CORRESPONDENCE ANALYSIS OF INCOMPLETE CONTINGENCY TABLES

JAN DE LEEUW
DEPARTMENT OF DATA THEORY
UNIVERSITY OF LEIDEN

PETER G. M. VAN DER HEIJDEN
DEPARTMENT OF PSYCHOMETRICS
UNIVERSITY OF LEIDEN

Correspondence analysis can be described as a technique which decomposes the departure from independence in a two-way contingency table. In this paper a form of correspondence analysis is proposed which decomposes the departure from the quasi-independence model. This form seems to be a good alternative to ordinary correspondence analysis in cases where the use of the latter is either impossible or not recommended, for example, in case of missing data or structural zeros. It is shown that Nora's reconstitution of order zero, a procedure well-known in the French literature, is formally identical to our correspondence analysis of incomplete tables. Therefore, reconstitution of order zero can also be interpreted as providing a decomposition of the residuals from the quasi-independence model. Furthermore, correspondence analysis of incomplete tables can be performed using existing programs for ordinary correspondence analysis.

Key words: Correspondence analysis, data analysis, quasi-independence, structural zeros.

1. Introduction

In this paper we introduce a modification of correspondence analysis (CA) which can be used in combination with the quasi-independence models familiar from loglinear analysis. The technique we propose decomposes the residuals that are left after fitting a quasi-independence model. The decomposed residuals are represented geometrically. Thus our paper interprets CA as a technique which can be used complementary to loglinear analysis. A similar approach has been adopted by Daudin and Trécourt (1980), Isaëls and Sikkel (1982), Lauro and Decarli (1982), and Caussinus and de Falguerolles (1986). It was also suggested by Aitkin (discussion of Deville & Malinvaud, 1983). CA can also be introduced as a model in its own right, or as an approximation to existing models. This is the approach taken by Goodman (1985, 1986), for example.

The French approach to CA, originated by Benzécri (1973, 1980), and described in considerable detail by Greenacre (1984), interprets CA as a multidimensional scaling technique which makes pictures of data matrices. In this presentation no statistical model is involved. Although we think that this geometrical interpretation of CA is in many cases the most natural one, we also think that combination and comparison with current statistical modeling approaches for frequency tables is quite useful. This is illustrated in van der Heijden and de Leeuw (1985) and van der Heijden and Worsley (1988). In this complementary interpretation of CA we study it as a technique to represent residuals of a
loglinear analysis in a picture. Both the geometrical and the statistical aspects are present in this approach, but clearly the statistics are predominant. We only apply CA to the variation that is left after the model is fitted. A model with a good fit leaves very little variation, and thus CA will be quite useless in such cases. This is more or less true by definition: A model fits well if there is no systematic variation in the residuals. As a consequence CA is most useful in combination with models that do not fit well. Thus we must combine the use of CA with the use of fairly restrictive models. This agrees closely with recommendations made by Aitkin: "CA would be particularly useful when considerable structure remains, as indicated by a large deviance, but no useful explanatory variables are available. The component plots may help identify the nature of the structure and other variables which should have been measured" (discussion of Deville & Malinvaud, 1983, p. 357). Ordinary CA is, in our interpretation, complementary to the complete independence model, which is of course highly restrictive.

We shall make use of a generalization of CA to decompose residuals from the quasi-independence model. It is supposed in this paper that the reader is familiar with the theory and applications of quasi-independence models for two-way tables. We merely indicate our notation. The model states that the theoretical probabilities \( \pi_{ij} \) in a bivariate contingency table satisfy \( \pi_{ij} = \alpha_i \beta_j \) for a subset \( S \) of all index pairs \( (i,j) \). There are various reasons why we may not want to require \( \pi_{ij} = \alpha_i \beta_j \) for all pairs. The first one is that some elements of the table are missing. A second one is that some elements may be zero by definition, the so-called structural zeros. Thirdly we may know from a previous analysis that some cells fit the independence model badly. And finally we may have the idea that for some parts of the table the independence model may be true, while for other parts (for instance the diagonal) independence is not plausible at all. For a thorough discussion we refer to Caussinus (1965), Mosteller (1968), Goodman (1968), Bishop, Fienberg and Holland (1975, pp. 177–210), and Haberman (1979, pp. 444–486).

2. Correspondence Analysis

CA will be discussed briefly here. For a longer discussion from a comparable perspective we refer to van der Heijden and de Leeuw (1985). In order to discuss correspondence analysis (CA) of incomplete tables later, we first define ordinary CA in terms of the Fisher-Lancaster decomposition of an observed table. This is sometimes called the canonical analysis of a contingency table (for instance in Kendall & Stuart, 1967, chap. 33), while the French call it the reconstitution formula. Suppose \( P \) is the observed table, with positive entries that add up to one. The diagonal matrix \( D_r \) contains row marginals, \( D_c \) contains the column marginals, \( t \) is a vector with all elements equal to one. Then we can find \( R \) and \( C \) such that \( t'D_r R = 0, t'D_c C = 0, R'D_r R = I, C'D_c C = I, \) and

\[
P = D_r (t' + RAC')D_c,
\]

with \( A \) diagonal. The sum of squares of the elements of \( \Lambda \) is equal to Pearson's index of mean square contingency. If \( P \) is based on a sample of size \( n \), then \( n \) times this index is equal to the chi-square statistic for testing independence. Thus we can say that CA, if interpreted as computing the Fisher-Lancaster decomposition (1), studies the deviations from the independence model.

In the introduction we said that CA gives a geometrical representation of the residuals, in this case of the residuals from independence. This can be explained most easily by introducing the chi-square distances between the rows of \( D_r^{-1}P \). Rows of \( D_r^{-1}P \) add up to 1, and are usually referred to as profiles (Benzécri, 1973, 1980; Greenacre, 1984). The