About some remarkable properties of generalized canonical analysis

Gilbert Saporta

ABOUT SOME REMARKABLE PROPERTIES OF GENERALIZED CANONICAL ANALYSIS

Gilbert SAPORTA

Institut Universitaire de Technologie
143, Avenue de Versailles - 75016 PARIS France

The method and the criterion proposed by J.D. Carroll for generalizing canonical analysis to more than two groups of variables, provide a very general model which unify various techniques of multi-dimensional data analysis.

Let X_i $i = 1,2,..,$ p be p sets of m_i variables on n individuals ; each X_i will be considered as a nxm_i matrix. Various attempts (Horst (1961) ; Kettenring (1971) , Masson (1974))has been made to generalize canonical analysis when $p \geq 3$ by finding directly p- uples of canonical variables $\underline{\xi}_i$ ($\underline{\xi}_i = X_i \underline{a}_i$) satisfying to some criteria of optimality. Most of them lead to tricky algorithms and the $\underline{\xi}_i$, generally, are no longer solutions of an eigenequation.

Carroll's method consists in finding an auxiliary variable, the most correlated with the p groups, in the following sense :

$$(1) \qquad \sum_{i=1}^{p} R^2 (\underline{z} , X_i) \quad \text{is maximal}$$

where R^2 is the square multiple correlation coefficient. This leads to the eigenequation :

$$(2) \qquad \sum_{i=1}^{p} X_i (X_i' X_i)^{-1} X_i' \underline{z} = \lambda \underline{z} \quad \lambda \max$$

where $X_i (X_i' X_i)^{-1} X_i' = A_i$ is the orthogonal projector onto the subspace spanned by X_i .

Regressing $\underline{z}$ onto the X_i gives then the canonical variates $\underline{\xi}_i$ and $\underline{z}$ is both the mean and the first principal component of the $\underline{\xi}_i$. It may be noted (Ten BERGE (1977))that this solution (equivalent to solutions proposed by HORST, Mc DONALD and others) is an improper generalization of canonical analysis because, if the $\underline{z}^{(k)}$ corresponding to successive eigenvalues λ_k of equation (2) are orthogonal the $\underline{\xi}_i^{(k)}$ are not orthogonal when $p > 2$.

.../...

Nevertheless, the main interest of Carroll's method stay in its criterion and in finding $\underline{z}$ variates (more than the $\underline{\xi}_i$) which leads to various interpretations :

First it is a generalization of principal components analysis to groups of variables instead of single variables : in normalized p.c.a. the first principal component $\underline{c}$ of $\underline{x}_1$; $\underline{x}_2$ $\underline{x}_p$ provides a maximum for $\sum_{i=1}^{p} r^2 (\underline{c} ; \underline{x}_i)$

Thus Carroll's method comes down to a p.c.a. of the Σm_i variables with a special metric which takes into account the grouping of variables.

Then, applying Carroll's generalized canonical analysis to p nominal variables (each X_i is the array of the indicator variables of the categories of the i^{th} variable $\mathcal{X}_i$) leads exactly to the principal components of scales analysis of Guttman (1941), to the optimal quantification of Hayashi (1950), to the Benzecri-Lebart (1972) multiple correspondence analysis (See also Bouroche - Saporta - Tenenhaus (1975) and to the optimal scaling method Homals of J. De Leeuw (1978) which are all equivalent.

Since the multiple correlation coefficient with the indicator variables is nothing but a correlation ratio, all these methods may be presented in a fairly simple way : they consist in finding a numerical variable $\underline{z}$ maximizing $\sum_{i=1}^{p} \eta^2 (\underline{z} ; \mathcal{X}_i)$

We have thus a general criterion for analyzing sets of nominal or numerical variables.

Neither J.D. Carroll nor others gave a simple equation for the $\underline{\xi}_i$: we prove here that the canonical variables can be obtained as the eigenvector of an np x np matrix whose blocks are the products of projectors $A_i A_j$; this is a generalization of a seemingly new property of ordinary canonical analysis. From this result, it can be easily shown that the $\underline{\xi}_i$ satisfy a mean orthogonality property :

$$\sum_{i=1}^{p} \operatorname{cov} (\underline{\xi}_i^{(k)} , \underline{\xi}_i^{(\ell)}) = 0 \quad \text{if } k \neq 1$$

.../...

which completes the weak orthogonality :

$$\operatorname{cov}\left(\sum_{i=1}^{p} \underline{\xi}_i^{(k)}, \sum_{i=1}^{p} \underline{\xi}_i^{(1)}\right) = 0 \text{ if } k \neq 1$$

of course writing the equation :

$$(3)\quad \begin{pmatrix} A_1 & A_1A_2 & \cdots & A_1A_p \\ A_2A_1 & A_2 & \cdots & \\ \vdots & & \ddots & \\ & & & \end{pmatrix} \begin{pmatrix} \hat{\underline{\xi}}_1 \\ \underline{\xi}_2 \\ \vdots \\ \underline{\xi}_p \end{pmatrix} = \lambda \begin{pmatrix} \underline{\xi}_1 \\ \underline{\xi}_2 \\ \vdots \\ \underline{\xi}_p \end{pmatrix}$$

has only a theoretical interest since it is of dimension np , the solution is in fact performed in dimension Σm_i by solving the equation (4

$$(4)\quad \begin{pmatrix} V_{11} & & & \\ & V_{22} & 0 & \\ & 0 & & \\ & & & V_{pp} \end{pmatrix}^{-1} \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & \cdots & \\ \vdots & & \ddots & \\ & & & V_{pp} \end{pmatrix} \begin{pmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \vdots \\ \underline{a}_p \end{pmatrix} = \lambda \begin{pmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \vdots \\ \underline{a}_p \end{pmatrix}$$

where $\underline{\xi}_i = X_i \underline{a}_i$ and $V_{ij} = X_i' X_j$

REFERENCES

J.M.F. Ten BERGE (1977) Optimizing factorial invariance.Thesis.University of Groningen.

J.P. BENZECRI (1972) - Sur l'analyse des tableaux binaires associés à une correspondance multiple Lab. Stat. Math. Université Paris VI.

J.M. BOUROCHE - G. SAPORTA - M. TENENHAUS (1975) - Generalized canonical analysis of qualitative data. US Japan seminar on multidimensionnal scaling. San Diego.

J.D. CARROLL (1968) - Generalization of canonical correlation analysis to three or more sets of variables.Proceedings 76th convention American Psychological Association 227-228

Y. ESCOUFIER - F. CAILLIEZ - J.P. PAGES (1978) Geometrie et techniques particulieres en analyse factorielle. European Meeting of the Psychometric Society. Uppsala and Revue de Statistique appliquée XXVII n° 1 5-28 (1979)

L.L. GUTTMAN (1941) The quantification of a class of attributes. A theory and method of scale construction in P. Horst ed. "the prediction of personal adjustment" Social Science Rescarch Council.

C. HAYASHI (1950) On the quantification of qualitative data from the mathematico - statistical point of view. Annals Inst-Stat. Math 2 n° 1.

P. HORST (1961) Relation among m sets of measures. Psychometrika 26 129-149.

R.J. KETTENRING (1971) Canonical analysis of several sets of variables.Biometrika 58 433-451

L. LEBART - A. MORINEAU - N. TABARD (1977) Techniques de la description statistique Dunod (Paris).

J. De LEEUW (1978) Homals.Dept of Data Theory.University Of Leyden

R.P. MAC DONALD (1968) A unified treatment of the weighting problem Psychometrika 33 351-381.

M. MASSON (1974) Processus linéaires et analyse de données non linéaire. Thèse d'état.Université Paris VI

G. SAPORTA (1975) Liaison entre plusieurs ensembles de variables et codages de données qualitatives. Thèse 3e cycle Université Paris V.

KEYWORDS : DATA ANALYSIS ; MULTIVARIATE ANALYSIS ; CANONICAL ANALYSIS ; PRINCIPAL COMPONENT ANALYSIS ; QUALITATIVE DATA; CORRESPONDENCE ANALYSIS.