Biometrika (1991), 78, 1, *pp.* 229–32
Printed in Great Britain

Reduced rank models for contingency tables

By JAN DE LEEUW
Departments of Psychology and Mathematics, University of California, Los Angeles, California 90024-1563, U.S.A.
AND PETER G. M. VAN DER HEIJDEN
Department of Empirical and Theoretical Sociology, University of Utrecht, 3508 TC Utrecht, The Netherlands

Summary

Reduced rank models for the analysis of two-way contingency tables are introduced. Two classes of reduced rank models are discerned, with well-known exponents canonical analysis and latent class analysis. The relation between these two classes is discussed. Results on the subject mentioned earlier in the literature are shown to be either redundant or inaccurate.

Some key words: Canonical analysis; Correspondence analysis; Latent class analysis; Reduced rank models.

1. Introduction

In recent years much attention has been given to models for two-way contingency tables that can be formulated in terms of reduced rank of a matrix with probabilities. A well-known reduced rank model is the independence model, where the rank is one. For rank higher than one distinct classes of reduced rank models are possible. Each has the independence model as the special case for rank one. A first class of such models is closely related to what is known under names as canonical analysis or correspondence analysis. Recently much attention has been given to the maximum likelihood estimation of versions of these models by Goodman (1985, 1986, 1987) and Gilula & Haberman (1986, 1988). A second class of models that can be formulated in terms of reduced rank is latent class analysis, LCA, for two-way tables. Latent class analysis was proposed by Lazersfeld (1950a, b). See Clogg (1981) for a more recent review.

In this paper we relate these classes of models to each other. The relation has been discussed earlier by Gilula (1979, 1983, 1984), Gilula & Haberman (1986), Goodman (1987), and van der Heijden, Mooijaart & de Leeuw (1989). We summarize existing results in a simple way using new proofs. Gilula (1979) provided conditions that had to hold for rank-2 correspondence analysis to imply rank-2 latent class analysis. We show here that rank-2 correspondence analysis always implies rank-2 latent class analysis. This implies that the theorem and the example given by Gilula (1979) are incorrect.

2. General reduced rank models

The basic model studied in this paper assumes that an $n \times m$ probability matrix Π has rank ρ, where $\rho \leq \min(n, m)$. We call this model R_ρ. The probability matrix Π has all elements nonnegative, while the sum of the π_{ij} is equal to one. We suppose, unless indicated otherwise, that Π is full, in the sense that its row sums π_{i+} and its column sums π_{+j} are all positive. Thus no row or column is equal to zero.

We compare this model with the canonical model C_ρ, in which at most $\rho - 1$ of the canonical correlations between the row and the column variables of the table are nonzero. These canonical correlations are the stationary values of the product moment correlation coefficient, seen as a function of scores for rows and scores for columns.

We also compare R_ρ with the model suggested by correspondence analysis, written as A_ρ, in which

$$\pi_{ij} = \omega_i \theta_j \left(1 + \sum_{s=1}^{\rho-1} \lambda_s x_{is} y_{js}\right).$$

Another way of formulating A_ρ is by saying that Π has a Fisher-decomposition of rank $\rho - 1$ (Lancaster, 1958).

THEOREM 1. *We have that C_ρ, A_ρ and R_ρ are equivalent.*

Proof. If $\rho - 1$ canonical correlations are nonzero, then Π can be written in the form A_ρ, with ω_i and θ_j equal to the marginals π_{i+} and π_{+j}, with λ_s equal to the canonical correlations, and with x_{is} and y_{js} equal to the canonical scores (Lancaster, 1958). Thus C_ρ implies A_ρ. It is obvious, moreover, that A_ρ implies R_ρ. We now prove that R_ρ implies C_ρ. Suppose rank $(\Pi) = \rho$. By the Lagrange theorem, for instance Guttman (1944), we know that $\pi_{ij} - \pi_{i+}\pi_{+j}$ has rank exactly equal to $\rho - 1$ and is doubly centred. The canonical correlations are computed from the singular value decomposition of the matrix of normalized residuals Z, given by

$$z_{ij} = \frac{\pi_{ij} - \pi_{i+}\pi_{+j}}{\surd(\pi_{i+}\pi_{+j})}.$$

The matrix Z is of rank $\rho - 1$, and thus has $\rho - 1$ nonzero canonical correlations.

3. Reduced rank models with nonnegativity constraints

Let us now look at the model R_ρ^*, which assumes that rank $(\Pi) = \rho$, and moreover that there exists a full rank decomposition $\Pi = AB'$, with $A \geqslant 0$ and $B \geqslant 0$. Clearly R_ρ^* implies R_ρ, but in general the reverse implication is not true, at least not obvious.

There are some interesting alternative ways to write R_ρ^*. In the first place the latent class model LCA_ρ, mentioned by Good (1965), is such that Π is a mixture of ρ bivariate distributions with independence. Thus

$$\pi_{ij} = \sum_{s=1}^{\rho} \eta_s \alpha_{is} \beta_{js},$$

with $\eta_+ = \alpha_{+s} = \beta_{+s} = 1$. Moreover all parameters are nonnegative. There is also the latent budget model LBA_ρ (van der Heijden et al., 1989), in which

$$\frac{\pi_{ij}}{\pi_{i+}} = \sum_{s=1}^{\rho} \alpha_{is} \beta_{js},$$

with $\alpha_{i+} = \beta_{+s} = 1$, and again all parameters are nonnegative.

THEOREM 2. *We have that R_ρ^*, LBA_ρ and LCA_ρ are equivalent.*

Proof. Suppose Π satisfies R_ρ^*. Thus $\Pi = AB'$, with $A \geqslant 0$ and $B \geqslant 0$. Suppose Φ is a diagonal matrix of order ρ, with the b_{+s} on the diagonal. Let $\tilde{A} = A\Phi$ and $\tilde{B} = B(\Phi^{-1})'$. Then clearly $\Pi = \tilde{A}\tilde{B}'$. Moreover $\tilde{b}_{+s} = 1$ and $\tilde{a}_{i+} = \pi_{i+}$. If we define $\beta_{js} = \tilde{b}_{js}$ and $\alpha_{is} = \tilde{a}_{is}/\tilde{a}_{i+}$, then we satisfy LBA_ρ. Let $\beta_{js} = \tilde{b}_{js}$ and $\alpha_{is} = \tilde{a}_{is}/\tilde{a}_{+s}$, and $\eta_s = \tilde{a}_{+s}$. These quantities satisfy LCA_ρ. Thus R_ρ^* implies LBA_ρ, and LBA_ρ and LCA_ρ imply each other. It is trivial that LCA_ρ implies R_ρ^*. □

4. Existence of nonnegative decompositions

As we said above, in general R_ρ^* implies R_ρ, but the reverse implication is not necessarily true. The relationship between these models was already mentioned by Good (1965, p. 64), and studied by Gilula (1979, 1983, 1984).

THEOREM 3. *We have that R_2 and R_2^* are equivalent.*

Proof. We know that R_2^* implies R_2, so we merely have to prove the reverse. Because of R_2 the columns of Π are m vectors in a two-dimensional subspace of R^n. Because all columns are nonnegative they are actually in a pointed cone in this plane. Two-dimensional cones are simplicial, i.e. they have exactly two extreme rays. The bundle of rays corresponding with the columns of Π has two extremes, all other columns are positive linear combinations of these two columns. But this means that R_2^* is true, with A equal to these two extreme columns. □

This very simple geometric proof is due to Paul Bekker. It replaces a lengthy computational proof we first had, and a complicated algebraic proof by Thomas (1974) we subsequently discovered. Thomas (1974) also gave a necessary and sufficient condition for R_p to imply R_p^*, which reformulates the problem in terms of the existence of certain polyhedral convex cones. He also provided the counterexample

$$\begin{bmatrix} 0\cdot125 & 0\cdot125 & 0\cdot0 & 0\cdot0 \\ 0\cdot125 & 0\cdot0 & 0\cdot125 & 0\cdot0 \\ 0\cdot0 & 0\cdot125 & 0\cdot0 & 0\cdot125 \\ 0\cdot0 & 0\cdot0 & 0\cdot125 & 0\cdot125 \end{bmatrix}.$$

This matrix satisfies R_3, but not R_3^*.

It follows from our result that the Theorem and Corollary 1 of Gilula (1979) are not correct. This result also shows that van der Heijden et al. (1989) are incorrect in stating that latent class analysis and correspondence analysis are always equivalent, i.e. for any rank p.

The example Gilula (1979) gives is supposed to satisfy R_2 and not R_2^*. The probability matrix is

$$\begin{bmatrix} 0\cdot165 & 0\cdot005 & 0\cdot030 \\ 0\cdot015 & 0\cdot580 & 0\cdot105 \\ 0\cdot020 & 0\cdot065 & 0\cdot015 \end{bmatrix}.$$

In this example the first two columns of Π are the extreme columns, and thus columns π_1, π_2 and π_3 satisfy the relationship $\pi_3 = \frac{3}{17}(\pi_1 + \pi_2)$. Consequently

$$\Pi = \begin{bmatrix} 0\cdot165 & 0\cdot005 \\ 0\cdot015 & 0\cdot580 \\ 0\cdot020 & 0\cdot065 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0\cdot1765\ldots \\ 0 & 1 & 0\cdot1765\ldots \end{bmatrix},$$

which counters Gilula's counterexample.

ACKNOWLEDGEMENTS

We are grateful to P. Bekker and J. F. B. M. Kraaijevanger for helpful comments.

REFERENCES

CLOGG, C. C. (1981). Latent structure models of mobility. *Am. J. Sociol.* **86**, 836–68.

GILULA, Z. (1979). Singular value decomposition of probability matrices: probabilistic aspects of latent dichotomous variables. *Biometrics* **66**, 339–44.

GILULA, Z. (1983). Latent conditional independence in two-way contingency tables: a diagnostic approach. *Br. J. Math. & Statist. Psychol.* **36**, 114–22.

GILULA, Z. (1984). On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika* **71**, 523–9.

GILULA, Z. & HABERMAN, S. J. (1986). Canonical analysis of contingency tables by maximum likelihood. *J. Am. Statist. Assoc.* **81**, 780–8.

GILULA, Z. & HABERMAN, S. J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Am. Statist. Assoc.* **83**, 760–71.

GOOD, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Cambridge, Mass: MIT Press.

GOODMAN, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.* **13**, 10–69.

GOODMAN, L. A. (1986). Some useful extensions to the usual correspondence analysis approach and the usual loglinear approach in the analysis of contingency tables (with comments). *Int. Statist. Rev.* **54**, 243–309.

GOODMAN, L. A. (1987). New methods for analyzing the intrinsic character of qualitative variables using cross-classified data. *Am. J. Sociol.* **93**, 529–83.

GUTTMAN, L. (1944). General theory and methods for matrix factoring. *Psychometrika* **9**, 1–16.

LANCASTER, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* **29**, 719–36.

LAZERSFELD, P. F. (1950a). The logical and mathematical foundations of latent structure analysis. In *Measurement and Prediction*, Ed. S. A. Stouffer et al., pp. 362–412. Princeton University Press.

LAZERSFELD, P. F. (1950b). The interpretation and computation of some latent structures. In *Measurement and Prediction*, Ed. S. A. Stouffer et al., pp. 413–72. Princeton University Press.

THOMAS, L. B. (1974). Solution problem 73–14, Rank factorization of nonnegative matrices, by A. Berman and R. J. Plemmons. *SIAM Rev.* **16**, 393–4.

VAN DER HEIJDEN, P. G. M., MOOIJAART, A. & DE LEEUW, J. (1989). Latent budget analysis. In *Statistical Modelling. Proceedings of GLIM 89 and the 4th International Workshop on Statistical Modelling*, Ed. A. Decarli, B. J. Francis, R. Gilchrist and G. U. H. Seeber, Lecture Notes in Statistics 57, pp. 301–13. Berlin: Springer-Verlag.

[*Received August* 1989. *Revised June* 1990]