MTISD 2008 – Methods, Models and Information Technologies for Decision Support Systems
Università del Salento, Lecce, 18-20 September 2008

Model Based Methods for Performing Multi-Way Non-Symmetric Correspondence Analysis

Eric J. Beh
University of Western Sydney, Australia
E.Beh@uws.edu.au

Rosaria Lombardo
Second University of Naples, Capua, Italy
Rosaria.lombardo@unina2.it

Biagio Simonetti
University of Sannio, Italy
simonett@unisannio.it

Abstract: Traditionally multiple correspondence analysis involves transforming a contingency table to its indicator or Burt matrix form then performing the classical two-way approach. Alternatively, one may also consider techniques that are more model based such as the partitions associated with the PARAFAC/CANDECOMP models or the Tucker3 model. Traditionally these procedures have proven to be of benefit in studies where the variables are nominally structured. This paper will demonstrate how they can be adapted for ordinal variables by incorporating orthogonal polynomials into the partitions and a graphical description of the association can be obtained by considering correspondence analysis. This paper will also consider the case where the variables of a three-way contingency table are asymmetrically associated.

Keywords: Multi-way contingency tables, PARAFAC/CANDECOMP, Tucker3, Orthogonal Polynomials

1 Introduction

Correspondence analysis has proven to be a rich and abundant area of research for data analysts. The purpose of considering such an analytic tool is to understand the association structure between two or more categorical variables that form a contingency table. This is commonly achieved by applying singular value decomposition (SVD) to a transformation of the cells of the contingency table. For two-way contingency tables SVD is applied to the matrix of Pearson residuals, or alternatively Pearson ratio's (depending on the scaling involved). For multiple categorical variables, a multi-way contingency table can be transformed into its indicator matrix form, or its Burt matrix form, both of which give a two-way matrix where the traditional approach to correspondence analysis can be performed.

An alternative strategy to performing multiple correspondence analysis, and one that is largely applicable for the analysis of three categorical variables, is to consider a more model based approach to categorical data analysis. For example, one may consider the partition of cell transformations based on the Tucker3 model, or the CANDECOMP/PARAFAC models. Such procedures have commonly been applied in cases where it is assumed that the variables being studied are symmetrically associated. For asymmetric variables, variations of these partitions need to be considered. This paper will discuss the role of some of these procedures for asymmetrically related variables, specifically focusing on its link with the Marcotorchino index (Marcotorchino, 1985; Lombardo, Carlier, D'Ambra, 1996; Simonetti, Beh, D'Ambra, 2006). Such a procedure can

be modified by incorporating orthogonal polynomials that reflect the structure of ordinal categorical variables.

2 Classical Correspondence Analysis

Consider an $I \times J$ two-way contingency table, N, where the (i, j)'th cell entry is given by n_{ij} for i= 1, 2, . . . , I and j= 1, 2, . .. , J. Let the grand total of N be n and the (i, j)'th cell of relative frequencies be denoted by $p_{ij} = n_{ij} / n$. Define the i.th row marginal proportion by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and define the j'th column marginal proportion as $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$. Let $\alpha_{ij} = p_{ij} / p_{i\bullet} p_{\bullet j}$ be the Pearson ratio of the (i, j)'th cell. Beh (2004) described that classical correspondence analysis can be performed by applying singular value decomposition to the Pearson ratio by

$$\alpha_{ij} = \sum_{m=1}^{M^*} a_{im} \lambda_m b_{jm} \tag{1}$$

where $M^* = \min(I, J) - 1$ is the maximum number of dimensions required to graphically depict the association between the responses of the two variables. Here, $\mathbf{a}_m$ is the m.th row singular vector associated with the I row categories while $\mathbf{b}_m$ is the m.th column singular vector associated with the J column categories, they are orthonormal with respect to the weighted metrics $D_I = \{p_{i\bullet}\}$ and $D_J = \{p_{\bullet j}\}$, respectively. The elements of $\lambda_m = (1, \lambda_1, \ldots, \lambda_{M^*})$ are the singular values arranged in descending order such that the Pearson chi-squared statistic can be partitioned so that $X^2 = n \sum_{m=1}^{M^*} \lambda_m^2$.

3 Model Based Correspondence Analysis of Three-way Contingency Tables

One limitation of viewing correspondence analysis from the perspective given in section 2 is that, since SVD is used as the central tool for dimension reduction, this approach is limited in its ability to consider more than two categorical variables without some modification to the data (Greenacre, 1984).

To avoid this problem one may consider a three-way extension of the SVD given by (1). Suppose we consider performing a multiple correspondence analysis on a three-way contingency table where the first two variables are those defined in section 2, and the third variable, consisting of K tubes, is denoted in the notation with a subscript k, or a "$\bullet$" when considering the summation of a measure over this variable. Therefore, complete independence between the all three variables will occur when

$$p_{ijk} = p_{i\bullet\bullet} p_{\bullet j\bullet} p_{\bullet\bullet k}$$

where the three-way Pearson ratio is defined by $\alpha_{ijk} = p_{ijk} / (p_{i\bullet\bullet} p_{\bullet j\bullet} p_{\bullet\bullet k})$. Here we shall consider some decompositions of the three-way Pearson ratio (Carlier and Kroonenberg, 1996).

3.1 The CANDECOMP/PARAFAC & Tucker3 Models

One way to extend the idea of what singular value decomposition does for allowing the researcher to reduce the number of dimensions required to visualise multivariate associations is to consider the two most common three-way SVD. The first one is the CANDECOMP (CANonical DECOMPosition) model of Harshman (1970) or the mathematically equivalent PARAFAC (PARAllel FACtor analysis) model of Carroll and Chang (1970). These models allow for the three-way Pearson ratio to be decomposed such that

$$\alpha_{ijk} = \frac{p_{ijk}}{p_{i\bullet\bullet} p_{\bullet j\bullet} p_{\bullet\bullet k}} = \sum_{m=0}^{M^*} a_{im} b_{jm} c_{km} \lambda_{mmm} \tag{2}$$

where λ_{mmm} is the three-way analogues of the singular value λ_m; $M^* = \min(I, J, K) - 1$, the row scores $\{a_{im}\}$, the column scores $\{b_{jm}\}$ and the tube scores $\{c_{km}\}$ are assumed to have unit lengths, where orthonormality is defined with respect to diagonal metrics whose general elements are given by marginal proportions $\{p_{i\bullet\bullet}\}$, $\{p_{\bullet j\bullet}\}$ and $\{p_{\bullet\bullet k}\}$, respectively.

Note that (2) is an extension of SVD (1) for the case where the association between three categorical variables is of interested. Thus, the decomposition (2) has been referred to as a "generalised three-way singular value decomposition" (Carlier and Kroonenberg, 1996).

The second alternative approach to performing correspondence analysis on a three-way contingency table is to consider the partition considered by psychometrician Ledyard R. Tucker in 1963. For the three-way Pearson ratio, the Tucker3 approach consists in decomposing the ratio as

$$\begin{aligned}\alpha_{ijk} &= \sum_{u=0}^{I-1}\sum_{v=0}^{J-1}\sum_{w=0}^{K-1} a_{iu} b_{jv} c_{kw} \lambda_{uvw} \\ &= 1 + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1} a_{iu} b_{jv} \lambda_{uv0} + \sum_{u=1}^{I-1}\sum_{w=1}^{K-1} a_{iu} c_{kw} \lambda_{u0w} + \sum_{v=1}^{J-1}\sum_{w=1}^{K-1} b_{jv} c_{kw} \lambda_{0vw} + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} a_{iu} b_{jv} c_{kw} \lambda_{uvw}\end{aligned} \tag{3}$$

where the $\{\lambda_{uvw}\}$ are another generalization of the singular values, known as elements of the core matrix (Kroonenberg, 1989). As in the PARAFAC model, the row scores $\{a_{im}\}$, the column scores $\{b_{jm}\}$ and the tube scores $\{c_{km}\}$ are subject to the weighted orthonormality constraints. For three orthogonal models it can be shown that the Pearson chi-squared statistic of the three-way contingency table may be so partitioned so that $X^2 = n\sum_{m=1}^{M^*} \lambda_{mmm}^2$ for the PARAFAC model and $X^2 = n\sum_{uvw} \lambda_{uvw}^2$ for the Tucker3 model.

4 Ordered Categorical Variables and Asymmetry

In many studies, categorical variables will consist of ordered responses (eg. income or age bracket, measurements on a Likert scale). However such a structure is not taken into consideration using the above methods. One may consider instead using orthogonal polynomials (Emerson, 1968) which are generated using scores to reflect the ordinal structure of the variable. When the row, column and tube variables are ordered, Beh and Davy (1998; 1999) showed that measures analogous to α_{ijk} can be partitioned using the polynomials. Suppose that our three-way contingency table, **N**, consists of ordered column categories and this ordinal structure is reflected by the set of column scores s(*j*): *j*=1,…,*J*.

Setting $\mathbf{b}^*_{(-1)}(v)=0$ and $\mathbf{b}^*_{(0)}(v)=1$, and using the natural set of scores, the column orthogonal polynomials of generic degree *v* (for *v=1,…,J-1*) are calculated using the following general recurrence formula

$$b^*_{j(v)} = A_v[(j - B_v) b^*_{j(v-1)} - C_v b^*_{j(v-2)}]$$

where

$$B_v = \sum\nolimits_{j=1}^{J} p_{\bullet j\bullet}\, j\, b^{*2}_{j,v-1}$$

$$C_v = \sum\nolimits_{j=1}^{J} p_{\bullet j\bullet}\, j\, b^{*}_{j,v-1} b^{*}_{j,v-2}$$

$$A_v = \left\{\sum\nolimits_{j=1}^{J} p_{\bullet j\bullet}\, j^2 b^{*2}_{j,v-1} - B_v^2 - C_v^2\right\}^{-1/2}$$

These Emerson polynomials are orthogonal with respect to the marginal proportion $p_{\bullet j \bullet} = \sum_{i=1}^{I} \sum_{k=}^{K} p_{ijk}$. When the row and tube variables are ordinal, we can compute the polynomials $\mathbf{a}^{*}_{i(u)}$ of order *u* (for *u=1,…,I-1*) and $\mathbf{c}^{*}_{k(w)}$ (for *w=1,…,K-1)* in a similar manner.

In this context, the Tucker3 partition, (3), can be generalised by considering the case where the categorical variables are asymmetrically associated – that is, where one set of categories for a response variable is dependent on two predictor category sets. Rather than considering the Pearson chi-squared statistic, such an asymmetric structure can be measured using the Marcotorchino index (Marcotorchino, 1985; Lombardo, Carlier, D'Ambra, 1996; Simonetti, D'Ambra and Beh, 2006) that is based on the differences

$$\alpha_{ijk} = \frac{p_{ijk}}{p_{\bullet j \bullet} p_{\bullet \bullet k}} - p_{i \bullet \bullet}$$

While Beh, Simonetti and D'Ambra (2007) considered the partition of the Marcotorchino index for ordinal categorical variables using orthogonal polynomials, in this paper we will explore the use of orthogonal polynomials with the Tucker3 decomposition in order to depict the asymmetric association between three categorical variables. It will also demonstrate how correspondence analysis can be adapted to provide a graphical interpretation of this asymmetric measure.

Bibliography

Beh, E. J. (2004), Simple correspondence analysis: A bibliographic review, *International Statistical Review*, 72, 257-284.

Beh, E. J., Simonetti, B. and D'Ambra, L. (2007) Partitioning a non-symmetric measure of association for three-way contingency tables, *Journal of Multivariate Analysis*, 98, 1391–1411.

Beh, E. J. and Davy, P. J. (1998), Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table, *The Australian and New Zealand Journal of Statistics*, 40, 465-477.

Beh, E. J. and Davy, P. J. (1999), Partitioning Pearson's chi-squared statistic for a partially ordered three-way contingency table, *The Australian and New Zealand Journal of Statistics*, 41, 233-246.

Carlier, A. and Kroonenberg, P. M. (1996), Decompositions and biplots in three-way correspondence analysis, *Psychometrika*, 61, 355-373.

Carroll, J. D. and Chang, J.-J. (1970), Analysis of individual differences in multi-dimensional scaling via an N-way generalisation of "Eckart-Young" decomposition, *Psychometrika*, 35, 283-319.

Emerson, P. L. (1968). Numerical construction of orthogonal polynomials from general recurrence formula. *Biometrics*, 24, 696-701.

Greenacre, M. J. (1984), *Theory and Application of Correspondence Analysis*, Academic Press: London.

Harshman, R. A. (1970), Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis, *UCLA Working Papers in Phonetics*, 16, 1-84.

Lombardo, R., Carlier, A., D'Ambra L. (1996), Nonsymmetric Correspondence Analysis for three-way contingency tables. In *Methodologica*, n. 4, 59-80.

Kroonenberg, P. M. (1989), Singular value decomposition of interactions in three-way contingency tables, In *Multiway Data Analysis* (Coppi, R. and Bolasco, S.), 169-184.

Marcotorchino, F. (1984), *Ulitisation des Comparaisons par Paires en Statistique des Contingencies: Partie III*, Report # F 071, Etude du Centre Scientifique, IBM, France.

Simonetti, B., D'Ambra, L. and Beh, E. J. (2006), The analysis of dependence for three way tables with ordinal variables in food industry evaluation, *9emes Journees Agro-Industrie et Methodes Statistiques*, 171–179.

Tucker, L. R. (1963), Implications of factor analysis of three-way matrices for measurement of change, In *Problems in Measuring Change* (C. W. Harris, ed), 122-137.

