Generalized canonical analysis of qualitative data

Jean-Marie Bouroche, Gilbert Saporta, Michel Tenenhaus

COREF

CONSEIL, RECHERCHE ET FORMATION

U.S. Japan Seminar

THEORY, METHODS AND APPLICATIONS OF MULTIDIMENSIONAL SCALING AND RELATED TECHNIQUES

August 20-24, 1975
University of California
San Diego, U.S.A.

Sponsors : *National Science Foundation, U.S.A.*
Japan Society for the Promotion of Science

GENERALIZED CANONICAL ANALYSIS OF QUALITATIVE DATA

J.M. BOUROCHE, G. SAPORTA, M. TENENHAUS

COREF

CONSEIL, RECHERCHE ET FORMATION

August, 1975

GENERALIZED CANONICAL ANALYSIS OF QUALITATIVE DATA (*)

J.M. BOUROCHE, G. SAPORTA, M. TENENHAUS

I - INTRODUCTION

A set of n qualitative nominal variables $\tilde{X}_1$, ..., $\tilde{X}_i$, ..., $\tilde{X}_n$, is measured on a common set of k individuals.

We suppose that $\tilde{X}_i$ has m_i levels. Classical factor analysis fails to find out the relationships beetween the $\tilde{X}_i$.

J.P. BENZECRI (1) has proposed to replace each $\tilde{X}_i$ by the corresponding set of dummy variables and to perform on these data a formal correspondance analysis.
Another way is to perform on the n sets of dummy variables a generalized canonical correlation analysis.

J.D. CARROLL's CANCOR (3) leads to the same results as the first approach.
In order to reduce computing time, we use some modified procedure to compute the CANCOR solution : results are obtained in a Σm_i (rather than a k) dimensional vector space.

II - MODIFIED CANCOR PROCEDURE

Using CARROLL's earlier notation, we assume n matrices X_1, ... , X_i, ..., X_n, where :

k is the common number of columns,

m_i is the number of rows of X_i.

---

* Project number n° 75.7.0230, supported by D.G.R.S.T., France.

We assume that all the rows are centered and that the X_i are non singular. The procedure finds Z (k components vector) and n transformation vectors A_i (m_i components vector) so that :

$$R^2 = \sum_{i=1}^{n} \left[\overline{r\ (Z,\ A_i\ X_i)} \right]^2$$

is maximized.

CARROLL shows that Z is the eigenvector of :

$$Q = \sum_{i=1}^{n} Q_i$$

$$Q_i = X_i' \ (X_i\ X_i')^{-1}\ X_i$$

associated with the largest eigenvalue, and that :

$$A_i = Z\ X_i'\ (X_i\ X_i')^{-1}$$

The procedure produces more than one Z by extracting the successive eigenvectors of Q.

Let V and H the super-matrices $(\Sigma m_i) \times (\Sigma m_i)$ defined by :

$$V = \begin{pmatrix} V_{11} & & & 0 \\ & \ddots & & \\ & & V_{ii} & \\ & & & \ddots \\ 0 & & & V_{nn} \end{pmatrix}$$

$$H = \begin{pmatrix} V_{11} & \cdots\cdots & V_{1i} & \cdots\cdots & V_{1n} \\ \vdots & & \vdots & & \vdots \\ V_{i1} & \cdots\cdots & V_{ii} & \cdots\cdots & V_{in} \\ \vdots & & \vdots & & \vdots \\ V_{n1} & \cdots\cdots & V_{ni} & \cdots\cdots & V_{nn} \end{pmatrix}$$

with $V_{j\ell} = \frac{1}{k} X_j X'_\ell$

and let A the super-vector :

$A = (A_1, \ldots, A_i, \ldots, A_n)$

It can be showed that A is the eigenvector of HV^{-1} associated with the largest eigenvalue, which is the same that the largest eigenvalue of Q.

We can also compute Z since :

$A_i X_i = Z Q_i$

$\sum_i A_i X_i = Z (\sum_i Q_i) = \lambda Z$

and $Z = \frac{1}{\lambda} \sum_i A_i X_i$

If we choose A such that $A V A' = \lambda$, we will have $Z Z' = 1$.
This procedure produces the same others Z and A as the former.
The successive A are orthogonal for the inner product defined by V.
A computing problem arises since HV^{-1} is not a symetric matrix.

It can be by-passed by putting $H = C' C$ and solving :

$$(A\ C')\ (C\ V^{-1}\ C') = \lambda\ (A\ C')$$

III - THE ANALYSIS OF NOMINAL VARIABLES

We assume now that the n matrices X_i countain only $\{0, 1\}$ values.
The m_i rows of X_i correspond to the m_i levels of the nominal variable $\tilde{X}_i$ and the k columns correspond to the k individuals.
In one column there is only one non-zero value (equal to one) in the row corresponding to the level of $\tilde{X}_i$ for the concerned individual.
We don't center the rows, so that the rank of X_i is equal to m_i ($m_i - 1$ if centered). We will see latter the consequences of non-centering.
Matrices V and H can be computed and we see that :

$$V_{ii} = \frac{1}{k}\ X_i\ X_i'$$

has all its components equal to zero except those on the diagonal wich are equal to the proportion of individuals at each level of $\tilde{X}_i$. The V_{ii} (and V) are non singular.

In the same way,

$V_{j\ell} = \frac{1}{k}\ X_j\ X_\ell'$ is the normalized contingency table of $\tilde{X}_j$ and $\tilde{X}_\ell$.

When non centering, one can show that the largest eigenvalue of $H\ V^{-1}$ is equal to n and that the corresponding eigenvector A has all its components equal to one, (it's the same for $Z = \frac{1}{\lambda} \sum_i A_i\ X_i$).

The second eigenvalue and the others are the same as if we had first centered the rows and suppressed one row in each X_i.
The corresponding Z are centered (orthogonal to a 1 vector).
The weighted sums of the components of A are equal to zero (the weigth of each component is equal to the proportion of individuals at the corresponding level of the corresponding variable).
It is now necessary to look at the centering problem.

It is necessary, in the general case, that the Z and A_i X_i be centered. In the nominal case, the X_i columns span a m_i dimensional vector subspace in $\mathbb{R}^k$. After centering, they span a $m_i - 1$ dimensional subspace, so that the V_{ii} matrices are singular. The problem is to choose a "pseudo-inverse".

An infinity of solutions exists, for example :

- centering first and suppress one column,
- centering and impose constraints like : the non-weighted sums of each A_i components are equal to zero (same as ANOVA),
- non centering, so that the first eigenvector is 1, and the others are centered.

The non-trivial eigenvalues are independant of this choice, but not the eigenvectors. So the values we will obtain for Z and A depend on it.

The choice we have done lead us to exactly the same numerical results as the use of correspondance analysis on the super-matric $X' = (X_1', \ldots, X_i', \ldots, X_n')$, like BENZECRI has proposed in (1).

BIBLIOGRAPHIE

(1) BENZECRI J.P. — L'analyse des données
Tome II, DUNOD, Paris (1973)

(2) BURT C. — The factorial analysis of qualitative data.
British Journal of Statistical Psychology, 3, (1950)

(3) CARROLL J.D. — A generalization of canonical correlation analysis to three or more sets of variables.
Proc. 76th Conv. American Psychological Association (1968).

(4) KETTENRING R.J. — Canonical analysis of several sets of variables.
BIOMETRIKA, 58 (1971)

(5) de LEEUW J. — Canonical analysis of categorical data
Ph. D. Thesis. Université de Leyde (1973)

(6) MASSON M. — Processus linéaires et analyse de données non linéaires.
Thèse de Doctorat d'Etat Paris (1974)

(7) SAPORTA G. — Liaisons entre plusieurs ensembles de variables et codage de données qualitatives.
Thèse de Doctorat de 3ème cycle
Paris (1975)

(8) TENENHAUS M. — Utilisation de l'analyse canonique généralisée pour le traitement de variables qualitatives.
Note de Travail CESA (1975)