_1,x_3),d(x_1,x_4),d(x_1,x_5),d(x_3,x_4),d(x_3,x_5),d(x_4,x_5)\right] + \\ & \max\left[d(x_2,x_3),d(x_2,x_4),d(x_2,x_5),d(x_3,x_4),d(x_3,x_5),d(x_4,x_5)\right]. \end{aligned} \tag{14.26}$$

Inequality (14.26) is true if the triangle inequality holds, which completes the proof for $k = 4$.
General proof: From the proof for $k = 3$ and $k = 4$, the following pattern becomes apparent. After filling in (14.22) in (14.13), there are $k(k + 1)/2$ different two-way dissimilarities to consider. The proof is immediate if $d(x_i, x_{k+1})$ for $i = 1, 2, ..., k$ is the largest dissimilarity. This part of the proof does not require the triangle inequality. If any of the other dissimilarities is the largest, then (14.22) satisfies (14.13) if the triangle inequality holds. □

14.5 Epilogue

In this chapter multi-way generalizations of two three-way functions, the perimeter distance and the maximum distance, were presented. The extended perimeter distance is based on two-way dissimilarities or on $(k-1)$-way dissimilarities. The resulting multi-way perimeter models are different and posses different properties. We studied decompositions of the perimeter models for ordered tuples, not for tuples with distinct elements. The decomposition of the three-way perimeter model for triples with distinct elements can be found in Chepoi and Fichet (2007), Bennani-Dosse (1993) and Gower and De Rooij (2003). The case has not been studied here, but it may be noted that the decompositions of the two four-way perimeter models defined on tuples with distinct elements, provide similar and interesting formulas.

The maximum function may also be defined on two-way dissimilarities or on $(k-1)$-way dissimilarities; the different definitions are equivalent. Both the generalized perimeter distance and the maximum distance satisfy polyhedral inequality (12.4).

Validity of a multi-way metric inequality for $k \geq 3$ appears not to be important for methods used for multi-way dimensional scaling (Cox, Cox and Branco, 1991; Heiser and Bennani, 1997; Gower and De Rooij, 2003). The results in Section 14.3 therefore appear to be of theoretical interest only. From a theoretical point of view it is unfortunate that no well-established basic multi-way metric structure emerged from the study.

Perimeter models are simple functions with a straightforward interpretation. However, some empirical evidence suggests that using perimeter models is not the best approach to evaluating complex relationships between three or more objects at a time. Gower and De Rooij (2003) used the three-way perimeter model and compared multidimensional scaling of three-way distances to the scaling of two-way distances. These authors concluded that, when the three-way distances were linear transformations of the two-way information, the three-way analysis gained little or nothing over the conventional multidimensional scaling. De Rooij (2001, Chapter 5; 2002) noted that the problem seems to be that definitions of three-way distances in terms of two-way distances do not model true three-way interactions.

CHAPTER 15

Generalizations of Theorem 10.3

For the properties in this section we have a new use of the symbols a, b, c, and d already used for the 2×2 contingency table in Part I. With two-way dissimilarities, a function is called metric if it satisfies, among other things, the triangle inequality. Theorem 10.3 states that which states that if c is a positive constant and the two-way dissimilarity d satisfies the triangle inequality, then the function $d/(c+d)$ satisfies the triangle inequality. In this chapter generalizations of Theorem 10.3 for the triangle inequality are considered.

For the use in this chapter it suffices to define a multi-way metric on the number of dissimilarities involved. Multi-way dissimilarities can be used to measure the resemblance between two or more, say k, objects. Let d_i, $i = 1, 2, ..., n, n+1$ denote $n+1$ multi-way dissimilarities. A generalization of Theorem 10.3 is presented for the inequality

$$d_{n+1} \leq \sum_{i=1}^{n} d_i. \tag{15.1}$$

Furthermore, Conjecture 15.1 below is an attempt to generalize Theorem 10.3 to polyhedral inequality

$$(n-1) \times d_{n+1} \leq \sum_{i=1}^{n} d_i. \tag{15.2}$$

Inequality (15.2) portraits inequality (12.4) and (14.13) in the present simpler notation.

15.1 A generalization of Theorem 10.3.

Proposition 15.1 below is a first attempt to generalize Theorem 10.3, which states that if c is a positive constant and the two-way dissimilarity d satisfies the triangle inequality, then the function $d/(c+d)$ satisfies the triangle inequality. In Proposition 15.1 we consider the multi-way metrics that are characterized by (15.1). We first consider the proofs for $n = 2, 3, 4$. A general proof for Proposition 15.1 is straightforward after considering these proofs.

Proposition 15.1 *If the dissimilarities d_i for $i = 1, 2, ..., n, n+1$ satisfy n-way symmetry, then*

$$\frac{d_{n+1}}{c + d_{n+1}} \leq \sum_{i=1}^{n} \frac{d_i}{c + d_i} \quad \textit{if} \quad d_{n+1} \leq \sum_{i=1}^{n} d_i \quad \textit{holds.}$$

Proof for $n = 2$: It must be shown that the quantity a given by

$$\begin{aligned} a &= (c+d_1)(c+d_2)(c+d_3)\left[\frac{d_1}{c+d_1} + \frac{d_2}{c+d_2} - \frac{d_3}{c+d_3}\right] \\ &= d_1(c+d_2)(c+d_3) + d_2(c+d_1)(c+d_3) - d_3(c+d_1)(c+d_2) \\ &= c^2(d_1 + d_2 - d_3) + 2cd_1d_2 + d_1d_2d_3. \end{aligned}$$

is positive. Since $d_3 \leq d_1 + d_2$ under the conditions of the assertion, the quantity a is positive, which completes the proof for $n = 2$.
Proof for $n = 3$: It must be shown that the quantity a given by

$$\begin{aligned} a =& (c+d_1)(c+d_2)(c+d_3)(c+d_4)\left[\frac{d_1}{c+d_1} + \frac{d_2}{c+d_2} + \frac{d_3}{c+d_3} - \frac{d_4}{c+d_4}\right] \\ =& d_1(c+d_2)(c+d_3)(c+d_4) + \\ & d_2(c+d_1)(c+d_3)(c+d_4) + \\ & d_3(c+d_1)(c+d_2)(c+d_4) - d_4(c+d_1)(c+d_2)(c+d_3) \end{aligned}$$

is positive. Expanding the equation in polynomial form we obtain

$$\begin{aligned} a =& c^3(d_1 + d_2 + d_3 - d_4) + 2c^2(d_1d_2 + 2d_1d_3 + 2d_2d_3) + \\ & c(3d_1d_2d_3 + d_1d_2d_4 + d_1d_3d_4 + d_2d_3d_4) + 2d_1d_2d_3d_4. \end{aligned}$$

Only the coefficient of c^3 needs to be checked since all other coefficients are positive. The coefficient of c^3 is positive if $d_4 \leq d_1 + d_2 + d_3$ (the condition of the assertion). This completes the proof for $n = 3$.

Proof for $n = 4$: It must be shown that the quantity a given by

$$\begin{aligned} a =& \prod_{i=1}^{5}(c+d_i)\left[\sum_{i=1}^{4}\frac{d_i}{c+d_i} - \frac{d_5}{c+d_5}\right] \\ =& d_1(c+d_2)(c+d_3)(c+d_4)(c+d_5)+ \\ & d_2(c+d_1)(c+d_3)(c+d_4)(c+d_5)+ \\ & d_3(c+d_1)(c+d_2)(c+d_4)(c+d_5)+ \\ & d_4(c+d_1)(c+d_2)(c+d_3)(c+d_5) - d_5(c+d_1)(c+d_2)(c+d_3)(c+d_4) \end{aligned}$$

is positive. Expanding the equation in polynomial form we obtain

$$\begin{aligned} a =& c^4(d_1+d_2+d_3+d_4-d_4) \\ & + 2c^3(d_1d_2+d_1d_3+d_1d_4+d_2d_3+d_2d_4+d_3d_4) \\ & + 3c^2(d_1d_2d_3+d_1d_2d_4+d_1d_3d_4+d_2d_3d_4) \\ & + 2c^2d_5(d_1d_2+d_1d_3+d_1d_4+d_2d_3+d_2d_4+d_3d_4) \\ & + 4cd_1d_2d_3d_4 + 2d_5(d_1d_2d_3+d_1d_2d_4+d_1d_3d_4+d_2d_3d_4) \\ & + 3d_1d_2d_3d_4d_5. \end{aligned}$$

Only the coefficient of c^4 needs to be checked since all other coefficients are positive. The coefficient of c^4 is positive under the conditions of the assertion. This completes the proof for $n = 4$.
Outline general proof: It must be shown that the quantity

$$a = \prod_{i=1}^{n+1}(c+d_i) \times \left[\sum_{i=1}^{n}\frac{d_i}{c+d_i} - \frac{d_{n+1}}{c+d_{n+1}}\right]$$

is positive. After expanding the equation in polynomial form only the coefficient of c^n needs to be checked. This coefficient is positive under the conditions of the assertion. □

Conjecture 15.1 in Section 15.3 is a (potentially) stronger result compared to Proposition 15.1. With Conjecture 15.1 we attempt to prove Proposition 15.1 not for inequality (15.1), but for inequality (15.2). Before presenting this attempt, the next section is first used to present some auxiliary results.

15.2 Auxiliary results

We first repeat Proposition 12.1 in Proposition 15.2, using the more convenient notation.

Proposition 15.2. *If the dissimilarities d_i for $i = 1, 2, ..., n, n+1$ satisfy n-way symmetry, then (for $n \geq 3$) (15.2) implies*

$$(n-2)d_n \leq \sum_{i=1}^{n-1} d_i.$$

Proof: Interchanging the roles of d_n and d_{n+1} and dividing by $n-1$ in (15.2), we may obtain the inequalities

$$(n-1)d_n \leq d_{n+1} + \sum_{i=1}^{n-1} d_i$$

and

$$d_{n+1} \leq \left[\frac{1}{n-1}\right] \sum_{i=1}^{n} d_i.$$

Adding the two inequalities and multiplying by $(n-1)/n$ gives the required inequality. □

The inequality in Proposition 15.4 below concerns one of the inequalities required in Conjecture 15.1 below. First, we present a stronger result, which is then used in the proof of Proposition 15.4.

Proposition 15.3. *Dissimilarities d_i for $i = 1, 2, ..., n, n+1$ satisfy*

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n+1} d_i d_j \geq \left[\frac{n^2 - n - 1}{2(n-1)}\right] \sum_{i=1}^{n+1} d_i^2$$

if (15.2) *holds.*
Proof: Inequality (15.2) can be written as

$$d_1 \geq (n-1)d_{n+1} - \sum_{i=2}^{n} d_i. \tag{15.3}$$

Squaring both sides of (15.3) we obtain

$$d_1^2 \geq (n-1)^2 d_{n+1}^2 + \sum_{i=2}^{n} d_i^2 - 2(n-1)d_{n+1} \sum_{i=2}^{n} d_i + 2 \sum_{i=2}^{n} \sum_{j=i+1}^{n+1} d_i d_j \tag{15.4}$$

(for $n = 2$ the last term of the inequality equals zero).

There are $(n+1)$ variants of d_i^2 in (15.4) and $n(n+1)/2$ variants of $d_i d_j$. The number of variants of the inequality is given by the smallest common multiple of $(n+1)$ and $n(n+1)/2$. Instead, consider the multiple $n(n+1)^2/2$. Adding up all $n(n+1)^2/2$ variants of (15.4) we obtain

$$\begin{aligned}\frac{n(n+1)}{2}\sum_{i=1}^{n+1} d_i^2 \geq{}& \frac{(n-1)^2 n(n+1)}{2}\sum_{i=1}^{n+1} d_i^2 \\ &+ \frac{(n-1)n(n+1)}{2}\sum_{i=1}^{n+1} d_i^2 \\ &- 2(n-1)^2(n+1)\sum_{i=1}^{n}\sum_{j=i+1}^{n+1} d_i d_j \\ &+ (n-1)(n-2)(n+1)\sum_{i=1}^{n}\sum_{j=i+1}^{n+1} d_i d_j\end{aligned}$$

which equals the required inequality. This completes the proof. □

The inequality in Proposition 15.4 is one of the inequalities required in Conjecture 15.1 in Section 15.3. The proof of this inequality makes use of the stronger result in Proposition 15.3.

Proposition 15.4. *Dissimilarities d_i for $i = 1, 2, ..., n, n+1$ satisfy*

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_i d_j \geq \left[\frac{n-2}{2}\right] d_{n+1}\sum_{i=1}^{n} d_i$$

if (15.2) *holds.*
Proof: Using the equality

$$\left[\sum_{i=1}^{n} d_i\right]^2 - \sum_{i=1}^{n} d_i^2 = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_i d_j$$

the quantity a given by

$$a = 2(n-1)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_i d_j - (n-1)(n-2)d_{n+1}\sum_{i=1}^{n} d_i$$

can be written as $a = b_1 + b_2$, where

$$b_1 = (n-2)\left[\sum_{i=1}^{n} d_i\right]\left[\sum_{i=1}^{n} d_i - (n-1)d_{n+1}\right]$$

and

$$b_2 = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_i d_j - (n-2)\sum_{i=1}^{n} d_i^2.$$

The assertion follows if it can be shown that the quantity a is positive. Under the condition of the proposition (the last term of) quantity b_1 is positive. Furthermore, in the light of Proposition 15.3, quantity b_2 is positive since

$$\frac{n^2-n-1}{2(n-1)} \geq \frac{n-2}{2}.$$

Hence quantity a is positive, which completes the proof. □

15.3 A stronger generalization of Theorem 10.3

Conjecture 15.1 below is an attempt to generalize Theorem 10.3, which states that if c is a positive constant and the two-way dissimilarity d satisfies the triangle inequality, then the function $d/(c+d)$ satisfies the triangle inequality. Below, proofs for small n are presented, but no proof is offered for any n. With respect to Conjecture 15.1, it is assumed that the multi-way dissimilarities satisfy n-way symmetry, which makes the use of Proposition 15.2 possible. Note that also for $n = 2$, Theorem 10.3 is a special case of Conjecture 15.1.

Conjecture 15.1 *If the dissimilarities d_i for $i = 1, 2, ..., n, n+1$ satisfy n-way symmetry, then*

$$\frac{(n-1)d_{n+1}}{c+d_{n+1}} \leq \sum_{i=1}^{n} \frac{d_i}{c+d_i}$$

if (15.2) *holds.*
Proof for $n = 2$: It must be shown that the quantity a given by

$$\begin{aligned} a &= (c+d_1)(c+d_2)(c+d_3)\left[\frac{d_1}{c+d_1} + \frac{d_2}{c+d_2} - \frac{d_3}{c+d_3}\right] \\ &= d_1(c+d_2)(c+d_3) + d_2(c+d_1)(c+d_3) - d_3(c+d_1)(c+d_2) \\ &= c^2(d_1+d_2-d_3) + 2cd_1d_2 + d_1d_2d_3 \end{aligned}$$

is positive. Since $d_3 \leq d_1 + d_2$ by Proposition 15.2, the quantity a is positive, which completes the proof for $n = 2$.

Proof for $n = 3$: It must be shown that the quantity a given by

$$\begin{aligned} a =&(c+d_1)(c+d_2)(c+d_3)(c+d_4)\left[\frac{d_1}{c+d_1}+\frac{d_2}{c+d_2}+\frac{d_3}{c+d_3}-\frac{2d_4}{c+d_4}\right] \\ =&d_1(c+d_2)(c+d_3)(c+d_4)+ \\ &d_2(c+d_1)(c+d_3)(c+d_4)+ \\ &d_3(c+d_1)(c+d_2)(c+d_4)-2d_4(c+d_1)(c+d_2)(c+d_3) \end{aligned}$$

is positive. Expanding the equation in polynomial form we obtain

$$\begin{aligned} a =&c^3(d_1+d_2+d_3-2d_4)+ \\ &c^2(2d_1d_2+2d_1d_3+2d_2d_3-d_1d_4-d_2d_4-d_3d_4)+ \\ &3cd_1d_2d_3+d_1d_2d_3d_4. \end{aligned}$$

The coefficient of c^3 is positive if $2d_4 \leq d_1+d_2+d_3$. The coefficient of c^2 is positive if $d_3 \leq d_1+d_2$, since it can be written as

$$d_1(d_2+d_3-d_4)+d_2(d_1+d_3-d_4)+d_3(d_1+d_2-d_4).$$

Thus, the quantity a is positive by Proposition 15.2, which completes the proof for $n = 3$.

Proof for $n = 4$: It must be shown that the quantity a given by

$$\begin{aligned} a =&\prod_{i=1}^{5}(c+d_i)\left[\sum_{i=1}^{4}\frac{d_i}{c+d_i}-\frac{3d_5}{c+d_5}\right] \\ =&d_1(c+d_2)(c+d_3)(c+d_4)(c+d_5)+ \\ &d_2(c+d_1)(c+d_3)(c+d_4)(c+d_5)+ \\ &d_3(c+d_1)(c+d_2)(c+d_4)(c+d_5)+ \\ &d_4(c+d_1)(c+d_2)(c+d_3)(c+d_5)-3d_5(c+d_1)(c+d_2)(c+d_3)(c+d_4) \end{aligned}$$

is positive. Expanding the equation in polynomial form we obtain

$$\begin{aligned} a =\ & c^4(d_1+d_2+d_3+d_4-3d_4) \\ &+2c^3(d_1d_2+d_1d_3+d_1d_4+d_2d_3+d_2d_4+d_3d_4-d_1d_5-d_2d_5-d_3d_4-d_4d_5) \\ &+3c^2(d_1d_2d_3+d_1d_2d_4+d_1d_3d_4+d_2d_3d_4) \\ &-c^2d_5(d_1d_2+d_1d_3+d_1d_4+d_2d_3+d_2d_4+d_3d_4) \\ &+4cd_1d_2d_3d_4+d_1d_2d_3d_4d_5. \end{aligned}$$

The coefficient of c^4 is positive if $3d_5 \leq d_1+d_2+d_3+d_4$. The coefficient of c^3 is positive if $2d_4 \leq d_1+d_2+d_3$, since it can be written as

$$\begin{aligned} &d_1(d_2+d_3+d_4-2d_5)+d_2(d_1+d_3+d_4-2d_5)+ \\ &d_3(d_1+d_2+d_4-2d_5)+d_4(d_1+d_2+d_3-3d_5). \end{aligned}$$

Alternatively, the coefficient of c^3 is positive by Proposition 15.4.

The coefficient of c^2 is positive if $d_3 \leq d_1 + d_2$, since it can be written as

$$d_1d_2(d_3 + d_4 - d_5) + d_1d_3(d_2 + d_4 - d_5) + d_1d_4(d_2 + d_3 - d_5) +$$
$$d_2d_3(d_1 + d_4 - d_5) + d_2d_4(d_1 + d_3 - d_5) + d_3d_4(d_1 + d_2 - d_5).$$

Thus, the quantity a is positive by Proposition 15.2, which completes the proof for $n = 4$.

Proof for $n = 5$: It must be shown that the quantity a given by

$$a = \prod_{i=1}^{6}(c + d_i)\left[\sum_{i=1}^{5}\frac{d_i}{c + d_i} - \frac{4d_6}{c + d_6}\right]$$

is positive. Quantity a can be written as

$$\begin{aligned} a = {} & d_1(c + d_2)(c + d_3)(c + d_4)(c + d_5)(c + d_6) \\ & + d_2(c + d_1)(c + d_3)(c + d_4)(c + d_5)(c + d_6) \\ & + d_3(c + d_1)(c + d_2)(c + d_4)(c + d_5)(c + d_6) \\ & + d_4(c + d_1)(c + d_2)(c + d_3)(c + d_5)(c + d_5) \\ & + d_5(c + d_1)(c + d_2)(c + d_3)(c + d_4)(c + d_5) \\ & - 4d_6(c + d_1)(c + d_2)(c + d_3)(c + d_4)(c + d_5). \end{aligned}$$

Expanding the equation in polynomial form we obtain

$$\begin{aligned} a = {} & c^5\left[\sum_{i=1}^{5} d_i - 4d_6\right] \\ & + c^4\left[2\sum_{i=1}^{4}\sum_{j=i+1}^{5} d_id_j - 3d_6\sum_{i=1}^{5} d_i\right] \\ & + c^3\left[3\sum_{i=1}^{3}\sum_{j=i+1}^{4}\sum_{r=j+1}^{5} d_id_jd_r - 2d_6\sum_{i=1}^{4}\sum_{j=i+1}^{5} d_id_j\right] \\ & + c^2\left[4\sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{r=j+1}^{4}\sum_{s=r+1}^{5} d_id_jd_rd_s - d_6\sum_{i=1}^{3}\sum_{j=i+1}^{4}\sum_{l=j+1}^{5} d_id_jd_l\right] \\ & + 5c\prod_{i=1}^{5} d_i + \prod_{i=1}^{6} d_i. \end{aligned}$$

The coefficient of c^5 is positive if $4d_6 \leq \sum_{i=1}^{5} d_i$. The coefficient of c^4 is positive if $3d_5 \leq \sum_{i=1}^{4} d_i$, since it can be written as

$$d_1(d_2 + d_3 + d_4 + d_5 - 3d_6) + d_2(d_1 + d_3 + d_4 + d_5 - 3d_6) +$$
$$d_3(d_1 + d_2 + d_4 + d_5 - 3d_6) + d_4(d_1 + d_2 + d_3 + d_5 - 3d_6) +$$
$$d_5(d_1 + d_2 + d_3 + d_4 - 3d_6).$$

Alternatively, the coefficient of c^4 is positive by Proposition 15.4.

The coefficient of c^3 is positive if $2d_4 \leq \sum_{i=1}^{3} d_i$, since it can be written as

$$\begin{aligned}
&d_1d_2(d_3+d_4+d_5-2d_6)+d_1d_3(d_2+d_4+d_5-2d_6)+\\
&d_1d_4(d_2+d_3+d_5-2d_6)+d_1d_5(d_2+d_3+d_4-2d_6)+\\
&d_2d_3(d_1+d_4+d_5-2d_6)+d_2d_4(d_1+d_3+d_5-2d_6)+\\
&d_2d_5(d_1+d_3+d_4-2d_6)+d_3d_4(d_1+d_2+d_5-2d_6)+\\
&d_3d_5(d_1+d_2+d_4-2d_6)+d_4d_5(d_1+d_2+d_3-2d_6).
\end{aligned}$$

The coefficient of c^2 is positive if $d_3 \leq d_1 + d_2$, since it can be written as

$$\begin{aligned}
&d_1d_2d_3(d_4+d_5-d_6)+d_1d_2d_4(d_3+d_5-d_6)+\\
&d_1d_2d_5(d_3+d_4-d_6)+d_1d_3d_4(d_2+d_5-d_6)+\\
&d_1d_3d_5(d_2+d_4-d_6)+d_1d_4d_5(d_2+d_3-d_6)+\\
&d_2d_3d_4(d_1+d_5-d_6)+d_2d_3d_5(d_1+d_4-d_6)+\\
&d_2d_4d_5(d_1+d_3-d_6)+d_3d_4d_5(d_1+d_2-d_6).
\end{aligned}$$

Hence, a is positive, which completes the proof $n = 5$.
Outline general proof: It must be shown that the quantity

$$a = \prod_{i=1}^{n+1}(c+d_i) \times \left[\sum_{i=1}^{n} \frac{d_i}{c+d_i} - \frac{(n-1)d_{n+1}}{c+d_{n+1}}\right]$$

is positive. Due to Proposition 15.2 each metric inequality also implies all weaker metric inequalities. The quantity a can be written as a polynomial function of $c^n, c^{n-1}, ..., c^2, c$ and a constant $\prod_{i=1}^{n+1} d_i$. The coefficient belonging to the linear part c and the constant $\prod_{i=1}^{n+1} d_i$ are always positive. It must be shown that the remaining $(n-1)$ coefficients are also positive. The coefficient corresponding to c^n appears to be positive if the metric inequality $(n-1)d_{n+1} \leq \sum_{i=1}^{n} d_i$ holds.

15.4 Epilogue

Theorem 10.3, which states that if two-way dissimilarity d satisfies the triangle inequality, then so does the function $d/(c+d)$, was generalized to the multi-way case in this chapter. In the first generalization, Proposition 15.1, multi-way metrics were considered that are characterized by inequality $d_{n+1} \leq \sum_{i=1}^{n} d_i$. In the second attempt, Conjecture 15.1, we tried to proof the generalization for the stronger class of multi-way metrics characterized by $(n-1)d_{n+1} \leq \sum_{i=1}^{n} d_i$. The proof of Proposition 15.1 turned out to be straightforward, especially in contrast to the proof of Conjecture 15.1.

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

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

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

$$S^{(k)}_{\text{BB}} = \frac{a^{(k)}}{\max(p_1, p_2, ..., p_k)} \qquad \text{(minimum)}$$

$$S^{(k)}_{\text{Kul}} = \frac{1}{k} \sum_{i=1}^{k} \frac{a^{(k)}}{p_i} \qquad \text{(arithmetic mean)}$$

$$S^{(k)}_{\text{DK}} = \frac{a^{(k)}}{\prod_{i=1}^{k} p_i^{1/k}} \qquad \text{(geometric mean)}$$

$$S^{(k)}_{\text{Sim}} = \frac{a^{(k)}}{\min(p_1, p_2, ..., p_k)} \qquad \text{(maximum)}.$$

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

$$S_{\text{Phi}} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$
$$S_{\text{Cohen}} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}$$
$$S_{\text{Loe}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)}$$
$$S_{\text{Yule1}} = \frac{ad - bc}{ad + bc}.$$

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 $S^{(k)*}_{\text{SM}}$, $S^{(k)**}_{\text{Gleas}}$, and $S^{(k)}_{\text{Cohen}}$ 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

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

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 $E(a) = p_1p_2$ 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 $P\left(x_{1,k}^{1}\right)$ denote the proportion of 1s in variables x_1 to x_k. Furthermore, let $P\left(x_{1,i,k}^{1,0,1}\right)$ denote the proportion of 1s in variables x_1 to x_k and 0 in variable x_i. Moreover, denote by $P\left(x_{1,i,k}^{1,-,1}\right)$ 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

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

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

$$S_{\text{RR}}^{(k)} = a^{(k)} = P\left(x_{1,k}^{1}\right).$$

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

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

The quantity $2/k(k-1)$ in the definition of $S_{\text{RR}}^{(k)*}$ is used to ensure that $0 \leq S_{\text{RR}}^{(k)*} \leq 1$. 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 $1 - S_{\text{RR}}^{(k)}$ 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*

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

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

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

which equals

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

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

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

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

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

In Proposition 18.6 the metric property of $1-S^{(k)}_{\mathrm{GL1}}(\theta)$ is considered. In order to proof the assertion, the result in Proposition 18.5 on $1 - S^{(k)}_{\mathrm{Jac}}$ 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 $1 - S^{(k)}_{\mathrm{GL1}}(\theta)$.

Proposition 18.6. *The function*

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

satisfies (18.1) *for* $0 < \theta \leq 1$.

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

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

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

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

Because $1 - S^{(k)}_{\mathrm{Jac}}$ 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.

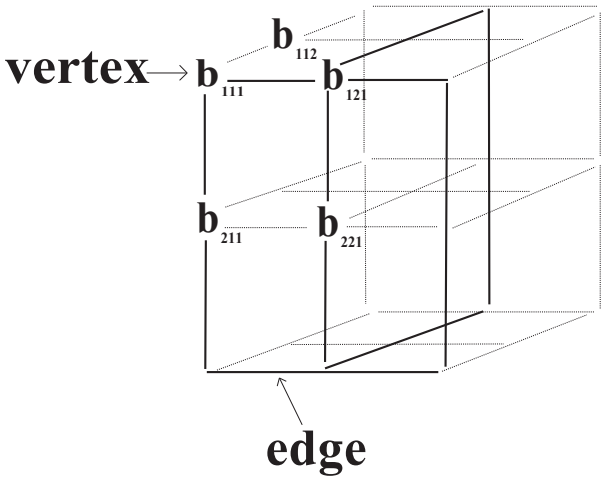

Figure 19.1: *Several aspects of a cube.*

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.

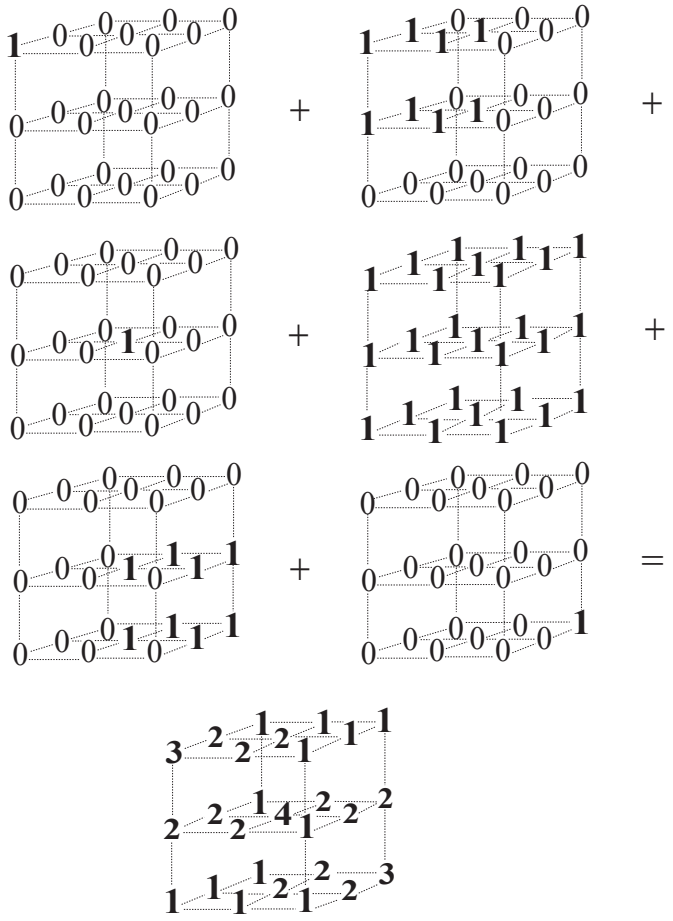

Figure 19.2: *The sum of the six regular Robinson cubes is a regular Robinson cube.*

References

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, *23*, 301-313.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Baroni-Urbani, C., & Buser, M. W. (1976). Similarity of binary data. *Systematic Zoology*, *25*, 251-259.

Barthélemy, J.-P., Brucker, F., & Osswald, C. (2004). Combinatorial optimization and hierarchical classifications. *4OR*, *2*, 179-219.

Batagelj, V., & 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.

Benini, R. (1901). *Principii di Demografie. no. 29 of Manuali Barbèra di Science Giuridiche Sociali e Poltiche.* Firenzi: G. Barbèra.

Bennani-Dosse, M. (1993). *Analyses Métriques á Trois Voies.* Unpublished doctoral dissertation, Université de Haute Bretagne Rennes II, France.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories and Mental Test Scores.* Reading: Addinson-Wesley.

Blackman, N. J. M., & Koval, J. J. (1993). Estimating rater agreement in 2×2 tables: Correction for chance and intraclass correlation. *Applied Psychological Measurement*, *17*, 211-223.

Bloch, D. A., & Kraemer, H. C. (1989). 2×2 Kappa coefficients: Measures of agreement or association. *Biometrics*, *45*, 269-287.

Boorman, S. A., & Arabie, P. (1972). Structural measures and the method of sorting. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, vol. 1: Theory* (p. 225-249). New York: Seminar Press.

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

Bray, J. R. (1956). A study of mutual occurrence of plant species. *Ecology*, *37*, 21-28.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs*, *27*, 325-349.

Brennan, R. L., & Light, R. J. (1974). Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, *27*, 154-163.

Brito, P. (1991). *Analyse de Données Symboliques: Pyramides d'Heritage.* Unpublished doctoral dissertation, Université Paris 9, Paris, France.

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

Buneman, P. (1974). A note on metric properties of trees. *Journal of Combinatorial Theory, Series B*, *17*, 48-50.

Burt, C. (1948). The factorial study of temperamental traits. *British Journal of Psychology (Statistical Section)*, *1*, 178-203.

Cain, A. J., & Harrison, G. A. (1958). An analysis of the taxonomist's judgment of affinity. *Proceedings of Zoological Society London*, *131*, 85-98.

Cheetham, A. H., & Hazel, J. E. (1969). Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, *43*, 1130-1136.

Chepoi, C., & Fichet, B. (2007). A note on three-way dissimilarities and their relationship with two-way dissimilarities. In P. Brito, P. Bertrand, G. Cucumel, & F. de Caravalho (Eds.), *Selected Contributions in Data Analysis and Classification* (p. 465-476). Berlin: Springer.

Chepoi, V., & Fichet, B. (1997). Recognition of Robinsonian dissimilarities. *Journal of Classification*, *14*, 311-325.

Cheung, K. C., & Mooi, L. C. (1994). A comparison between the rating scale model and dual scaling for Likert scales. *Applied Psychological Measurement*, *18*, 1-13.

Clement, P. W. (1976). A formula for computing inter-observer agreement. *Psychological Reports*, *39*, 257-258.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *32*, 113-120.

Cole, L. C. (1949). The measurement of interspecific association. *Ecology*, *30*, 411-424.

Coombs, C. H. (1964). *A Theory of Data.* New York: Wiley.

Cox, T. F., & Cox, M. A. A. (2000). A general weighted two-way dissimilarity coefficient. *Journal of Classification*, *17*, 101-121.

Cox, T. F., Cox, M. A. A., & Branco, J. A. (1991). Multidimensional scaling of n-tuples. *British Journal of Mathematical and Statistical Psychology*, *44*, 195-206.

Critchley, F. (1994). On exchangeability-based equivalence relations induced by strongly Robinson and, in particular, by quadripolar Robinson dissimilarity matrices. In B. Van Cutsem (Ed.), *Classification and Dissimilarity Analysis, Lecture Notes in Statistics.* New York: Springer-Verlag.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cureton, E. E. (1959). Note on $\phi/\phi_{\max}$. *Psychometrika*, *24*, 89-91.

Czekanowski, J. (1932). Coefficient of racial likeliness und Durchschnittliche Differenz. *Anthropologischer Anzeiger*, *9*, 227-249.

Davenport, E. C., & El-Sanhurry, N. A. (1991). Phi/phimax: Review and synthesis. *Educational and Psychological Measurement*, *51*, 821-828.

De Gruijter, D. N. M., & Van der Kamp, L. J. T. (2008). *Statistical Test Theory for the Behavioral Sciences.* New York: Chapman & Hall.

De Gruijter, D. N. M. (1984). Homogeneity analysis of test score data: A confrontation with the latent trait approach. *Applied Psychological Measurement*, *8*, 385-390.

De Rooij, M. (2001). *Distance Models for Transition Frequency Data.* Doctoral dissertation: Leiden University, Leiden, The Netherlands.

De Rooij, M. (2002). Distance models for three-way tables and three-way association. *Journal of Classification*, *19*, 161-178.

De Rooij, M., & Gower, J. C. (2003). The geometry of triadic distances. *Journal of Classification*, *20*, 181-220.

Deza, M.-M., & Rosenberg, I. G. (2000). n-Semimetrics. *European Journal of Combinatorics*, *Special Issue Discrete Metric Spaces 21-6*, 797-806.

Deza, M.-M., & Rosenberg, I. G. (2005). Small cones of m-hemimetrics. *Discrete Mathematics*, *291*, 81-97.

Diatta, J. (2006). Description-meet compatible multiway dissimilarities. *Discrete Applied Mathematics*, *154*, 493-507.

Diatta, J. (2007). Galois closed entity sets and k-balls of quasi-ultrametric multi-way dissimilarities. *Advances in Data Analysis and Classification*, *1*, 53-65.

Diatta, J., & Fichet, B. (1998). Quasi-ultrametrics and their 2-ball hypergraphs. *Discrete Mathematics*, *192*, 87-102.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297-302.

Diday, E. (1984). *Une Représentation Visuelle des Classes Empiétantes: Les Pyramides.* INRIA: Research report 291.

Diday, E. (1986). Orders and overlapping clusters in pyramids. In J. de Leeuw, W. J. Heiser, J. J. Meulman, & F. Critchley (Eds.), *Multidimensional Data Analysis* (p. 201-234). Leiden: DSWO Press.

Diday, E., & Bertrand, P. (1986). An extension of hierarchical clustering: The pyramidal representation. In E. Gelsema & L. Kanal (Eds.), *Pattern Recognition in Practice II* (p. 411-424). Amsterdam: North-Holland.

Digby, P. G. N. (1983). Approximating the tetrachoric correlation coefficient. *Biometrics*, *39*, 753-757.

Doolittle, M. H. (1885). The verification of predictions. *Bulletin of the Philosophical Society of Washington*, *7*, 122-127.

Driver, H. E., & Kroeber, A. L. (1932). Quantitative expression of cultural relationship. *The University of California Publications in American Archaeology and Ethnology*, *31*, 211-256.

Fager, E. W., & McGowan, J. A. (1963). Zooplankton species groups in the North Pacific. *Science*, *140*, 453-460.

Farkas, G. M. (1978). Correction for bias present in a method of calculating inter-observer agreement. *Journal of Applied Behavior Analysis*, *11*, 188.

Fichet, B. (1984). Sur une extension de la notion de hiérarchie et son équivalence avec quelques matrices de Robinson. *Actes des "Journées de Statistique de la Grande Motte"*, 12-12.

Fichet, B. (1986). Distances and Euclidean distances for presence-absence characters and their application to factor analysis. In J. de Leeuw, W. J. Heiser, J. J. Meulman, & F. Critchley (Eds.), *Multidimensional Data Analysis* (p. 23-46). Leiden: DSWO Press.

Fisher, R. A. (1922). On the interpretation of the χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*, 87-94.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*, 651-659.

Forbes, S. A. (1907). On the local distribution of certain Illinois fishes: An essay in statistical ecology. *Bulletin of the Illinois State Laboratory for Natural History*, *7*, 273-303.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*, 553-569.

Gantmacher, F. R. (1977). *Matrix Theory* (Vol. II). New York: Chelsea.

Gantmacher, F. R., & Krein, M. G. (1950). *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems.* Washington: translation from Russian, issued (1961), AEC-tr-4481, by US Atomic Energy Commision.

Gaul, W., & Schader, M. (1994). Pyramidal classification based on incomplete dissimilarity data. *Journal of Classification*, *11*, 171-193.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* Chicester: Wiley.

Gleason, H. A. (1920). Some applications of the quadrat method. *Bulletin of the Torrey Botanical Club*, *47*, 21-33.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732-764.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*, 325-338.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857-874.

Gower, J. C. (1986). Euclidean distance matrices. In J. de Leeuw, W. J. Heiser, J. J. Meulman, & F. Critchley (Eds.), *Multidimensional Data Analysis* (p. 11-22). Leiden: DSWO Press.

Gower, J. C. (1990). Fisher's optimal scores and multiple correspondence analysis. *Biometrics*, *46*, 947-961.

Gower, J. C., & De Rooij, M. (2003). A comparison of the multidimensional scaling of triadic and dyadic distances. *Journal of Classification*, *20*, 115-136.

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

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* New York: Academic Press.

Guilford, J. P. (1965). The minimal phi coefficient and the maximal phi. *Educational and Psychological Measurement*, *25*, 3-8.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The Prediction of Personal Adjustment.* New York: SSRC.

Guttman, L. (1950). The principal components of scale analysis. In S. A. Stouffer (Ed.), *Measurement and Prediction* (p. 312-361). Princeton: Princeton University Press.

Guttman, L. (1954). The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), *Mathematical Thinking in the Social Sciences* (p. 216-257). Glencoe: Free Press.

Hamann, U. (1961). Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen. *Willdenowia*, *2*, 639-768.

Harris, F. C., & Lahey, B. B. (1978). A method for combining occurrence and nonoccurrence agreement scores. *Journal of Applied Behavioral Analysis*, *11*, 523-527.

Hawkins, R. P., & Dotson, V. A. (1968). Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of interobserver agreement scores in interval coding. In E. Ramp & G. Semb (Eds.), *Behavior Analysis: Areas of Research and Application* (p. 539-376). Englewood Cliffs, N.J.: Prentice-Hall.

Heiser, W. J. (1981). *Unfolding Analysis of Proximity Data.* Unpublished doctoral dissertation, Leiden University, Leiden, The Netherlands.

Heiser, W. J., & Bennani, M. (1997). Triadic distance models: Axiomatization and least squares representation. *Journal of Mathematical Psychology*, *41*, 189-206.

Heuvelmans, A. P. J. M., & Sanders, P. F. (1993). Beoordelaarsovereenstemming. In T. J. H. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de Praktijk* (p. 443-470). Arnhem: Cito Instituut voor Toetsontwikkeling.

Holley, J. W., & Guilford, J. P. (1964). A note on the $G-$index of agreement. *Educational and Psychological Measurement*, *24*, 749-753.

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

Hubert, L. J. (1977). Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, *30*, 98-103.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193-218.

Jaccard, P. (1912). The distribution of the flora in the Alpine zone. *The New Phytologist*, *11*, 37-50.

Janson, S., & Vegelius, J. (1979). On the generalization of the $G-$index and the phi coefficient to nominal scales. *Multivariate Behavioral Research*, *14*, 255-269.

Janson, S., & Vegelius, J. (1981). Measures of ecological association. *Oecologia*, *49*, 371-376.

Janson, S., & Vegelius, J. (1982). The $J-$index as a measure of nominal scale response agreement. *Applied Psychological Measurement*, *6*, 111-121.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241-254.

Joly, S., & Le Calvé, G. (1995). Three-way distances. *Journal of Classification*, *12*, 191-205.

Karlin, S. (1968). *Total Positivity I.* Stanford: Stanford Univeristy Press.

Kendall, D. G. (1971). Seriation from abundance matrices. In F. R. Hodson, D. G. Kendall, & P. Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences* (p. 215-252). Edinburgh: University Press.

Kent, R. N., & Foster, S. L. (1977). Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of Behavioral Assessment* (p. 279-328). New York: John Wiley & Sons.

Krippendorff, K. (1987). Association, agreement, and equity. *Quality and Quantity*, *21*, 109-123.

Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis.* Hoboken, New Jersey: Wiley.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151-160.

Kulczynski, 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.

Lerman, I. C. (1988). Comparing partitions (mathematical and statistical aspects). In H. H. Bock (Ed.), *Classification and Related Methods of Data Analysis* (p. 121-131). North-Holland: Elsevier Science Publishers B.V.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*, 365-377.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychometrika*, *Monograph No. 4*.

Loevinger, J. A. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, *45*, 507-530.

Lord, F. M. (1952). A theory of mental test scores. *Psychometrika*, *Monograph No. 7*.

Lord, F. M. (1958). Some relations between Guttman's principal components analysis and other psychometric tests. *Psychometrika*, *36*, 109-133.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading: Addinson-Wesley.

Mak, T. K. (1988). Analysing intraclass correlation for dichotomous variables. *Applied Statistics*, *37*, 344-352.

Maxwell, A. E., & 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.

McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika*, *48*, 377-391.

Meulman, J. (1982). *Homogeneity Analysis.* Leiden: DSWO Press.

Michael, E. L. (1920). Marine ecology and the coefficient of association. *Journal of Animal Ecology*, *8*, 54-59.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis.* The Hague, The Netherlands: Mouton.

Montgomery, A. C., & Crittenden, K. S. (1977). Improving coding reliability for open-ended questions. *Public Opinion Quarterly*, *41*, 235-243.

Morey, L. C., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, *44*, 33-37.

Mountford, M. D. (1962). An index of similarity and its applications to classificatory problems. In P. W. Murphy (Ed.), *Progress in Soil Zoology* (p. 43-50). London: Butterworths.

Murtagh, F. (2004). On ultrametricity, data coding, and computation. *Journal of Classification*, *21*, 167-184.

Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*, 5269-5273.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications.* Toronto: University of Toronto Press.

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.

Odum, E. P. (1950). Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology*, *31*, 587-605.

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika*, *34*, 139-167.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika*, *9*, 105-117.

Pearson, K., & Heron, D. (1913). On theories of association. *Biometrika*, *9*, 159-315.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, *4*, 453-454.

Popping, R. (1983a). *Overeenstemmingsmaten voor Nominale Data.* Unpublished doctoral dissertation, Rijksuniversiteit Groningen, Groningen, The Netherlands.

Popping, R. (1983b). Traces of agreement. On the dot-product as a coefficient of agreement. *Quality and Quantity*, *17*, 1-18.

Popping, R. (1984). Traces of agreement. On some agreement indices for open-ended questions. *Quality and Quantity*, *18*, 147-158.

Post, W. J., & Snijders, T. A. B. (1993). Nonparametric unfolding models for dichotomous data. *Methodika*, *7*, 130-156.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846-850.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* New York: Wiley.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests. Studies in Mathematical Psychology.* Copenhagen: Danish Institute for Educational Research.

Restle, F. (1959). A metric and an ordering on sets. *Psychometrika*, *24*, 207-220.

Robinson, W. S. (1951). A method for chronologically ordering archaeological deposits. *American Antiquity*, *16*, 293-301.

Rogers, D. J., & Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, *132*, 1115-1118.

Rogot, E., & Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Disease*, *19*, 991-1006.

Russel, P. F., & 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.

Schouten, H. J. A. (1980). Measuring pairwise agreement among many observers. *Biometrical Journal*, *22*, 497-504.

Schriever, B. F. (1986). Multiple correspondence analysis and ordered latent structure models. *Kwantitatieve Methoden*, *21*, 117-131.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321-325.

Sepkoski, J. J. (1974). Quantified coefficients of association and measurement of similarity. *Mathematical Geology*, *6*, 135-152.

Serlin, R. C., & Kaiser, H. F. (1978). A method for increasing the reliability of a short multiple-choice test. *Educational and Psychological Measurement*, *38*, 337-340.

Sibson, R. (1972). Order invariant methods for data analysis. *Journal of the Royal Statistical Society, Series B*, *34*, 311-349.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory.* Thousand Oaks: Sage.

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science*, *241*, 1-31.

Sneath, P. H. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, *17*, 201-226.

Snijders, T. A. B., Dormaar, M., Van Schuur, W. H., Dijkman-Caes, C., & Driessen, G. (1990). Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes. *Journal of Classification*, *7*, 5-31.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409-1438.

Sokal, R. R., & Sneath, P. H. (1963). *Principles of Numerical Taxonomy.* San Francisco: W. H. Freeman and Company.

Sørenson, T. (1948). A method of stabilizing groups of equivalent amplitude in plant sociology based on the similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab Biologiske Skrifter*, *5*, 1-34.

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

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, *9*, 386-396.

Stiles, H. E. (1961). The association factor in information retrieval. *Journal of the Association for Computing Machinery*, *8*, 271-279.

Thissen, D., Chen, W. H., & Bock, D. (2003). *Multilog 7: Analysis of multiple-category response data.* Scientific Software International.

Torgerson, W. S. (1958). *Theory and Methods of Scaling.* New York: Wiley.

Tucker, L. R. (1951). *A Method for Synthesis of Factor Analysis Studies.* Personnel research section report No. 984. Washington, D.C.: Department of the Army.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* Berlin, Germany: Springer.

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

Wallace, D. L. (1983). A method for comparing two hierarchical clusterings: Comment. *Journal of the American Statistical Association*, *78*, 569-576.

Warrens, M. J., De Gruijter, D. N. M., & Heiser, W. J. (2007). A systematic comparison between classical optimal scaling and the two-parameter IRT model. *Applied Psychological Measurement*, *31*, 106-120.

Warrens, M. J., & Heiser, W. J. (2006). Scaling unidimensional models with multiple correspondence analysis. In M. J. Greenacre & J. Blasius (Eds.), *Multiple Correspondence Analysis and Related Methods* (p. 219-235). Boca Raton: Chapman & Hall.

Warrens, M. J., Heiser, W. J., & De Gruijter, D. N. M. (2006). Reparametrization of homogeneity analysis to accommodate item response functions. *Behaviormetrika*, *32*, 127-139.

Wilkinson, E. M. (1971). Archaeological seriation and the traveling salesman problem. In F. R. Hodson, D. G. Kendall, & P. Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences* (p. 276-283). Edinburgh: University Press.

Williams, W. T., Lambert, J. M., & Lance, G. N. (1966). Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. *Journal of Ecology*, *54*, 427-445.

Yamada, F., & Nishisato, S. (1993). Several mathematical properties of dual scaling as applied to dichotomous item-category data. *Japanese Journal of Behaviormetrics*, *20*, 56-63.

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society, A*, *75*, 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.

Yule, G. U., & Kendall, M. G. (1950). *An Introduction to the Theory of Statistics.* London: Charles Griffin and Co. Ltd.

Zegers, F. E. (1986). *A General Family of Association Coefficients.* Unpublished doctoral dissertation, Rijksuniversiteit Groningen, Groningen, The Netherlands.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika*, *50*, 17-24.

Zysno, P. V. (1997). The modification of the phi-coefficient reducing its dependence on the marginal distributions. *Methods of Psychological Research Online*, *2*, 41-52.

List of similarity coefficients

In this appendix we present a list of the two-way coefficients for binary data that one may find in the literature. The coefficients are ordered on year of appearance.

Peirce (1884):

$$S_{\text{Peir1}} = \frac{ad - bc}{p_1 q_1} \quad \text{and} \quad S_{\text{Peir2}} = \frac{ad - bc}{p_2 q_2}$$

Doolittle (1885), Pearson (1926):

$$S_{\text{Doo}} = \frac{(ad - bc)^2}{p_1 p_2 q_1 q_2}$$

Yule (1900), Montgomery and Crittenden (1977):

$$S_{\text{Yule1}} = \frac{ad - bc}{ad + bc}$$

Pearson (1905) (quoted by Yule and Kendall, 1950):

$$\text{Chi-square} \qquad \chi^2 = \frac{n(ad - bc)^2}{p_1 p_2 q_1 q_2}$$

Forbes (1907):

$$S_{\text{Forbes}} = \frac{na}{p_1 p_2}$$

Jaccard (1912):

$$S_{\text{Jac}} = \frac{a}{a + b + c}$$

Yule (1912), Pearson and Heron (1913):

$$\text{phi coefficient} \qquad S_{\text{Phi}} = \frac{ad - bc}{\sqrt{p_1 p_2 q_1 q_2}}$$

Yule (1912):

$$S_{\text{Yule2}} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Gleason (1920), Dice (1945), Sørenson (1948), Nei and Li (1979):

$$S_{\text{Gleas}} = \frac{2a}{p_1 + p_2}$$

Michael (1920):

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

Kulczyński (1927), Driver and Kroeber (1932):

$$S_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{p_1} + \frac{a}{p_2}\right) \quad \text{and} \quad S_{\text{Kul2}} = \frac{a}{b+c}$$

Braun-Blanquet (1932):

$$S_{\text{BB}} = \frac{a}{\max(p_1, p_2)}$$

Driver and Kroeber (1932), Ochiai (1957), Fowlkes and Mallows (1983):

$$S_{\text{DK}} = \frac{a}{\sqrt{p_1 p_2}}$$

Kuder and Richardson (1937), Cronbach (1951) for two binary variables:

$$S_{\text{KR}} = \frac{4(ad - bc)}{p_1 q_1 + p_2 q_2 + 2(ad - bc)}$$

Russel and Rao (1940):

$$S_{\text{RR}} = \frac{a}{a + b + c + d}$$

Simpson (1943):

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

Dice (1945), Wallace (1983), Post and Snijders (1993):

$$S_{\text{Dice1}} = \frac{a}{p_1} \quad \text{and} \quad S_{\text{Dice2}} = \frac{a}{p_2}$$

Loevinger (1947, 1948), Mokken (1971), Sijtsma and Molenaar (2002):

$$S_{\text{Loe}} = \frac{ad - bc}{\min(p_1 q_2, p_2 q_1)}$$

Cole (1949):

$$S_{\text{Cole1}} = \frac{ad - bc}{p_1 q_2} \quad \text{and} \quad S_{\text{Cole2}} = \frac{ad - bc}{p_2 q_1}$$

Goodman and Kruskal (1954):

$$S_{\text{GK}} = \frac{2\min(a, d) - b - c}{2\min(a, d) + b + c}$$

Scott (1955):

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

Sokal and Michener (1958), Rand (1971), Brennan and Light (1974):

$$\text{Simple matching coefficient} \qquad S_{\text{SM}} = \frac{a+d}{a+b+c+d}$$

Sorgenfrei (1958), Cheetham and Hazel (1969):

$$\text{Correlation ratio} \quad S_{\text{Sorg}} = \frac{a^2}{p_1 p_1}$$

Cohen (1960):

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

Rogers and Tanimoto (1960), Farkas (1978):

$$S_{\text{RT}} = \frac{a+d}{a+2(b+c)+d}$$

Stiles (1961):

$$S_{\text{Sti}} = \log_{10} \frac{n\left(|ad-bc| - \frac{n}{2}\right)^2}{p_1 p_2 q_1 q_2}$$

Hamann (1961), Holley and Guilford (1964), Hubert (1977):

$$S_{\text{Ham}} = \frac{a-b-c+d}{a+b+c+d}$$

Mountford (1962):

$$S_{\text{Mount}} = \frac{2a}{a(b+c)+2bc}$$

Fager and McGowan (1963):

$$S_{\text{FM}} = \frac{a}{\sqrt{p_1 p_2}} - \frac{1}{2\sqrt{\max(p_1, p_2)}}$$

Sokal and Sneath (1963):

$$S_{\text{SS1}} = \frac{a}{a+2(b+c)} \qquad S_{\text{SS2}} = \frac{2(a+d)}{2a+b+c+2d}$$

$$S_{\text{SS3}} = \frac{1}{4}\left(\frac{a}{p_1} + \frac{a}{p_2} + \frac{d}{q_1} + \frac{d}{q_2}\right) \qquad S_{\text{SS4}} = \frac{ad}{\sqrt{p_1 p_2 q_1 q_2}}$$

$$\text{and} \quad S_{\text{SS5}} = \frac{a+d}{b+c}$$

McConnaughey (1964):

$$S_{\text{McC}} = \frac{a^2 - bc}{p_1 p_2}$$

Rogot and Goldberg (1966):

$$S_{\text{RG}} = \frac{a}{p_1 + p_2} + \frac{d}{q_1 + q_2}$$

Johnson (1967):

$$S_{\text{John}} = \frac{a}{p_1} + \frac{a}{p_2}$$

Hawkins and Dotson (1968):

$$S_{\mathrm{HD}} = \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right)$$

Maxwell and Pilliner (1968):

$$S_{\mathrm{MP}} = \frac{2(ad-bc)}{p_1q_1 + p_2q_2}$$

Fleiss (1975):

$$S_{\mathrm{Fleiss}} = \frac{(ad-bc)[p_1q_2 + p_2q_1]}{2p_1p_2q_1q_2}$$

Clement (1976):

$$S_{\mathrm{Clem}} = \frac{aq_1}{p_1} + \frac{dp_1}{q_1}$$

Baroni-Urabani and Buser (1976):

$$S_{\mathrm{BUB}} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}} \quad \text{and} \quad S_{\mathrm{BUB2}} = \frac{a-b-c+\sqrt{ad}}{a+b+c+\sqrt{ad}}$$

Kent and Foster (1977):

$$S_{\mathrm{KF1}} = \frac{-bc}{bp_1 + cp_2 + bc} \quad \text{and} \quad S_{\mathrm{KF2}} = \frac{-bc}{bq_1 + cq_2 + bc}$$

Harris and Lahey (1978):

$$S_{\mathrm{HL}} = \frac{a(q_1+q_2)}{2(a+b+c)} + \frac{d(p_1+p_2)}{2(b+c+d)}$$

Digby (1983):

$$S_{\mathrm{Digby}} = \frac{(ad)^{3/4} - (bc)^{3/4}}{(ad)^{3/4} + (bc)^{3/4}}$$

Some coefficients for which no source was found in the literature:

$$\frac{2a-b-c}{2a+b+c}, \quad \frac{2d}{b+c+2d}, \quad \frac{2d-b-c}{b+c+2d}$$

$$\frac{4ad}{4ad + (a+d)(b+c)}$$ which is the harmonic mean of $\frac{a}{p_1}$, $\frac{a}{p_2}$, $\frac{d}{q_1}$ and $\frac{d}{q_2}$

$$\frac{ad-bc}{\min(p_1p_2, q_1q_2)}$$ for which its minimum value of -1 is tenable.

Summary of coefficient properties

For some of the vast amount of similarity coefficients in the appendix entitled "List of similarity coefficients", several mathematical properties were studied in this thesis. Seven coefficients stand out in the sense that for these coefficients multiple attractive properties were established in this thesis. A practical conclusion is that in most data-analytic applications the choice for the right coefficient for binary variables can probably be limited to the following seven coefficients.

Source	Jaccard (1912)
Formula	$S_{\text{Jac}} = a/(a+b+c)$
Properties	– Value indeterminate if $d = 1$ – Member of parameter family $S_{\text{GL1}} = a/[a+\theta(b+c)]$; members are interchangeable with respect to an ordinal comparison – Bounded below by correlation ratio $S_{\text{Sorg}} = a^2/p_1p_2$ – Bounded above by $S_{\text{BB}} = a/\max(p_1, p_2)$ – $D_{\text{Jac}} = 1 - S_{\text{Jac}}$ satisfies the triangle inequality – Coefficient matrix is a Robinson matrix if $\mathbf{X}$ is double Petrie – A multivariate generalization satisfies a strong generalization of the triangle inequality

Source	Gleason (1920), Dice (1945), Sørenson (1948), Bray (1956), Bray and Curtis (1957), Nei and Li (1979)
Formula	$S_{\text{Gleas}} = 2a/(p_1 + p_2)$
Properties	– Value indeterminate if $d = 1$ – Member of parameter family $S_{\text{GL1}} = a/[a + \theta(b + c)]$; members are interchangeable with respect to an ordinal comparison – Special case of a coefficient by Czekanowski (1932) – Bounded below by $S_{\text{BB}} = a/\max(p_1, p_2)$ – Bounded above by $S_{\text{DK}} = a/\sqrt{p_1 p_2}$ – Becomes S_{Cohen} after correction for chance using $E(a + d) = p_1 p_2 + q_1 q_2$ – Coefficient matrix is a Robinson matrix if $\mathbf{X}$ is double Petrie – Three straightforward multivariate generalizations

Source	Braun-Blanquet (1932)
Formula	$S_{\text{BB}} = a/\max(p_1, p_2)$
Properties	– Value indeterminate if $d = 1$ – Special case of a coefficient by Robinson (1951) – Bounded below by $S_{\text{Jac}} = a/(a + b + c)$ – Bounded above by $S_{\text{Gleas}} = 2a/(p_1 + p_2)$ – Coefficient matrix is a Robinson matrix if $\mathbf{X}$ is double Petrie – Coefficient matrix is a Robinson matrix with a monotonic stochastic model – First eigenvector of coefficient matrix reflects a stochastic model

Source	Russel-Rao (1940)
Formula	$S_{\mathrm{RR}} = a/(a+b+c+d)$
Properties	– No indeterminate values – $D_{\mathrm{RR}} = 1 - S_{\mathrm{RR}}$ satisfies the triangle inequality – Coefficient matrix is a Robinson matrix if $\mathbf{X}$ is row Petrie – Coefficient matrix is totally positive of order 2 if $\mathbf{X}$ is double Petrie – First eigenvector of coefficient matrix reflects an ordering of a stochastic model – Two multivariate generalizations satisfy a strong generalization of the triangle inequality

Source	Loevinger (1947, 1948)
Formula	$S_{\mathrm{Loe}} = (ad - bc)/\min(p_1q_2, p_2q_1)$
Properties	– $S_{\mathrm{Loe}} = [a - E(a)]/[a_{\max} - E(a)]$ with $E(a) = p_1p_2$ and $a_{\max} = \min(p_1, p_2)$ – Coefficient $S_{\mathrm{Sim}} = a/\min(p_1, p_2)$ becomes S_{Loe} after correction for chance using $E(a) = p_1p_2$ – Various coefficients, including S_{Cohen} and S_{Phi}, become S_{Loe}, after correction for maximum value – Coefficients that are linear in $(a+d)$ become S_{Loe} after correction for chance using $E(a+d) = p_1p_2 + q_1q_2$ and correction for maximum value; the result is irrespective of what correction is applied first

Source	Sokal and Michener (1958)
Formula	$S_{\text{SM}} = (a+d)/(a+b+c+d)$ "Simple matching coefficient"
Properties	– No indeterminate values – Is a special case of proportion of agreement for two nominal variables – Is equivalent to coefficients by Rand (1971) and Brennan and Light (1974) – Member of parameter family $S_{\text{GL2}} = (a+d)/[a+\theta(b+c)+d]$; members are interchangeable with respect to an ordinal comparison – Becomes S_{Cohen} after correction for chance using $E(a+d) = p_1p_2 + q_1q_2$ – $D_{\text{SM}} = 1 - S_{\text{SM}}$ satisfies the triangle inequality – Two multivariate generalizations satisfy a strong generalization of the triangle inequality

Source	Cohen (1960)
Formula	$S_{\text{Cohen}} = 2(ad-bc)/(p_1q_2 + p_2q_1)$
Properties	– S_{Cohen} is a special case of Cohen's kappa for two nominal variables – Bounded below by $S_{\text{Scott}} = (4ad-(b+c)^2)/(p_1+p_2)(q_1+q_2)$ – A variety of coefficients that are linear in $(a+d)$, like S_{SM} and S_{Gleas}, become S_{Cohen} after correction for chance using $E(a+d) = p_1p_2 + q_1q_2$ – Is equivalent to the Adjusted Rand index by Hubert and Arabie (1985)

Coefficient index

Author index

Summary in Dutch (Samenvatting)

We spreken van binaire of dichotome data als er sprake is van een reeks getallen die slechts twee waardes aannemen. De twee waardes kunnen gezien worden als twee, elkaar uitsluitende, categorieën die voor het gemak als 1 en 0 kunnen worden gecodeerd. Een binaire reeks, bijvoorbeeld $\{0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0\}$, kan verkregen worden door van een aantal personen het geslacht vast te stellen, waarbij bijvoorbeeld, 1=vrouw en 0=man. Een binaire reeks kan ook verkregen worden door voor een persoon te coderen welke vragen hij of zij goed of fout had op een toets. Duidelijk is dat binaire reeksen simpel te verkrijgen zijn door allerlei tweeledigheden te verzamelen: goed/fout, voor/tegen, wel/niet, nieuw/oud, ja/nee, aanwezig/niet aanwezig, of PSV-fan/geen kampioen worden.

Als er twee of meer binaire reeksen beschikbaar zijn kan het interessant zijn om te weten in hoeverre de twee reeksen op elkaar lijken. Een bioloog die voor twee gebieden heeft gecodeerd welke diersoorten er wel of niet leven, bijvoorbeeld $\{0, 1, 1, 0\}$ en $\{1, 0, 1, 0\}$, kan zich afvragen in hoeverre de twee gebieden (reeksen) op elkaar lijken. Om twee reeksen te kunnen vergelijken moeten de posities van de reeksen wel dezelfde diersoorten weergeven. De eerste reeks geeft bijvoorbeeld aan dat in het eerste gebied geen vogels, wel paarden, wel muizen, maar geen schildpadden leven; de tweede reeks geeft aan dat in het tweede gebied wel vogels, geen paarden, wel muizen, en geen schildpadden leven. Twee reeksen kunnen nu vergeleken worden met elkaar door na te gaan hoeveel 1n of 0n ze gemeenschappelijk hebben in dezelfde posities. In de biologie zijn twee leefomgevingen meer in overeenstemming naarmate er meer diersoorten in beide gebieden aanwezig zijn (het is niet gebruikelijk om overeenstemming te definiëren in termen van afwezigheid).

Essentieel bij het bestuderen van binaire reeksen is het uitgangspunt dat alle informatie in twee reeksen van gelijke lengte uitputtend kan worden samengevat in vier getallen: $a = \#(1, 1)$ = het aantal posities dat beide reeksen een 1 hebben $(1, 1)$, $d = \#(0, 0)$ = het aantal posities dat beide reeksen een 0 hebben $(0, 0)$, en $b = \#(1, 0)$ en $c = \#(0, 1)$ = de aantallen posities dat er een 1 staat in de ene reeks en een nul in de andere reeks. Willen we de overeenstemming van twee binaire reeksen quantificeren dan kan dat met behulp van overeenstemmingsmaten

of gelijkheidscoëfficiënten. Een overeenstemmingsmaat drukt de gelijkheid van twee binaire reeksen uit in een getal. Voor de vergelijking van twee binaire reeksen is door de jaren heen echter een groot aantal maten voorgesteld. Ondanks het groot aantal verschillende overeenstemmingsmaten, zijn ze allemaal een of andere functie van de getallen of variabelen a, b, c, en d. Een voorbeeld is de Jaccard coefficient met de formule $a/(a+b+c)$.

Omdat het niet altijd duidelijk is wat nu in welke situatie de meeste geschikte coëfficiënt of associatiemaat is, is het nuttig om de coëfficiënten en hun eigenschappen te bestuderen. Dit kan op een veelvoud van manieren, maar in dit proefschrift is gekozen voor een mathematische bestudering van de coëfficiënten en hun eigenschappen. Kortweg wordt hiermee bedoeld dat het niet echt uitmaakt welke waardes de getallen a, b, c, en d aannemen, maar dat het alleen van uitmaakt hoe a, b, c, en d zich (in een formule) tot elkaar verhouden. Eigenlijk worden in het gehele proefschrift verschillende combinaties (formules), allemaal functies van a, b, c, en d, met elkaar vergeleken.

Dit proefschrift bestaat uit negentien hoofdstukken verdeeld in vier delen. In deel I worden steeds eigenschappen van coëfficiënten bestudeerd waar alleen de individuele formule voor nodig is. In dit deel worden er slechts twee binaire reeksen tegelijk met elkaar vergeleken. In deel II en IV worden twee benaderingen besproken voor het geval dat we meer dan twee reeksen tegelijk beschouwen. In deel II worden niet individuele coëfficiënten maar matrices van coëfficiënten bestudeerd. Een matrix wordt verkregen door, bij meer dan twee binaire reeksen, tussen alle paren van reeksen de associatiemaat te bepalen. Deze coëfficiënten kunnen dan worden weergegeven in een coëfficiëntmatrix. In deel IV van dit proefschrift worden coëfficiënten gedefiniëerd die de mate van associatie of overeenstemming reflecteren van twee of meerdere binaire reeksen tegelijk. Voordat de meerweg coëfficiënten in deel IV worden behandeld, wordt deel III gebruikt om een aantal meerweg concepten te definiëren en te bestuderen.

Deel I bestaat uit vijf hoofdstukken. Notatie en enkele basisconcepten van overeenstemmingsmaten worden geïntroduceerd in Hoofdstuk 1. Hoofdstuk 2 stelt de overeenstemmingsmaten voor binaire data in een breder perspectief. De formules die in dit proefschrift worden behandeld zijn in veel gevallen een speciaal geval van een formule die geschikt is voor algemenere data dan binaire gegevens. In dit hoofdstuk wordt aangetoond dat men de Hubert-Arabie adjusted Rand index kan uitrekenen door eerst de 2×2 tabel te formeren door het aantal objectparen te tellen dat in hetzelfde cluster is geplaatst door beide methodes, dat in een cluster is geplaatst door een methode maar in verschillende clusters door de andere methode, en het aantal objectparen te tellen dat in verschillende clusters door beide methodes is geplaatst, en vervolgens Cohen's kappa uit te rekenen voor deze 2×2 tabel. Hoofdstuk 3 laat zien dat een aantal coëfficiënten behoren tot families van coëfficiënten. Het bestuderen van families in plaats van individuele coëfficiënten geeft ons vaak algemenere inzichten en resultaten. Een hoge waarde van een coëfficiënt kan ook komen door toeval. In hoofdstuk 4 worden coëfficiënten en correctie voor toeval bestudeerd en wordt, bijvoorbeeld, aangetoond dat de simple matching coëfficiënt, Cohen's kappa,

Goodman en Kruskal's lambda, Scott's pi, Hamann's eta, en overeenstemmingsmaten geintroduceerd door Gleason/Dice/Sørenson en Rogot en Goldberg, equivalent worden na correctie voor toeval, ongeacht de verwachte waarde die gebruikt wordt.

De maximale waarde van een coëfficiënt gegeven de marginale distributies (totaal aantal 1n van de ene en andere binaire reeks) wordt bestudeerd in hoofstuk 5. Voor sommige coëfficiënten is de maximale waarde niet onder alle omstandigheden gelijk aan 1. De formule van deze coëfficiënten wordt een andere formule na correctie voor de maximale waarde. Iedere overeenstemmingsmaat voor binaire data, waarvan de teller gelijk is aan de covariantie en de noemer een functie is van de marginale distributies, wordt gelijk aan de Loevinger coëfficiënt na correctie voor maximale waarde gegeven de marginale dsitributies.

Deel II bestaat uit vijf hoofdstukken. Hoofdstuk 6 beschrijft een aantal manieren waarop de 1n en 0n van twee of meer binaire reeksen aan elkaar gerelateerd kunnen zijn. De modellen en data structuren die hier beschreven worden, dienen in hoofdstukken 7 en 8 als voldoende voorwaarden voor bepaalde matrices van coëfficiënten om zekere eigenschappen te bezitten. Hoofdstuk 7 betreft Robinson matrices. Een vierkante coëfficiëntenmatrix wordt een Robinson matrix genoemd als de hoogste waardes in iedere rij en kolom op de hoofddiagonaal liggen, en wegbewegend van de hoofddiagonaal zijn de waardes nooit oplopend. In hoofdstuk 8 worden eigenwaardes en eigenvectoren van coëfficiëntenmatrices bestudeerd. Als het double monotonicity model voor binaire items opgaat, dan wordt de correcte ordering van de items weerspiegeld in de elementen van de eigenvector behorende bij de grootste eigenwaarde van de matrix met elementen $a(i,j)/p(j)$, waar $a(i,j)$ de proportie 1n is dat items i en j in dezelfde posities hebben, en $p(j)$ is de proportie item correct van item j. In hoofdstuk 9 wordt een systematische vergelijking gemaakt tussen een eigenwaarde techniek, homogeniteitsanalyse, en het logistische item response theory model met twee parameters. Hoofdstuk 10 is het eerste hoofdstuk waar metrische eigenschappen van coëfficiënten worden bestudeerd. Een functie wordt metrisch genoemd als deze voldoet aan de driehoeksongelijkheid. Dit hoofdstuk dient als opstap naar deel III, waar allerlei generalisaties van driehoeksongelijkheid worden gedefiniëerd en besproken.

Deel III bestaat uit vijf hoofdstukken. Voordat meerweg coëfficiënten bestudeerd kunnen worden in deel IV, wordt eerst een aantal meerweg concepten gedefiniëerd en bestudeerd in deel III. Ideeën voor de meerweg concepten zijn vooral verkregen door te kijken naar literatuur over drieweg data-analyse. Hoofdstuk 11 behandelt axioma's en basiseigenschappen die kunnen opgaan voor meerweg coëfficiënten en hun complementen, afstandsmaten. In dit hoofdstuk wordt onder andere bestudeerd wat mogelijk de kleinste sets van axioma's zijn. In hoofdstuk 12 wordt geëxploreerd op welke manieren de driehoeksongelijkheid kan worden gegeneraliseerd naar ongelijkheden voor vier of meer objecten. Een voorbeeld is hier een ongelijkheid gebaseerd op de tetraëder, waarbij het oppervlakte van een van de zijdes van de tetraëder altijd kleiner of gelijk is aan de som van de oppervlaktes van de drie overige zijdes. Deze ongelijkheden definiëren verschillende meerweg metrieken. In hoofdstuk 13

worden meerweg ultrametrieken bestudeerd en hoofdstuk 14 gaat over hoe twee specifieke drieweg functies gegeneraliseerd kunnen worden. Hoofdstuk 15 beschrijft twee manieren om een resultaat uit hoofdstuk 10 te generaliseren. Dit resultaat vertelt ons dat als een afstandsmaat k aan de driehoeksongelijkheid voldoet, dan voldoet de functie $k/(e+k)$ daar ook aan, waarbij e een positief getal is.

Als laatste bestaat deel IV uit vier hoofdstukken. In dit laatste deel worden meerweg formuleringen van coëfficiënten behandeld. In hoofdstuk 17 zijn de formuleringen functies van de tweeweg informatie, ofwel de coëfficiënten uit deel I. De meerweg coëfficiënten in hoofdstuk 16 zijn geen functies van de tweeweg informatie, maar in dit hoofdstuk wordt een poging om coëfficiënten te formuleren die een kern of basiseigenschap van de tweeweg coëfficiënten generaliseren. Metrische eigenschappen van de meerweg coëfficiënten worden onderzocht in hoofdstuk 18. Hoofdstuk 19 behandelt de drieweg uitbreiding van de Robinson matrices uit hoofdstuk 7, Robinson kubussen genoemd.

Curriculum vitae

Matthijs Joost Warrens werd geboren op 13 december 1978 te Rotterdam. In 1997 behaalde hij zijn Gymnasium diploma aan de Johannes Calvijn te Rotterdam. Hierna volgden de studies Informatica, Muziekwetenschap en Psychologie aan de universiteiten van Leiden en Amsterdam. De Leidse studie Psychologie werd in 2003 afgerond met het doctoraal examen in de afstudeerrichting Methoden en Technieken van psychologisch onderzoek. Van 2003 tot 2008 was Matthijs aangesteld als promovendus aan de afdeling Methoden en Technieken aan de Universiteit Leiden. Thans is hij als post-doctoraal medewerker verbonden aan dezelfde afdeling.