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

Warrens, M.J.

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

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

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*

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

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

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* $\mathbf{S}_{\mathrm{RR}}$, $\mathbf{S}_{\mathrm{DK}}$, $\mathbf{S}_{\mathrm{BB}} = \mathbf{S}_{\mathrm{Jac}} = \mathbf{S}_{\mathrm{Sorg}}$ and $\mathbf{S}_{\mathrm{Phi}}$ *are Green's matrices.*
Proof: If $a_{jk} = \min(p_j, p_k)$ and $p_j \geq p_{j+1}$, then

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

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

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

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

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 $S_{\text{GL1}}(\theta)$ and $S_{\text{GL2}}(\theta)$.

Proposition 7.5. *Let* $\mathbf{S}$ *and* $\mathbf{S}^*$ *be coefficient matrices corresponding to any two members of* $S_{\text{GL1}}(\theta)$. $\mathbf{S}$ *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 $S_{\text{GL1}}(\theta)$ 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* $S_{\text{GL2}}(\theta)$. $\mathbf{S}$ *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 $S_{\text{GL2}}(\theta)$ 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

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

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 $\mathbf{S}_{\text{BB}}$ 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* $\mathbf{S}_{\text{BB}}$ with $S_{\text{BB}} = a_{jk}/\max(p_j, p_k)$ *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

$$S_{\text{RR}} = \frac{a}{a+b+c+d} \qquad \text{(Russel and Rao, 1940)}$$

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 $\mathbf{S}_{\text{RR}}$.

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

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

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 $\mathbf{S}_{\text{Jac}}$ is a Robinson matrix for many row Petrie data matrices, and that in many practical cases it has approximately the same properties as $\mathbf{S}_{\text{RR}}$.

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.

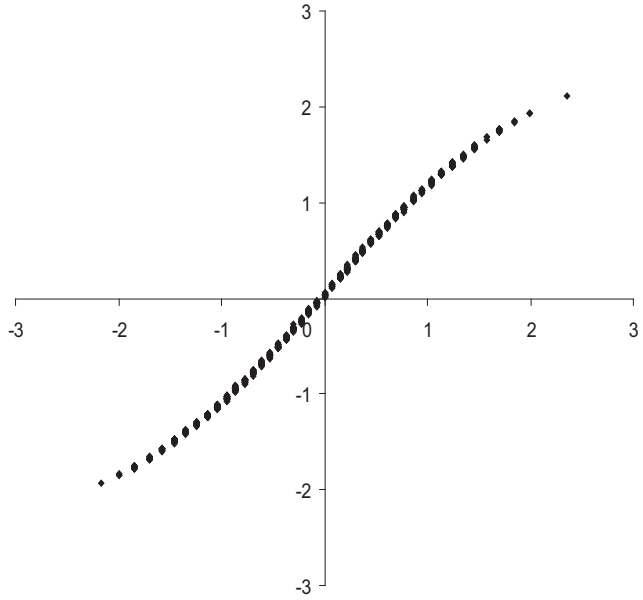

Figure 9.1: *Plot of maximum a posteriori person estimates (horizontal) versus homogeneity person scores (vertical) for the Rasch data set.*

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

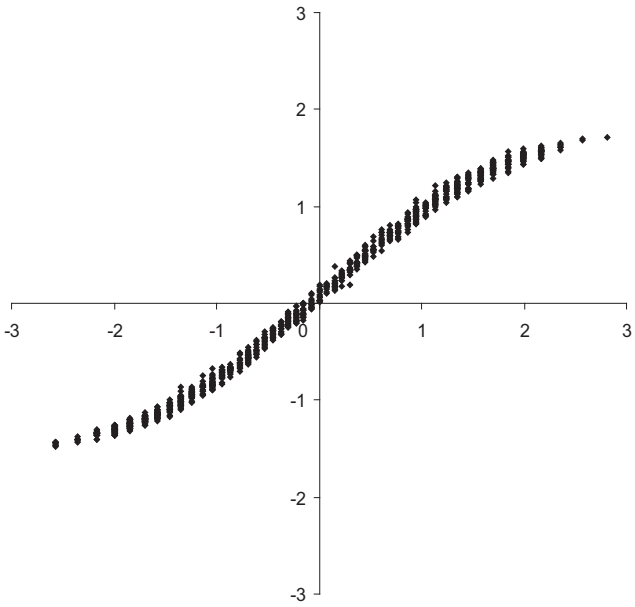

Figure 9.2: *Plot of maximum a posteriori person estimates (horizontal) versus homogeneity person scores (vertical) for the logistic 2-parameter model data set.*

$v_{j1}^{(t)}$ (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

$$D_{\mathrm{RR}} = 1 - a \quad \text{and} \quad D_{\mathrm{SM}} = b + c.$$

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

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

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.