Jan de Leeuw

Canonical analysis of multiple time series

RB - 001 - 72

0:

In this paper we apply the ideas of De Leeuw (1972 a, b) to a particular problem that seems to occur quite often.

1:

There are n categorical variables, m occasions (points in time), and a sample of size l. The variable v has g_v categories, let $N = \sum g_v$ be the total number of categories. The indicator matrix Y is three dimensional with elements y_{ijk} ($i=1,\ldots,N$; $j=1,\ldots,m$; $k=1,\ldots,l$).

2:

By using linear restrictions in the way explained in De Leeuw (1972 a, b) we reduce the matrix y_{ijk} to a smaller matrix z_{ijk} with $i=1,\ldots,M \leq N-n$. In particular we suppose $\sum z_{ijk} = 0$ for all $i=1,\ldots,M$; $j=1,\ldots,m$.

3:

A direct quantification of the variables and a linear combination of the quantified variables produces an induced quantification of occasions and elements by the formula $x_{jk} = \sum w_i z_{ijk}$. The procedures outlined in this paper choose the quantification w in such a way that some rational ANOVA-type criterion is optimized.

4:

The first criterion is a simple homogeneity criterion like the ones discussed in De Leeuw (1972 a, b). We investigate the stationary values of the homogeneity coefficient λ defined by the partition

Source	Sum of Squares	Written out
Between occasions	$m\sum_{k=1}^{l} x_{.k}^2$	$\frac{1}{m}\sum_{i=1}^{M}\sum_{i'=1}^{M} w_i w_{i'} \sum_{k=1}^{l}\sum_{j=1}^{m}\sum_{j'=1}^{m} z_{ijk} z_{i'j'k}$
Within occasions	$\sum_{k=1}^{l}\sum_{j=1}^{m}(x_{jk}-x_{.k})^2$	$\sum_{i=1}^{M}\sum_{i'=1}^{M} w_i w_{i'} \sum_{k=1}^{l}\sum_{j=1}^{m}\sum_{j'=1}^{m} (\delta^{jj'} - \frac{1}{m}) z_{i'j'k} z_{ijk}$
Total	$\sum_{k=1}^{l}\sum_{j=1}^{m} x_{jk}^2$	$\sum_{i=1}^{M}\sum_{i'=1}^{M} w_i w_{i'} \sum_{k=1}^{l}\sum_{j=1}^{m}\sum_{j'=1}^{m} \delta^{jj'} z_{i'j'k} z_{ijk}$

and

$$\lambda = \frac{B}{T}.$$

If we define the l matrices Z_k of order M x m by $(Z_k)_{ij} = z_{ijk}$, and the m x m matrix E with all elements equal to unity, then we can write

$$\lambda = \frac{w'Aw}{w'Bw},$$

with

$$A = \sum Z_k E Z_k',$$

$$B = m\sum Z_k Z_k'.$$

In general all stationary values of λ may be interesting. In some cases we may want to maximize homogeneity over time points, in others heterogeneity. The procedure could be called HOMANOVA. It is not difficult to see that it actually is a special case of the two-set theory developed in De Leeuw (1972a, b). In the first set of variables we use the M x ml matrix of data, for the second set we use the one-way classification of the ml objects into the m different occasions. This makes it clear that our procedure is a type of m-group canonical discriminant analysis with optimal scoring.

5:

Suppose the m x p matrix Y contains a number of orthonormal functions on the m discrete time points. In our second procedure we want to quantify the variables in such a way that the columns of X can be best fitted by linear combinations of the columns of Y in a least squares sense. This corresponds with the partition

Source	Sum of Squares
Explained	$\sum_{k=1}^{l} \sum_{j=1}^{m} \sum_{j'=1}^{m} x_{jk} x_{j'k} \sum_{s=1}^{p} y_{js} y_{j's}$
Residual	$\sum_{k=1}^{l} \sum_{j=1}^{m} \sum_{j'=1}^{m} x_{jk} x_{j'k} (\delta^{jj'} - \sum_{s=1}^{p} y_{js} y_{j's})$
Total	$\sum_{k=1}^{l} \sum_{j=1}^{m} x_{jk}^2$

and

$$\lambda = \frac{E}{T}.$$

Using the Z_k once again gives

$$\lambda = \frac{w'Cw}{w'Dw},$$

$$C = \sum Z_k YY' Z_k',$$

$$D = \sum Z_k Z_k'.$$

In this case we obviously want to maximize λ. If Y contains orthogonal polynomials the procedure could be called POMANOVA, otherwise ORMANOVA. Again it can be interpreted as a special case of the general procedures. The are two sets of variables: our original n observed variables and p additional numerical ones, one for each orthonormal function.

6:

Cluster analysis is not discussed in De Leeuw (1972a, b), but it can easily be fitted into the general framework. We have two sets of variables, one containing our original n observed variables, and the second set containing a single categorical variable with a fixed number of p categories, but with an unknown categorization.

We study the behaviour of λ over different direct quantifications of the variables and over different categorizations. This is equivalent to p-set canonical discriminant analysis with an unknown dependent classification. In our context we have a binary matrix t_{jks} $(s=1,\ldots,p)$, satisfying $\sum t_{jks} = 1$ for all $j=1,\ldots,m$; $k=1,\ldots,l$. Let Z_k be as usual, $(T_k)_{sj} = t_{jks}$. Collect them both in the $M \times (ml)$ supermatrix Z, and the $p \times (ml)$ supermatrix T. The relevant partition is

Source	Sum of squares
Between clusters	$w'ZT'(TT')^{-1}TZ'w$
Within clusters	$w'Z(I - T'(TT')^{-1}T)Z'w$
Total	$w'ZZ'w$

We can maximize

$$\lambda = \frac{B}{T}$$

over all w and T. Alternatively we can maximize the sum of all stationary values of λ, i.e.

$$\mathrm{tr}\left\{(ZZ')^{-\frac{1}{2}}ZT'(TT')^{-1}TZ'(ZZ')^{-\frac{1}{2}}\right\}$$

over T. If we relax the requirement that T must be binary and exclusive to the weaker requirement that TT' must be diagonal, we find that the optimizing $T(TT')^{-\frac{1}{2}}$ is equal to the first p normalized eigenvectors of $Z'(ZZ')^{-1}Z$. This is related to latent partition analysis (Wiley 1967) and provides a good starting point for any cluster analysis. Further improvements must be based on a more or less systematic search over the discrete set of all p-category categorizations. In our particular situation this can be combined with an ordening of the clusters $C_1 \succ \ldots \succ C_p$, and the restriction that an object can only move from C_q to C_{q+1} in a unit time step. This restricts the set of all admissible clusterings in an obvious way, and connects cluster analysis with Markov models and latent Markov chains in particular (Lazarsfeld & Henry 1968, chapter 9).

[illegible]:

The analysis in sections 4 and 5 can be refined by studying homogeneity after the effect of the orthonormal functions in Y is removed. The partition is

Source	Sum of squares
Between	$\frac{1}{m} w' \sum Z_k(I-YY')E(I-YY')Z_k' w$
Within	$\frac{1}{m} w' \sum Z_k(I-YY')(mI - E)(I-YY')Z_k' w$

Total	$w'\sum Z_k(I - YY')Z_k' w$

Consequently we partition the residual component in the analysis of section 5. We can integrate the two approaches by defining

$$A_B = \frac{1}{m} \sum Z_k YY'EYY'Z_k',$$

$$A_W = \frac{1}{m} \sum Z_k YY'(mI-E)YY'Z_k',$$

$$A = A_B + A_W = \sum Z_k YY'Z_k',$$

$$B_B = \frac{1}{m} \sum Z_k (I-YY')E(I-YY')Z_k',$$

$$B_W = \frac{1}{m} \sum Z_k (I-YY')(mI-E)(I-YY')Z_k',$$

$$B = B_B + B_W = \sum Z_k (I-YY')Z_k',$$

$$C = A + B = \sum Z_k Z_k'.$$

The complete partition is

Source	Sum of squares	Subtotal
Explained between	$w'A_Bw$	
Explained within	$w'A_Ww$	
Explained	.....	$w'Aw$
Residual between	$w'B_Bw$	
Residual within	$w'B_Ww$	
Residual	.....	$w'Bw$
Total	.....	$w'Cw$

Alternatively we can also define the subtotal 'between' with matrix $A_B + B_B$, and 'within' with matrix $A_W + B_W$. A general technique is to pick two quadratic froms and to maximize their ratio. The techniques of sections 4, 5, and 7 are special cases.

8:

Lazarsfeld, P.F. & Henry, N.W. — Latent structure analysis. Boston, Houghton Mifflin, 1968

Wiley, D.E. — Latent partition analysis. Psychometrika, 1967, 32, 183-198

De Leeuw, J. — Canonical analysis of relational data, parts 1 & 2. Univ. of Leiden, 1972, (mimeographed)