Hindawi Publishing Corporation
Journal of Mathematics
Volume 2015, Article ID 582731, 5 pages
http://dx.doi.org/10.1155/2015/582731

Research Article

Ordering Properties of the First Eigenvector of Certain Similarity Matrices

Matthijs J. Warrens and Alexandra de Raadt

GION, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, Netherlands

Correspondence should be addressed to Matthijs J. Warrens; m.j.warrens@rug.nl

Received 21 July 2015; Revised 24 October 2015; Accepted 1 November 2015

Academic Editor: Niansheng Tang

It is shown for coefficient matrices of Russell-Rao coefficients and two asymmetric Dice coefficients that ordinal information on a latent variable model can be obtained from the eigenvector corresponding to the largest eigenvalue.

1. Introduction

An important role in statistics and data analysis is played by similarity coefficients. A similarity coefficient is a measure of resemblance or association of two data vectors, such as score patterns, variables, and items. For example, in ecological biology similarity coefficients are used for measuring the degree of coexistence between two species types over different locations. In many research studies the data consist of binary (0, 1) vectors: presence or absence of disease; presence or absence of species characteristics; yes or no answers in questionnaires; pass or fail in high-stakes testing. For expressing the degree of resemblance of two binary vectors in a number, a variety of similarity coefficients has been proposed [1–3]. Examples are the Jaccard coefficient [4], the Russell-Rao coefficient [5], the Dice coefficient [6], and the simple matching coefficient [7, 8]. In choosing a coefficient, a measure has to be considered in the context of the data-analytic study of which it is a part [9]. Because there are so many similarity coefficients for binary data to choose from, it is important that the different coefficients and their properties are better understood.

Instead of studying properties of individual coefficients [10–13] one may also study properties of coefficient matrices [14]. Coefficient matrices are used as input in various techniques of multivariate data analysis, including factor or component analysis [15, 16], hierarchical cluster analysis, and techniques in classification and dissimilarity analysis [17]. Moreover, exploratory data-analytic methods such as principal coordinates analysis and (multiple) correspondence analysis can be defined as eigendecomposition of certain coefficient matrices [15, 16, 18]. It would be interesting to know what information, if any, is reflected in the eigenvectors of a coefficient matrix that is based on a similarity coefficient for binary vectors.

In this paper we show for several coefficient matrices that ordinal information on latent variable models can be obtained from the eigenvector corresponding to the largest eigenvalue. It is thus possible to uncover meaningful orderings of various models by using eigenvectors. The results are first of all of theoretical interest. They show that some coefficient matrices have more interesting eigenvectors than others. Coefficient matrices based on some coefficients may thus lead to more interesting data-analytic solutions than matrices corresponding to other coefficients. Furthermore, potentially, the results can enhance the interpretation of a data analysis that uses these coefficient matrices as input.

The paper is organized as follows. Notation and two latent variable models are introduced in the next section. In Section 3 several ordering properties of eigenvectors corresponding to a largest eigenvalue are presented. An illustration of the results is presented in Section 4. Section 5 contains a conclusion.

2. Latent Variable Models

Suppose the data consist of m binary $(0, 1)$ vectors of length n. It may be assumed that the scores in the binary vectors are realizations of a latent variable model. In this section

we introduce two models in the context of nonparametric item response theory [19, 20]. In item response theory the m vectors are often viewed as m items that, for instance, contain the responses (pass, fail) of a high-stakes test for n subjects. The m items will be indexed by i and j.

Let θ denote a one-dimensional latent variable and let $L(\theta)$ be its probability density function. Let $p_i(\theta)$ denote the response function corresponding to the response 1 on item i. The unconditional probability of a response 1 on item i is then given by

$$p_i = \int_{-\infty}^{\infty} p_i(\theta)\, dL(\theta). \tag{1}$$

Next, assume local independence; that is, conditionally on θ the responses of a subject on the items are stochastically independent. The joint probability of items i and j for a value of θ is then given by $p_i(\theta)p_j(\theta)$. The corresponding unconditional probability is

$$a_{ij} = \int_{-\infty}^{\infty} p_i(\theta)\, p_j(\theta)\, dL(\theta). \tag{2}$$

Throughout the paper we assume that $0 < a_{ij} \leq 1$.

Next, we define the latent variable models. Both models have monotone response functions and are frequently applied in the context of measuring ability. The first model is characterized by requirements (3) and (4). The first requirement is that $p_i(\theta)$ are monotonically increasing on θ; that is,

$$p_i(\theta_1) \leq p_i(\theta_2) \tag{3}$$

for $\theta_1 < \theta_2$. The second requirement is that the m items can be ordered such that $p_i(\theta)$ are nonintersecting; that is,

$$p_i(\theta) \geq p_j(\theta) \tag{4}$$

for $i < j$. The case that assumes (3) and (4), together with the assumptions of local independence and a single latent variable, is called the double monotonicity model in nonparametric item response theory [19, 20]. A well-known result is that if the double monotonicity model holds, then the items can be ordered such that we have

$$p_i > p_j \tag{5}$$

for $i < j$, and

$$a_{ij} \geq a_{i'j} \tag{6}$$

for $i < i'$ and $j \neq i'$ [19, 20]. The second model is characterized by requirements (3) and (7). The response functions $p_i(\theta)$ may satisfy various orders of total positivity [21]. If the functions $p_i(\theta)$ are totally positive of order 2, the items can be ordered such that

$$p_i(\theta_1)\, p_j(\theta_2) - p_i(\theta_2)\, p_j(\theta_1) \geq 0 \tag{7}$$

holds for $\theta_1 < \theta_2$ and $i < j$. Schriever [22] proved the following result for a set of response functions that are both monotonically increasing and satisfy total positivity of order 2. If the vectors are ordered such that (3) and (7) hold, then

$$\frac{a_{ij}}{p_i} \leq \frac{a_{i'j}}{p_{i'}} \tag{8}$$

holds for $i < i'$ and $j \neq i'$.

We conclude this section with a parametric example that satisfies requirements (3), (4), and (7). A well-known model from the field of item response theory is the Rasch [23] model. A response function of this one-parameter logistic model is given by

$$p_i(\theta, b_i) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}}, \tag{9}$$

where b_i is a location parameter. In the context of item response theory the parameter b_i is usually called a difficulty parameter [19, 20]. The functions $p_i(\theta, b_i)$ form a location family.

3. Ordering Properties

In this section we present ordering properties for three coefficient matrices. The coefficient matrices of size $m \times m$ are

$$\begin{aligned} A &= \left(a_{ij}\right), \\ B &= \left(\frac{a_{ij}}{p_i}\right), \\ C &= \left(\frac{a_{ij}}{p_j}\right). \end{aligned} \tag{10}$$

An element of the matrix A is a Russell-Rao coefficient for two binary vectors i and j [5, 10]. Some data-analytic properties of the matrix A are discussed in Warrens [14]. The elements of the matrices B and C are conditional probabilities discussed and applied in Dice [6]. The harmonic mean of the two conditional probabilities is equal to the Dice coefficient [6]. Matrix C is also called the conditional adjacency matrix in Post and Snijders [24].

A specific result that will be used in the proofs of Theorems 2, 3, and 4 below is the Perron-Frobenius theorem [25, 26]. More precisely, only the following weaker version of the Perron-Frobenius theorem will be used.

Lemma 1. *If a square matrix D has strictly positive elements, then the eigenvector y corresponding to the largest eigenvalue of D has strictly positive elements.*

In the proof of Theorems 2, 3, and 4 we use certain special matrices. Let S denote the upper triangular matrix of size $k \times k$ ($2 \leq k \leq m$) with unit elements on and above the diagonal and all other elements zero. Its inverse S^{-1} is the matrix with

unit elements on the diagonal and with elements −1 adjacent and above the diagonal. Examples of S and S^{-1} of size 3×3 are

$$S = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad S^{-1} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{11}$$

Furthermore, let I be the identity matrix of size $(m-k)\times(m-k)$, and let T denote the diagonal block matrix of size $m\times m$ with diagonal elements S and I. Examples of T and T^{-1} of size 4×4 are

$$T = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad T^{-1} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{12}$$

We first consider the matrix C. Let y be the eigenvector corresponding to the largest eigenvalue λ of the matrix C. Theorem 2 shows that if the binary vectors can be ordered such that (3) and (4) hold, then this ordering is reflected in the corresponding elements of y.

Theorem 2. *Suppose that k of the m vectors, which without loss of generality can be taken as the first k, can be ordered such that (3) and (4) hold. Then the elements of y of C corresponding to these k vectors satisfy* $y_1 > y_2 > \cdots > y_k > 0$.

Proof. Since T is nonsingular, y is an eigenvector of C corresponding to λ if and only if $z = T^{-1}y$ is an eigenvector of $D = T^{-1}CT$ corresponding to λ. Under the conditions of the theorem, the elements of D are positive and the elements of D^2 are strictly positive. Application of Lemma 1 then yields that the eigenvector z of D (or D^2) has strictly positive elements. The assertion then follows from the identity $z = T^{-1}y$.

In the remainder of the proof we show that D has positive elements and D^2 has strictly positive elements. The matrix $U = T^{-1}C$ has elements

$$u_{ij} = \frac{a_{ij} - a_{i+1,j}}{p_j} \tag{13}$$

for $1 \le i < k$ and $1 \le j \le m$ and

$$u_{ij} = \frac{a_{ij}}{p_j} \tag{14}$$

for $k \le i \le m$ and $1 \le j \le m$. Under the conditions of the theorem properties (5) and (6) hold for the first k items. By (6), we have $a_{ij} \ge a_{i+1,j}$, and the matrix U has positive elements except for $u_{i,i+1}$ for $1 \le i \le k-1$. However by (5), we have $p_i > p_{i+1}$ and it follows that

$$\begin{aligned} u_{ii} + u_{i,i+1} &= \frac{p_{i+1}a_{ii} - p_{i+1}a_{i,i+1} + p_i a_{i,i+1} - p_i a_{i+1,i+1}}{p_i p_{i+1}} \\ &= \frac{a_{i,i+1}\left(p_i - p_{i+1}\right)}{p_i p_{i+1}} > 0 \end{aligned} \tag{15}$$

for $1 \le i \le k-1$. Hence, the matrix $D = UT$ has positive elements. Moreover, because the elements in the last row and last column of D are strictly positive, it follows that the elements of D^2 are strictly positive. □

An analogous result holds for the matrix A. Let y be the eigenvector corresponding to the largest eigenvalue λ of the matrix A. Theorem 3 shows that if the binary vectors can be ordered such that (3) and (4) hold, then this ordering is reflected in the corresponding elements of y of A.

Theorem 3. *Suppose that k of the m vectors, which without loss of generality can be taken as the first k, can be ordered such that (3) and (4) hold. Then the elements of y of A corresponding to these k vectors satisfy* $y_1 > y_2 > \cdots > y_k > 0$.

Proof. The proof is similar to the proof of Theorem 2. The matrix $U = T^{-1}A$ has elements

$$u_{ij} = a_{ij} - a_{i+1,j} \tag{16}$$

for $1 \le i < k$ and $1 \le j \le m$ and

$$u_{ij} = a_{ij} \tag{17}$$

for $k \le i \le m$ and $1 \le j \le m$. Under the conditions of the theorem properties (5) and (6) hold for the first k items. By (6), we have $a_{ij} \ge a_{i+1,j}$, and the matrix U has positive elements except for $u_{i,i+1}$ for $1 \le i \le k-1$. But by (5), we have $p_i > p_{i+1}$, and it follows that

$$\begin{aligned} u_{ii} + u_{i,i+1} &= a_{ii} - a_{i,i+1} + a_{i,i+1} - a_{i+1,i+1} = p_i - p_{i+1} \\ &> 0 \end{aligned} \tag{18}$$

for $1 \le i \le k-1$. □

Finally, Theorem 4 below presents an ordering property of the matrix B. The ordering holds for a slightly stronger model than the one considered in Theorems 2 and 3. Theorem 4 shows that if the binary vectors can be ordered such that (3), (4), and (7) hold, then this ordering is reflected in the corresponding elements of y of B.

Theorem 4. *Suppose that k of the m vectors, which without loss of generality can be taken as the first k, can be ordered such that (3), (4), and (7) hold. Then the elements of y of B corresponding to these k vectors satisfy* $0 < y_1 < y_2 < \cdots < y_k$.

Proof. The proof is similar to the proof of Theorems 2 and 3. Let $(T^{-1})'$ denote the transpose of T^{-1}. The matrix $U = (T^{-1})'B$ has elements

$$u_{ij} = \frac{p_{i-1}a_{ij} - p_i a_{i-1,j}}{p_{i-1}p_i} \tag{19}$$

for $2 \leq i \leq k$ and $1 \leq j \leq m$ and

$$u_{ij} = \frac{a_{ij}}{p_i} \tag{20}$$

for $k < i \leq m$ and $1 \leq j \leq m$. Under the conditions of the theorem properties (5) and (8) hold. By (8), we have $p_{i-1}a_{ij} \geq p_i a_{i-1,j}$, and the matrix U has positive elements except for u_{ii-1} for $2 \leq i \leq k$. However, by (5), we have $p_{i-1} > p_i$, and it follows that

$$\begin{aligned} u_{i,i-1} + u_{ii} &= \frac{p_{i-1}a_{i,i-1} - p_i a_{i-1,i-1} + p_{i-1}a_{ii} - p_i a_{i,i-1}}{p_{i-1}p_i} \\ &= \frac{a_{i,i-1}\left(p_{i-1} - p_i\right)}{p_{i-1}p_i} > 0 \end{aligned} \tag{21}$$

for $2 \leq i \leq k$. □

4. An Illustration

In this section we consider an example from educational testing to illustrate some of the results from Section 3. The data consist of responses of 1000 individuals to five items of the LSAT (Law School Admission Test). The test was designed to measure a one-dimensional latent variable. The example is part of a data set given by Bock and Lieberman [27]. The data set is distributed with the R package "ltm" written by Rizopoulos [28].

Requirements (3), (4), and (7) cannot be checked directly for real life data. However, it can be shown that the Rasch model in (9) fits these data quite well. Using subroutines from the "ltm" package we fitted the Rasch model and the so-called two-parameter logistic model [19, 20]. In the Rasch model the items are allowed to differ in location. In the more general two-parameter model the items are also allowed to differ in slope. For these data the two-parameter model has four additional parameters. The log likelihoods of the models are −2466.94 and −2466.65, respectively, and the corresponding likelihood ratio test has a p value of $p = .967$. Thus, the extra slope parameters are statistically not warranted.

Requirements (3), (4), and (7) can also be studied by verifying if conditions (5), (6), and (8) hold. The proportions of correct responses are $p_1 = .924$, $p_2 = .709$, $p_3 = .553$, $p_4 = .763$, and $p_5 = .870$ for items 1 to 5, respectively. For verifying conditions (6) and (8), we suppose that the items are ordered on the proportions of correct responses, from easiest to hardest item (1, 5, 4, 2, and 3). In other words, in the following we assume that the items are ordered such that condition (5) holds.

To study condition (6) we may inspect the matrix A of Russell-Rao coefficients. For the LSAT data this matrix is given by

$$A = \begin{pmatrix} .924 & .806 & .710 & .664 & .524 \\ .806 & .870 & .678 & .630 & .490 \\ .710 & .678 & .763 & .553 & .445 \\ .664 & .630 & .553 & .709 & .418 \\ .524 & .490 & .445 & .418 & .553 \end{pmatrix}. \tag{22}$$

The elements on the main diagonal are the proportions of correct responses. If we ignore the elements on the main diagonal it can be verified that the other four elements in each column of A are strictly decreasing. Hence, condition (6) holds.

Since conditions (5) and (6) hold for all five LSAT items it follows from Theorem 3 that the ordering of the items is reflected in the eigenvector corresponding to the largest eigenvalue of A. The largest eigenvalue is $\lambda = 3.191$ and the associated eigenvector is given by (.516, .495, .446, .420, .336). The item ordering is thus reflected in the elements of the eigenvector.

To verify whether condition (8) holds we may inspect the matrix B of Dice coefficients. For the LSAT data this matrix is given by

$$B = \begin{pmatrix} 1.00 & .872 & .768 & .719 & .567 \\ .926 & 1.00 & .779 & .724 & .563 \\ .931 & .889 & 1.00 & .725 & .583 \\ .937 & .889 & .780 & 1.00 & .590 \\ .948 & .886 & .805 & .756 & 1.00 \end{pmatrix}. \tag{23}$$

If we ignore the elements on the main diagonal it can be verified that the remaining four elements in the first, third, and fourth columns of B are strictly increasing. Furthermore, the elements in the second and fifth columns are roughly increasing. In both columns there is one anomaly. We may conclude that condition (8) holds approximately.

If the five LSAT items satisfy conditions (5) and (8) it follows from Theorem 4 that the ordering of the items is reflected in the eigenvector corresponding to the largest eigenvalue of B. The largest eigenvalue is $\lambda = 4.106$ and the associated eigenvector is given by (.424, .431, .446, .454, .478). The item ordering is thus reflected in the elements of the eigenvector.

5. Conclusion

Similarity coefficients for binary vectors are frequently used in statistics for analyzing the structure between objects. Examples that are commonly used are the Russell-Rao coefficient [5] and the Dice coefficient [6]. Since the choice of a coefficient depends on the context of the data-analytic study, it is important that the different coefficients and their properties are well understood.

In this paper we showed that ordinal information on latent variable models is reflected in the eigenvector corresponding to the largest eigenvalue of the coefficient matrices with Russell-Rao coefficients (Theorem 3) and two asymmetric coefficients used in Dice [6] (Theorems 2 and 4). For other well-known coefficients like the Jaccard coefficient [4] and the simple matching coefficient similar ordering properties could not been found. The results may indicate that the Russell-Rao coefficient and Dice coefficients may lead to more clearly interpretable output if used as input in clustering methods or principal coordinates analysis. However, more research on this topic is needed.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko, "On similarity indices and correction for chance agreement," *Journal of Classification*, vol. 23, no. 2, pp. 301–313, 2006.

[2] F. B. Baulieu, "A classification of presence/absence based dissimilarity coefficients," *Journal of Classification*, vol. 6, no. 2, pp. 233–246, 1989.

[3] M. J. Warrens, "On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions," *Psychometrika*, vol. 73, no. 4, pp. 777–789, 2008.

[4] P. Jaccard, "The distribution of the ora in the Alpine zone," *The New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[5] P. F. Russell and T. R. Rao, "On habitat and association of species of *Anopheline* larvae in South-Eastern Madras," *Journal of Malaria Institute India*, vol. 3, pp. 153–178, 1940.

[6] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[7] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, vol. 38, pp. 1409–1438, 1958.

[8] M. J. Warrens, "On similarity coefficients for 2×2 tables and correction for chance," *Psychometrika*, vol. 73, no. 3, pp. 487–502, 2008.

[9] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *Journal of Classification*, vol. 3, no. 1, pp. 5–48, 1986.

[10] M. J. Warrens, "Bounds of resemblance measures for binary (presence/absence) variables," *Journal of Classification*, vol. 25, no. 2, pp. 195–208, 2008.

[11] M. J. Warrens, "On the indeterminacy of resemblance measures for binary (presence/absence) data," *Journal of Classification*, vol. 25, no. 1, pp. 125–136, 2008.

[12] M. J. Warrens, "Corrected Zegers-ten Berge coefficients are special cases of Cohen's weighted kappa," *Journal of Classification*, vol. 31, no. 2, pp. 179–193, 2014.

[13] M. J. Warrens, "Properties of the quantity disagreement and the allocation disagreement," *International Journal of Remote Sensing*, vol. 36, pp. 1439–1446, 2015.

[14] M. J. Warrens, "On Robinsonian dissimilarities, the consecutive ones property and latent variable models," *Advances in Data Analysis and Classification*, vol. 3, no. 2, pp. 169–184, 2009.

[15] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, pp. 325–338, 1966.

[16] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, New York, NY, USA, 1984.

[17] B. Mirkin, *Mathematical Classification and Clustering*, Kluwer, Dordrecht, The Netherlands, 1984.

[18] A. Gifi, *Nonlinear Multivariate Analysis*, Wiley, Chichester, UK, 1990.

[19] W. J. Van der Linden and R. K. Hambleton, *Handbook of Modern Item Response Theory*, Springer, Berlin, Germany, 1997.

[20] K. Sijtsma and I. W. Molenaar, *Introduction to Nonparametric Item Response Theory*, SAGE Publications, Thousand Oaks, Calif, USA, 2002.

[21] S. Karlin, *Total Positivity*, Stanford Univeristy Press, Stanford, Calif, USA, 1968.

[22] B. F. Schriever, "Multiple correspondence analysis and ordered latent structure models," *Kwantitatieve Methoden*, vol. 21, pp. 117–131, 1986.

[23] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, Studies in Mathematical Psychology, Danish Institute for Educational Research, Copenhagen, Denmark, 1984.

[24] W. J. Post and T. A. B. Snijders, "Nonparametric unfolding models for dichotomous data," *Methodika*, vol. 7, pp. 130–156, 1993.

[25] F. R. Gantmacher, *Matrix Theory*, Chelsea, New York, NY, USA, 1977.

[26] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, NY, USA, 1973.

[27] R. D. Bock and M. Lieberman, "Fitting a response model for n dichotomously scored items," *Psychometrika*, vol. 35, no. 2, pp. 179–197, 1970.

[28] D. Rizopoulos, "ltm: an R package for latent variable modeling and item response theory analyses," *Journal of Statistical Software*, vol. 17, no. 5, pp. 1–25, 2006.