About maximal association criteria in linear analysis and in cluster analysis

Gilbert Saporta

Classification and Related Methods of Data Analysis
H. H. Bock (Editor)
Elsevier Science Publishers B.V. (North-Holland), 1988

ABOUT MAXIMAL ASSOCIATION CRITERIA IN LINEAR ANALYSIS AND IN CLUSTER ANALYSIS

Gilbert SAPORTA

Conservatoire National des Arts et Métiers, 292 rue Saint-Martin, 75141 Paris Cédex 03

Many well-known methods of multivariate analysis may be presented in terms of maximizing the sum of association measures between an unknown variable Y and several known variables $X_1\ X_2, \ldots\ X_p$:

$$\max_{Y} \sum_{k=1}^{p} \phi\ (Y\ ;\ X_j)\ .$$

According to the type of the variables and to the association measures we get various forms of cluster analysis if Y is categorical, or of component analysis if Y is numerical : principal components, redundancy analysis, correspondence analysis. We try here to present a survey of this approach and propose new criteriums.

I - Maximal association in linear analysis.

1.1 Principal components (p.c.a.) and generalized canonical analysis (g.c.a).

Everyone knows that the first principal component $\underline{c}$ of a set of p standardized variables $\underline{x}_1\ \underline{x}_2, \ldots\ \underline{x}_p$ maximizes :

$$\sum_{k\ =\ 1}^{p} r^2\ (\underline{c}\ ;\ \underline{x}_k) \qquad \text{(P1)}$$

and that the other components corresponds to the other stationary values of this criterium.

J.D. Carroll [1] proposed a similar criterium for generalized canonical analysis, when there are p sets of m_k numerical variables $(k = 1, 2, \ldots, p)$.
Let $X_1, X_2, \ldots, X_p$ be the p data matrices of zero-mean variables, then a set of canonical variables $\underline{\xi}_1, \underline{\xi}_2, \ldots\ \underline{\xi}_p$ may be derived by regressing an auxiliary variable $\underline{z}$ defined by :

$$\max_{\underline{z}} \sum_{k=1}^{p} R^2\ (\underline{z}\ ;\ X_k) \qquad \text{(P2)}$$

$$\underline{\xi}_k\ =\ X_k\ (X'_k\ X_k)^{-1}\ X'_k\ \underline{z}$$

where R^2 is the square multiple correlation coefficient.
This kind of canonical analysis comes down to usual p.c.a. when $m_k = 1$ for every k .

Since $\underline{z} = \sum_{k=1}^{p} X_k \underline{a}_k$, the component loadings $\underline{a}_k$ are given by :

$$\begin{bmatrix} V_{11}^{-1} & O \ldots\ldots O \\ O & V_{22}^{-1} \\ & \\ O & V_{pp}^{-1} \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} & V_{1p} \\ & V_{22} & \\ & & \\ V_{p1} & V_{p2} & V_{pp} \end{bmatrix} \begin{bmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \\ \underline{a}_p \end{bmatrix} = \lambda \begin{bmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \\ \underline{a}_p \end{bmatrix}$$

see [13] where V_{ij} is the covariance matrix between sets X_i and X_j.

One can remark that the max of $\sum_{k=1}^{p} r(\underline{c} ; \underline{x}^k)$ is attained when $\underline{c}$ is proportional to $\sum_{k=1}^{p} \underline{x}_k$.

What is the solution of $\max_{\underline{z}} \sum_{k=1}^{p} R(\underline{z} ; X_k)$? (P3)

If $A_k = X_k (X_k' X_k)^{-1} X_k'$, then $R(\underline{z} ; X_k) = \left(\frac{\underline{z}' A_k \underline{z}}{\underline{z}' \underline{z}} \right)^{1/2}$

and maximizing (P3) is equivalent to maximize the Lagrangian :

$$\sum_k (\underline{z}' A_k \underline{z})^{1/2} - \frac{1}{2} \lambda \underline{z}' \underline{z} .$$

Differentiating in $\underline{z}$ leads to :

$$\frac{A_k \underline{z}}{(\underline{z}' A_k \underline{z})^{1/2}} = \lambda \underline{z} . \text{ Let } r_{ij} = r(A_i \underline{z} ; A_j \underline{z}) = \frac{\underline{z}' A_i A_j \underline{z}}{(\underline{z}' A_i \underline{z})^{1/2} (\underline{z}' A_j \underline{z})^{1/2}} .$$

By straightforward computation we get :

$$\sum_i \sum_{j \neq} r_{ij} = \lambda^2$$

Since λ is to be maximized, the $A_i \underline{z}$ are the canonical variables defined by Horst [6] and $\underline{z}$ is proportional to the sum of these standardized canonical variables. However, this does not lead to simple algorithms for it is no longer an eigenvalue problem.

1.2 Multiple correspondence analysis (m.c.a.)

M.c.a., also known as "homogeneity analysis" [5] , dual scaling [11] or else (see [16] or [2]) is a technique that derives numerical scales $\underline{z}$ from a set of p categorical variables. $\mathcal{X}_1 , \mathcal{X}_2 , \ldots \mathcal{X}_p$ with respective number of categories $m_1 , m_2 , \ldots , m_p$.

m.c.a. is equivalent to p.c.a. in that respect that the numerical scales $\underline{z}$ are stationary solutions of :

$$\max \sum_{k=1}^{p} \eta^2 (\underline{z} ; \mathcal{X}_k) \qquad \text{(P4)}$$

where η^2 is the square correlation ratio (or variance ratio) between the numerical variable $\underline{z}$ and the categorical variable $\mathcal{X}_k$.

Let X_k be the indicator matrix of $\mathcal{X}_k$, since $\eta^2 (\underline{z} ; \mathcal{X}_k) = R^2 (\underline{z} ; X_k) = \max_{\underline{a}_k} r^2(\underline{z} ; X_k a_k)$

m.c.a. is also a particular case of g.c.a. .

The scales $\underline{z}$ are eigenvectors of the matrix Q of size n defined by

$Q = \sum_{k=1}^{p} A_k = X D^{-1} X'$ where X is the super indicator matrix. We will see further an interpretation of these matrices.

Tenenhaus [15] , by combining criteriums (P1) and (P4) for the case of a mixture of continuous and categorical variables derived an extension of both p.c.a. and m.c.a. :

$$\max \left(\sum r^2 (\underline{c} ; \underline{x}_j) + \sum \eta^2 (\underline{c} ; \mathcal{X}_k) \right) \qquad \text{(P5)}$$

1.3 Redundancy analysis or p.c.a. of instrumental variables

Proposed by C.R. Rao [12] and studied mainly by Escoufier [3] and Van Den Wollenberg [17] redundancy analysis consists in deriving a linear combination of variables of a first set X_1 $\underline{\xi} = X_1 \underline{a}$, that be the best predictor for a second set of variables X_2 .

One finds that $\underline{a}$ is solution of

$$V_{11}^{-1} V_{12} V_{21} \underline{a} = \lambda \underline{a} \quad \text{with } \lambda \text{ max} .$$

When the variables of X_2 have unit-variances, the $\underline{\xi}'_1$ are the linear combinations of variables of X_1 , maximally correlated with the variable of the second set according to the criterium :

$$\max \sum_{k=1}^{p_2} r^2 (\underline{\xi} ; \underline{x}_k^{(2)}) \qquad \text{(P6)}$$

The $\underline{\xi}$ are principal component of the data matrix $X_1 V_{11}^{-1} V_{12}$, i.e the matrix of the least-squares approximations of $\underline{x}_1^{(2)}$, ... $\underline{x}_{p_2}^{(2)}$ by the variables of the first set.

All these techniques have been generalized for time continuous data [2] , [14] , i.e. realizations of a stochastic process X_t .

For instance p.c.a. of a numerical process (Karhunen-Loeve decomposition) consists in finding variables z maximizing :

$$\int_0^T r^2 (\underline{z} ; X_t) \, dt \qquad \text{(P7)}$$

and qualitative harmonic analysis in finding the maximum of

$$\int_0^T \eta^2 (\underline{z} ; \mathcal{X}_t) \, dt \qquad (P8)$$

II - Maximal association for ordinal variables

When the p variables $X_1 , X_2 , \ldots X_p$ are ordinal, finding a new variable Y of the same kind (in general a complete order) may be done by maximizing the sum of functions of rank correlation coefficients.
This approach is close to consensus or ranking aggregation problems.

2.1 Maximizing sum of Spearman's r_s and of Kendall's τ

Since r_s is invariant by monotone transformation of the data, and is a product moment correlation coefficient

The solution of $\max \sum_{k=1}^{p} r_s (Y ; X_k)$ (P9)

is given by the order associated to the sum of the p rankings of the n observations. This is known as Borda's rule.

The solution of : $\max \sum_{k=1}^{p} \tau (Y ; X_k)$ (P10)

is the order given by the Condorcet aggregation rule for which efficient algorithms have been proposed by Marcotorchino and Michaud [9] , using a paired comparison approach.

2.2 Maximizing $\sum r_s^2$

Less attention has been paid in the literature to maximizing sum of squares of rank correlation coefficients. In a consensus framework, it is of course of nonsense to equally consider a ranking and its opposite. However this criterium may be interesting in order to robustify p.c.a.(see [7]) or in order to eliminate rankings which are uncorrelated to the others.

The solution of : $\max \sum_{k=1}^{p} r_s^2 (Y ; X_k)$ (P11)

is simply given by the order associated to the first principal component of the rank matrix.
The problem :

$$\max \sum_{k=1}^{p} \tau^2 (Y ; X_k) \qquad (P12)$$

seems an open one.

III - Cluster analysis

Partitioning a set of n observations is equivalent to look for an unknown categorical variable Y correlated in some respect to descriptive variables $X_1 , X_2 , \ldots X_p$.

3.1 Numerical variables

The most frequently used criterium for a partition is based upon the maximization of Trace B where B is the between-class covariance matrix.

For zero-mean variables, Trace $B = \frac{1}{n} \sum_{i=1}^{k} n_i \, \| \underline{g}_i \|^2$ where the $\underline{g}_i$ are the centroïds of the clusters. Since Trace B is also equal to the sum of between class variances, for standardized variables this comes down to :

$$\max_{Y} \sum_{j=1}^{p} \eta^2 (\underline{x}_j ; Y) \qquad \text{(P13)}$$

and we know that k , the number of categories of Y has to be fixed in order to prevent the trivial solution $k = n$, corresponding to $\eta^2 = 1$.

3.2 Categorical variables

Several partitioning methods may be rewritten in terms of the maximization of

$$\sum_{j=1}^{p} \emptyset (Y ; \mathcal{X}_j)$$

where $\emptyset$ is an adequate measure of association.
Marcotorchino [8] pointed out recently that the central partition problem, which consists in maximizing the number of agreements with p known partitions, may be set as :

$$\max_{Y} \sum_{j=1}^{p} R (Y ; \mathcal{X}_j) \qquad \text{(P14)}$$

Where R is the Rand's measure of association :

$$R = 1 + \frac{2 \sum\sum n_{uv}^2 - \sum n_{u.}^2 - \sum n_{.v}^2}{n^2}$$

Various other criteria may be defined using other measures of association, for instance :

$$\max \sum_{j=1}^{p} \chi^2 (Y ; \mathcal{X}_j) \qquad \text{(P15)}$$

However unlike (P14) , (P15) has a trivial solution $k = n$.

IV - Measures of association, similarity matrices and the RV coefficient.

4.1 Some measures of similarities between pairs of observations

Almost all the above mentioned coefficients can be defined as scalar products between two similarity matrices : Trace AB, where A and B are n x n matrices giving the similarities between pairs of observations according to the measurement level of the underlying variable.

It is well known, that r, r_s, τ are particular cases of Daniels coefficients :

$$r = \frac{\sum\sum a_{ij}\, b_{ij}}{\sqrt{\sum\sum a_{ij}^2 \;\sum\sum b_{ij}^2}} = \frac{\text{Trace } AB}{\sqrt{\text{Trace } A^2 \;\text{Trace } B^2}}$$

for various forms of the elements :

$a_{ij} = x_i - x_j$ for r

$a_{ij} = \text{rank } x_i - \text{rank } x_y$ for r_s

$a_{ij} = \dfrac{x_i - x_j}{|x_i - x_j|}$ for τ

When the matrices A et B are related to data arrays Y and X by $A = Y M Y'$ and $B = X N X'$ where M and N are suitable metrics, we find the approach advocated by Robert and Escoufier [4] and the scalar product between A and B leads to the RV coefficient.

For a single numerical variable $\underline{x}$ with zero mean we will use $b_{ij} = x_i x_j$, i.e. $B = \underline{x}\,\underline{x}'$.

The sum of similarity matrices for p centered numerical variables is thus equal to the matrix W of scalar products : $W = X X'$.

For a single categorical variable $\mathcal{X}$ with m categories we may use one of the three following similarities :

(1) $\begin{cases} b_{ij} = 1 & \text{if i and j belong to the same category} \\ b_{ij} = 0 & \text{else} \end{cases}$

Thus $B = X X'$ where $X_{n,m}$ is the binary indicator matrix.

Or (2) $t_{ij} = 2\, b_{ij} - 1$ $\qquad T = 2\, XX' - 1\,1'$

(3) $u_{ij} = \dfrac{2\, b_{ij}}{b_{i.} + b_{.j}} = \dfrac{b_{ij}}{b_{i.}} = \dfrac{b_{ij}}{\sqrt{b_{i.}\, b_{.j}}}$ which is the inverse of the frequency

of the category in common $U = X(X'X)^{-1} X'$ is the chi-square similarity matrix.

For p categorical variables, the sum of the U matrices is equal to Q (see part 1-2) and the sum of the T matrices is the majority matrix giving the number of times where elements of each pair belong to a same category minus the number of times they do not.

4.2 Cluster analysis and factor analysis : a few equivalences

We see easily from the preceding formulas that p.c.a. and m.c.a. consist in finding the dominant eigenvectors of the similarity matrices W and Q respectively, whilst cluster analysis consists in finding an unknown indicator matrix Y such that the associated similarity matrix be the closest possible to another one.

More precisely :

Problem P1 (p.c.a.) is equivalent to :

$$\max_{\underline{c}} \sum_{j=1}^{p} \text{Trace } (\underline{c}\,\underline{c}'\,\underline{x}_j\,\underline{x}'_j) = \max_{\underline{c}} \text{Trace } (\underline{c}\,\underline{c}'\,W)$$

Problem P4 (m.c.a.) is equivalent to :

$$\max_{\underline{c}} \sum_{j=1}^{p} \text{Trace } (\underline{c}\,\underline{c}'\, X_j\, (X_j'\, X_j)^{-1}\, X_j') \;=\; \max_{\underline{c}} \text{Trace } (\underline{c}\,\underline{c}'\, Q)$$

Problem P6 is equivalent to :

$$\max_{\underline{c}} \sum_{j=1}^{p} \text{Trace } \underline{c}\,\underline{c}'\, (P_1\, \underline{x}_j^{(2)})\; (P_1\, \underline{x}_j^{(2)})' \;=\; \text{Trace } \underline{c}\,\underline{c}'\, P_1\, W_2 \quad \text{where } P_1 = X_1\, (X_1'\, X_1)^{-1}\, X_1'$$

Problem P13 is equivalent to :

$$\max_{Y} \sum_{j} \text{Trace } (Y\, (Y'Y)^{-1}\, Y'\, \underline{x}_j\, \underline{x}_j') \;=\; \max_{Y} \text{Trace } (Y(Y'Y)^{-1}\, Y'\, W)$$

Problem P14 is equivalent to :

$$\max_{Y} \sum_{j} \text{Trace } (2\, Y\, Y' - \underline{1}\,\underline{1}')\; (2\, X_j\, X_j' - \underline{1}\,\underline{1}') \;=\; \max_{Y} \text{Trace } (2\, Y\, Y' - \underline{1}\,\underline{1}')\, T$$

The last one excepted, these problems are equivalent to the maximization of the RV coefficient, between the unknown variable and the known variables, with suitable normalizations.
Problem P13 has been studied by Nin [10] who was looking directly for the indicator matrix with an optimization algorithm.

4.3. Cluster analysis and the paired-comparison approach.

The methodology proposed by Marcotorchino-Michaud [9] consists in finding directly the similarity matrix $Y\,Y' = A$ instead of the binary indicator matrix Y_{nk} where k is generally unknown.
The elements of A must verify the constraints

$$\begin{aligned} a_{ij} &= a_{ji} && \text{reflexivity} \\ a_{ij} + a_{jk} - a_{ik} &\leq 1 && \text{transitivity} \end{aligned} \qquad \text{(C)}$$

When the maximization problems may be written with a linear objective function, a solution can be achieved by using a variant of linear programming for instance, and the number of clusters need not to be fixed in advance : it is an outcome of the maximization procedure.
Problems P13, P15 are not linearizable since the objective functions are :

$$\text{P13} \;:\; \max_{a_{ij}} \sum_i \sum_j \frac{a_{ij}}{a_{i.}}\; w_{ij}$$

$$\text{P15} \;:\; \max_{a_{ij}} \sum_i \sum_j \frac{a_{ij}}{a_{i.}}\; q_{ij}$$

unlike the Condorcet problem :

$$\text{P14} \;:\; \max_{a_{ij}} \sum_i \sum_j a_{ij}\; t_{ij}$$

4.4. A new criterium for numerical data

The necessity of fixing the number of clusters for criterium P13 is due to the presence of a denominator $a_{i.}$. The following criterium which, as far as we know, has not been previously used, allows a partition without fixing the number of classes :

$$\max \quad \text{Trace}\,(Y\,Y'\,W) \tag{P16}$$

Alternative formulations of this criterium are :

$$\max_{a_{ij}} \sum_i \sum_j a_{ij}\, w_{ij} \tag{P16'}$$

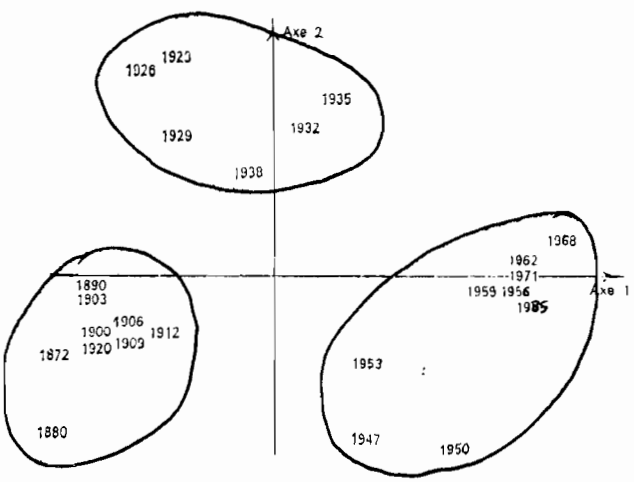
with the constraints (C) for which we may use the general algorithm of [9].
Since Trace $Y\,Y'W$ = Trace $Y\,Y'\,X\,X'$ = Trace $X'\,Y\,Y'\,X$ and

$$Y'X = \begin{bmatrix} n_1\,\underline{g}_1 \\ n_2\,\underline{g}_2 \\ \vdots \\ n_k\,\underline{g}_k \end{bmatrix}$$

$$\max \sum_{i=1}^{k} n_i^2 \left\| \underline{g}_i \right\|^2 \tag{P16''}$$

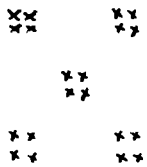
We see here the main difference with the Trace of between covariance matrix : there is here a compromise between the size of the classes and the distance of the centroïds to the overall mean.
At first sight, it seems that this criterium will tend to create clusters far from the centroïd, for the criterium is based on the scalar products "through the origin".
This is confirmed by the analysis of few exemples.

Exemple 1 is a 24 x 11 table of data studied by Bouroche, Saporta * about public expenditures between 1872 and 1971.
Criterium P16 gives three clusters, represented here with the first two principal components which accounted 64% of the variance.



* L'analyse des Données - Que sais-je - PUF, Paris - page 18.

Exemple 2 is an artificial set of data corresponding to the side "five" of a dice :



Here the criterium fails to recognize the five natural clusters and give two solutions whith two clusters.

V - Final remarks

It is our opinion that the explicitation of the underlying criteria of various techniques provide some unity between linear methods and cluster analysis. It offers also the possibility of inventing new criteria, but in this case there is a need for further experiments.

References

[1] CARROLL J.D. (1968). A generalization of canonical correlation analysis to three or more sets of variables. Proc. 76th Conv. Am. Psych. Ass., 227-228.

[2] DEVILLE J.C. , SAPORTA G. (1983). Correspondence analysis with an extension towards nominal series. J. of Econometrics, 22, 169-189.

[3] ESCOUFIER Y. (1979). New results and new uses in principal components of instrumental variables. Cont. papers. 42nd Session ISI, 149-152.

[4] ESCOUFIER Y. , ROBERT P. (1976). A unifying tool for linear multivariate analysis methods : the RV-coefficient. JRSS, B, 25, 257-265.

[5] GIFI A. (1981). Non linear multivariate analysis. Dept of Data Theory, Leyden University.

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

[7] LEBART L. , MORINEAU A. , TABARD N. (1977). Techniques de la description statistique. Dunod, Paris.

[8] MARCOTORCHINO F. (1986). Maximal Association as a Tool for Classification. In Classification as a Tool of Research. W. Gaul, M. Schader Editors. North Holland, 275-288.

[9] MARCOTORCHINO J.F. , MICHAUD P. (1979). Optimisation en analyse ordinale des données, Masson, Paris.

[10] NIN G. (1982). Cluster analysis based on the maximization of the RV coefficient .In Compstat 82. H. Caussinus, P. Ettinger, R. Tomassone Editors, Physica-Verlag, 359-363.

[11] NISHISATO S. (1980). Analysis of categorical data : dual scaling and its applications, Univ. Toronto. Press, Toronto.

[12] RAO C.R. (1964). The use and interpretation of principal components analysis in applied research, Sankhya, A, 26, 329-357.

[13] SAPORTA G. (1980). About some remarkable properties of generalized canonical analysis. European Meeting of the Psychometric Society, Groningen, Netherlands.

[14] SAPORTA G. (1981). Methodes exploratoires d'Analyse des Données temporelles, Cahiers du BURO, n° 37-38, Institut de Statistique. Paris.

[15] TENENHAUS M. (1977). Analyse en Composantes Principales d'un Ensemble de Variables Nominales ou Numériques. Revue de Statistique Appliquée, 25, 39 - 56.

[16] YOUNG F.W. , TENENHAUS M. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, etc... Psychometrika 50, 91 - 119.

[17] VAN DEN WOLLENBERG (1977). Redundancy analysis : an alternative for canonical correlation analysis. Psychometrika, 42, 207 - 219.

Acknowledgements : we are indebted to M. Petitjean who programmed the method of paragraph 4.4.