Computational Statistics and Data Analysis 53 (2009) 3139–3144

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Diagnostics for regression dependence in tables re-ordered by the dominant correspondence analysis solution

Matthijs J. Warrens *, Willem J. Heiser

Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands

ARTICLE INFO

Article history:
Available online 5 August 2008

ABSTRACT

Correspondence analysis is an exploratory technique for analyzing the interaction in a contingency table. Tables with meaningful orders of the rows and columns may be analyzed using a model-based correspondence analysis that incorporates order constraints. However, if there exists a permutation of the rows and columns of the contingency table so that the rows are regression dependent on the columns and, vice versa, the columns are regression dependent on the rows, then both implied orders are reflected in the first dimension of the unconstrained correspondence analysis [Schriever, B.F., 1983. Scaling of order dependent categorical variables with correspondence analysis. International Statistical Review 51, 225–238]. Thus, using unconstrained correspondence analysis, we may still find that the data fit an ordinal stochastic model. Fit measures are formulated that may be used to verify whether the re-ordered contingency table is regression dependent in either the rows or columns. Using several data examples, it is shown that the fit indices may complement the usual geometric interpretation of the unconstrained correspondence analysis solution in low-dimensional space.

1. Introduction

Correspondence analysis (CA; cf. Benzécri (1973, 1992) and Greenacre (1984, 1993)) is an exploratory technique that can be used for the graphical and numerical analysis of a matrix of counts or frequencies. Although there are different ways of thinking about CA, the geometric approach appears to be the dominant view nowadays (Blasius and Greenacre, 2006; Borg and Groenen, 2005, Section 24.2). In the geometric approach to CA the rows and columns of the contingency table are assumed to be points in high-dimensional space. CA finds multi-dimensional scores for the rows and columns that redefine the dimensions of the space so that the principal dimensions capture a maximal amount of variance. The emphasis lies on the well-defined geometric interpretation of these scores. Low-dimensional descriptions of the data are possible if a substantial proportion of the variance is captured by the principal dimensions.

Although CA was originally designed as a descriptive tool, it has been modified to incorporate order constraints. Special algorithms to obtain solutions of the CA problem under order constraints induced by the categories are considered in, e.g. Heiser (1981) or Gifi (1990). Goodman (1986), together with a group of discussants, presented a discussion on modified CA and the analysis of association in contingency tables via the log-linear approach. Since then CA has been developed into an efficient model-based approach (Gilula and Ritov, 1990; Ritov and Gilula, 1993). This development is in line with the interest for analyzing ordered and unordered tables by saturated and non-saturated models based on scores assigned to the rows and columns of the table (cf. Agresti (2002)).

* Corresponding author.
E-mail addresses: warrens@fsw.leidenuniv.nl (M.J. Warrens), heiser@fsw.leidenuniv.nl (W.J. Heiser).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.07.035

Table 1
The appreciations of five red Bordeaux wines by 200 judges using a four category system (Van Rijckevorsel, 1987, p. 60)

Wine	Category				Row totals
	Excellent	Good	Mediocre	Boring	
Grand Cru Classé	87	93	19	1	200
Cru Bourgeois	45	126	24	5	200
Bordeaux d'Origine	36	68	74	22	200
Vin de Marque	0	30	111	59	200
Vin de Table	0	0	52	148	200
Column totals	168	317	280	235	1000

An inconsistency in the original table (column total of 286 for 'Mediocre') has been adjusted.

Schriever (1983) presented a powerful result that connects CA to the concept of positive regression dependence (cf. Lehmann (1966)). Schriever (1983) has pointed to the fact that if there exists a permutation of the rows and columns of the contingency table so that the rows are regression dependent on the columns and, vice versa, the columns are regression dependent on the rows, then CA recovers both implied orders. The orders are reflected in the category scores of the first CA dimension. Thus, using unconstrained CA, instead of inferential CA, we may still find that the data fit an ordinal stochastic model. Because the gain in the interpretative value of an ordered solution might be large, identifying the ordinal model without imposing it is certainly worthwhile. Identification of a model in the presence of noise can be achieved by using a diagnostic that captures a critical feature of the model.

The paper is organized as follow: Regression dependence is discussed in the next section. To promote Schriever's (1983) work on CA and regression dependence, diagnostics that may facilitate the interpretation of the CA solution are formulated in Section 3, in terms of fit indices. Applications of the fit indices and data examples of regression dependence are presented in Section 4. We focus on interpreting the order of the rows and columns of each contingency table in terms of an ordinal model and will not inspect the usual two-dimensional plot of the CA solution. Section 5 contains the discussion.

2. Regression dependence

Let $\mathbf{X}$ be a contingency table of size $n \times m$ with nonnegative entries x_{ij}. Let

$$x_{i+} = \sum_{j=1}^{m} x_{ij} \quad \text{and} \quad x_{+j} = \sum_{i=1}^{n} x_{ij}$$

be, respectively, the row and column totals of table $\mathbf{X}$, and let

$$x_{++} = \sum_{i=1}^{n} x_{i+} = \sum_{j=1}^{m} x_{+j}$$

be the grand total. Furthermore, let $\mathbf{P}$ of size $n \times m$ have entries $p_{ij} = x_{ij}/x_{++}$, and let I and J denote the two variables indicating row and column number of $\mathbf{X}$ and $\mathbf{P}$.

Suppose the rows and columns of $\mathbf{P}$ can be indexed such that the family of induced distributions $J|I = i$ is stochastically increasing, i.e. $\mathbf{P}\{J \leq j_0 | I = i\}$ is decreasing in i for each j_0. The conditional distribution functions of the variable $J|I = i$ are then stochastically ordered in a sequence which is identical to the order given by the row index i. Tong (1980, p. 79) and Lehmann (1966) call this form of bivariate dependence, positive regression dependence of J on I. If J is regression dependent on I (if the columns are regression dependent on the rows), then the $n \times m$ table $\mathbf{X}$ must satisfy

$$1 \leq i < i' \leq n,\ 1 \leq j_0 \leq m-1 \quad \Rightarrow \quad \frac{1}{x_{i+}} \sum_{j=1}^{j_0} x_{ij} \geq \frac{1}{x_{i'+}} \sum_{j=1}^{j_0} x_{i'j}.$$

Schriever (1983) calls this case row regression dependence (RR) of $\mathbf{X}$. Thus, $\mathbf{X}$ is RR if for any two rows i and i' with $i < i'$, the partial row sum of column 1, column 1 + column 2, column 1 + column 2 + column 3, etc., each divided by the corresponding row sum, is larger for row i than for row i'. Column regression dependence of $\mathbf{X}$ is defined by interchanging the roles of the rows and columns. The table $\mathbf{X}$ is column regression dependent (CR) if for any two columns j and j' with $j < j'$, the partial column sum of row 1, row 1 + row 2, row 1 + row 2 + row 3, etc., each divided by the corresponding column sum, is larger for column j than for column j'.

Schriever (1983) proved that if there exists a permutation of the rows and columns of the contingency table $\mathbf{X}$ so that $\mathbf{X}$ is both RR and CR, then both implied orders are recovered by CA. The correct orders of rows and columns are reflected in the category scores of the first CA dimension. A data matrix that is both RR and CR is presented in Table 1. To illustrate the property derived in Schriever (1983), Van Rijckevorsel (1987) reported the results of a blind wine tasting of five typical red Bordeaux wines by 200 judges. Four categories were used, varying from excellent to boring. The rows and columns of Table 1 are permuted using the scores of the first CA dimension.

Table 2
The cumulative row percentages (a) and cumulative column percentages (b) of the wine data from Table 2

Wines	Category			
	Excellent	Good	Mediocre	Boring
(a)				
Grand Cru Classé	0.44	0.90	1.00	1.00
Cru Bourgeois	0.23	0.86	0.98	1.00
Bordeaux d'Origine	0.18	0.52	0.89	1.00
Vin de Marque	0.00	0.15	0.71	1.00
Vin de Table	0.00	0.00	0.26	1.00
(b)				
Grand Cru Classé	0.52	0.29	0.07	0.00
Cru Bourgeois	0.79	0.69	0.15	0.03
Bordeaux d'Origine	1.00	0.91	0.42	0.12
Vin de Marque	1.00	1.00	0.81	0.37
Vin de Table	1.00	1.00	1.00	1.00

Table 2 presents the cumulative row percentages (a) and cumulative column percentages (b) of the data in Table 1. We may verify that Table 1 is both RR and CR by inspecting panels (a) and (b) in Table 2. If Table 1 is RR, then each column in panel (a) of Table 2 should not increase moving from top to bottom. If Table 1 is CR, then each row in panel (b) of Table 2 should not increase moving from left to right.

Since Table 1 is both RR and CR, the interpretation of the wine data is not difficult. There exists a strong ordinal association between the categories of variables I and J. Moreover, all other information from a CA, i.e. dimensions higher than the first, may be considered noise in light of the stochastic ordinal model. The five wines can be perfectly ordered from excellent to boring: the Grand Cru Classé is considered excellent by the judges whereas the Vin de Table is considered boring. The first CA dimension explains 81% of the variance, and accounts for all of the ordering of the row and column categories.

3. Diagnostics for regression dependence

It seems useful to be able to verify in general whether a contingency table is either RR or CR after its rows and columns are re-ordered by a CA solution. We therefore formulate diagnostics that can be used to check for both forms of positive regression dependence.

Denote by $\mathbf{R}$ the diagonal matrix of order n with the row totals x_{i+} of $\mathbf{X}$ on its main diagonal. Furthermore, let $\mathbf{S}_n$ be the upper triangular matrix of order n with elements 1 on and above its main diagonal and elements 0 below its diagonal. Its inverse $\mathbf{S}_n^{-1}$ is a matrix with elements 1 on its main diagonal, elements -1 directly above and adjacent to the main diagonal, and all other elements zero.

The matrix $\mathbf{R}^{-1}\mathbf{X}$ is equal to table $\mathbf{X}$ divided by its row totals. Furthermore, the matrix $\mathbf{R}^{-1}\mathbf{X}\mathbf{S}_m$ is equal to the matrix with cumulative row percentages (see e.g. panel (a) of Table 2). The first $n-1$ rows of matrix $\mathbf{D} = \mathbf{S}_n^{-1}\mathbf{R}^{-1}\mathbf{X}\mathbf{S}_m$ contain the differences between adjacent rows of matrix $\mathbf{R}^{-1}\mathbf{X}\mathbf{S}_m$. Matrices $\mathbf{D}$ and $\mathbf{X}$ are both of size $n \times m$, and the bottom row of $\mathbf{D}$ is equal to the bottom row of $\mathbf{X}$.

Table $\mathbf{X}$ is RR when all elements of $\mathbf{D}$ are nonnegative. Regression dependence of the columns on the rows can thus be verified by inspecting the elements d_{ij} of $\mathbf{D}$. For verifying RR of $\mathbf{X}$, we use the fit index

$$\mathrm{RF} = \frac{\sum_{i=1}^{n-1}\sum_{i=1}^{m-1} d_{ij}}{\sum_{i=1}^{n-1}\sum_{i=1}^{m-1} \left|d_{ij}\right|}.$$

Index RF consists of the sum of all (relevant) elements of $\mathbf{D}$ divided by the sum of all absolute differences. Note that the elements of the last row and column of $\mathbf{D}$ are not included. The last row of $\mathbf{D}$ does not contain differences. Furthermore, the first $n-1$ elements of the last column of $\mathbf{D}$ are all zero and contain no information.

Fit index RF is an intuitively reasonable but nevertheless arbitrary measure and many alternative measures can be defined using the elements of $\mathbf{D}$. However, the current fit index has some appealing properties. Its maximum value of unity is obtained if $\mathbf{X}$ is RR. A value of minus unity is obtained if all d_{ij} have negative signs, which implies a reversed ordering. High positive values of RF are obtained if there are only small deviations from regression dependence but also large d_{ij} with positive sign. Values close to zero are obtained if the amount of differences with negative sign is approximately equal to the amount of differences with positive sign. Note that this case is a reasonable null model for RF since we are far from any form of positive dependence.

Table 3
Data tables from the literature with corresponding fit indices RF and CF

Data	Source	RF	CF
Occupational mobility	Glass (1954), Gifi (1990, p. 278)	1.00	1.00
Eye and hair color	Fisher (1940), Maung (1941), Greenacre (1984, p. 256)	0.97	1.00
Smoking	Greenacre (1984, p. 55)	1.00	0.87
Student population census	Benzécri (1992, p. 535)	0.84	1.00
Spot patterns	Guilford (1954, p. 203), Gifi (1990, p. 335)	0.84	1.00
Political party groups	Van Schuur (1984, p. 91)	0.33	1.00
Nuclear power	Formann (1988)	0.21	1.00
Brazil's import	Benzécri (1992, p. 424)	0.44	0.99
Readership	Greenacre (1984, p. 4)	0.40	0.98
Religious practice	Sugiyama (1975), Gifi (1990, p. 291)	0.14	0.98
Developmental processes	Leik and Matthews (1968)	0.29	0.97
Scientific disciplines	Greenacre (1993, p. 75)	0.30	0.96

Denote by $\mathbf{C}$ the diagonal matrix of order m with the column totals x_{+j} of $\mathbf{X}$ on its main diagonal. Furthermore, let $\mathbf{S}_n$ be the lower triangular matrix of order n with elements 1 on and below its main diagonal and elements 0 above its diagonal. Similar to $\mathbf{D}$ and RF for verifying RR of $\mathbf{X}$, we define $\mathbf{E} = \mathbf{S}_n\mathbf{X}\mathbf{C}^{-1}\mathbf{S}_m^{-1}$ and

$$\mathrm{CF} = \frac{\sum_{i=1}^{n-1}\sum_{i=1}^{m-1} e_{ij}}{\sum_{i=1}^{n-1}\sum_{i=1}^{m-1} |e_{ij}|}$$

for verifying CR of $\mathbf{X}$. Measure CF possesses the same properties as index RF. Some applications of the two fit indices RF and CF are presented in the next section.

4. Applications

Schriever (1983) noted that data tables that are RR or CR are encountered in various domains of data analysis. He presented examples of discretized bivariate distributions to support this claim. If a contingency table is either RR or CR, or both, there exists a strong ordinal association between the categories of I and J, i.e the rows and columns. If the data table is CR, the mass of the conditional distribution functions $I|J = j$ concentrates on the row categories with larger indices as the index j of the column categories gets higher. The conditional distribution functions do not cross each other and can be ordered perfectly.

Table 3 consists of a list of data tables from the CA literature. The data tables are either subject by attribute matrices or cross-classifications of subjects. For each example we have re-ordered the row and columns using the category scores of the first CA dimension. For each re-ordered data table we have provided the fit indices RF and CF in Table 3, and conclude that regression dependence is (approximately) present in the variable with the smallest number of categories. The first seven re-ordered data tables listed in Table 3 are strictly RR or strictly CR. The last five data examples in Table 3 are approximately CR. We discuss the first three examples in Table 3 in more detail.

Glass (1954) reported a 7×7 table of occupational status of fathers versus occupational status of sons for a sample of 3497 British families (see also Gifi (1990, p. 278)). For these occupational mobility data, the first CA dimension (69% variance explained) assigns the same order to both father and son categories: professional and high administrative, managerial and executive, higher supervisory, lower supervisory, skilled manual and routine nonmanual, semi-skilled manual, and unskilled manual labor. The first CA dimension orders the categories of occupational status from high to low. Index RF $= 1.00$ and index CF $= 1.00$, indicating that the table is both RR and CR.

Fisher (1940) and Maung (1941) reported a contingency table of 5387 schoolchildren from Caithness, Scotland, classified according to the two nominal variables, eye color and hair color (see also Greenacre (1984, p. 256). The first CA dimension (87% variance explained) orders the $n = 4$ eye color categories as, dark eyes, medium eyes, blue eyes, and light eyes, and orders the $m = 5$ hair color categories as black hair, dark hair, medium hair, red hair, and fair hair. The first CA dimension can be interpreted as dark to light/fair dimension of color. Index RF $= 0.97$ and index CF $= 1.00$, indicating that the re-ordered table is CR, and RR for all practical purposes.

Greenacre (1984, p. 55) reported an artificial table of different types of smokers in a sample of personnel from a fictitious organization. The first CA dimension (88% variance explained) orders the $m = 4$ smoking categories as expected (heavy, medium, light, and non smoking), and orders the $n = 5$ personnel categories as follows: junior managers, junior employees, senior managers, secretaries, and senior employees. Index RF $= 0.87$ and CF $= 1.00$, indicating that the conditional distribution functions of the smoking categories can be ordered perfectly. The first CA dimension can be interpreted as a heavy to non smoking dimension: the amount of smoking depends on seniority and the type of employee.

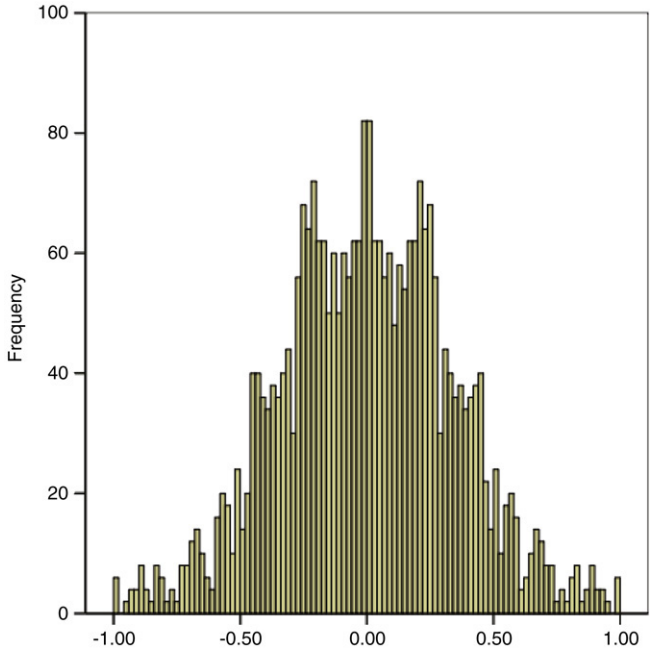

Fig. 1. Distribution of the values of the column fit index for all possible (2880) rearrangements of the rows and columns of Table 1.

5. Discussion

In this paper fit indices were defined that may be used to verify whether the re-ordered contingency table is row regression dependent or column regression dependent. Both forms of bivariate dependence imply that there exists a strong ordinal association between the categories of variables I and J. The usefulness of the fit indices was tested on several data examples. We may conclude that the interpretative value of an ordered solution can indeed be large. The data-analyst that works within the boundaries of the usual geometric interpretation of correspondence analysis, may use these indices as complementary statistics.

The fit measures defined in this paper are arbitrary, and a variety of alternative (possibly weighted) coefficients may be formulated. The measures defined here were only used to verify whether the re-ordered contingency table is (approximately) regression dependent or not. We have only used values $\geq$0.96, and have not attempted to interpret lower values. Statistical testing can, e.g., be incorporated by using permutation tests. For example, we can obtain 2880 ($=5! \times 4! = 120 \times 24$) different versions of Table 1 by considering all rearrangements of the rows and columns. For each variant, we may calculate the values of the fit measures to obtain a reference distribution under the hypothesis that all orders are equally likely. For the column fit index (CF) these values are plotted in Fig. 1. The reference distribution in Fig. 1 is symmetric and shows that the obtained value of unity, CF $= 1.00$, is rare.

In this paper, we focused on regression dependence, because it was related to correspondence analysis by Schriever (1983). For other forms of bivariate dependence, other diagnostics, and ways of testing positive dependence, see e.g. Cohen et al. (2006) or Kimeldorf and Sampson (1989).

References

Agresti, A., 2002. Categorical Data Analysis. Wiley, New York.
Benzécri, J.-P., 1973. Analyse de Données. Dunod, Paris.
Benzécri, J.-P., 1992. Correspondence Analysis Handbook. Marcel Dekker, New York.
Blasius, J., Greenacre, M.J., 2006. Correspondence analysis and related methods in practice. In: Greenacre, M.J., Blasius, J. (Eds.), Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC, Boca Raton, pp. 3–40.
Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling. Theory and Applications, 2nd edn. Springer, New York.
Cohen, A., Kolassa, J., Sackrowitz, H.B., 2006. A new test for stochastic order of $k \geq 3$ ordered multinomial populations. Statistics & Probability Letters 76, 1017–1024.
Fisher, R.A., 1940. The precision of discriminant functions. Annals of Eugenics 10, 422–429.
Formann, A.K., 1988. Latent class models for nonmonotone dichotomous items. Psychometrika 53, 45–62.
Gifi, A., 1990. Nonlinear Multivariate Analysis. Wiley, Chicester.
Gilula, Z., Ritov, Y., 1990. Inferential ordinal correspondence analysis: Motivation, derivation, and limitations. International Statistical Review 58, 99–108.
Glass, D.V., 1954. Social Mobility in Britain. Free Press, Glencoe.
Goodman, L.A., 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. International Statistical Review 54, 243–309.
Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. Academic Press, New York.
Greenacre, M.J., 1993. Correspondence Analysis in Practice. Academic Press, London.

Guilford, J.P., 1954. Psychometric Methods, 2nd edn. McGraw-Hill, New York.
Heiser, W.J., 1981. Unfolding analysis of proximity data. Ph.D. Thesis, Leiden University.
Kimeldorf, G., Sampson, A.R., 1989. A framework for positive dependence. Annals of the Institute of Statistical Mathematics 41, 31–45.
Lehmann, E.L., 1966. Some concepts of dependence. Annals of Mathematical Statistics 37, 1137–1153.
Leik, R.K., Matthews, M.A., 1968. A scale for developmental processes. American Sociological Review 33, 62–75.
Maung, K., 1941. Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. Annals of Eugenics 11, 189–223.
Ritov, Y., Gilula, Z., 1993. Analysis of contingency tables by correspondence models subject to order constraints. Journal of the American Statistical Association 88, 1380–1387.
Schriever, B.F., 1983. Scaling of order dependent categorical variables with correspondence analysis. International Statistical Review 51, 225–238.
Sugiyama, M., 1975. Religious behavior of the Japanese: Execution of a partial order scalogram analysis based on quantification theory. In: US–Japan Seminar of Multidimensional Scaling and Related Techniques, La Jolla.
Tong, Y.L., 1980. Probability Inequalities in Multivariate Distributions. Academic Press, New York.
Van Rijckevorsel, J., 1987. The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis. DSWO Press, Leiden.
Van Schuur, W.H., 1984. Structure in Political Beliefs. A New Model for Stochastic Unfolding With Application to European Party Activists. CT Press, Amsterdam.