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 $X_1, X_2, \ldots, X_p$:

$$\max_Y \sum_{k=1}^p \Phi (Y; X_k).$$

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.

1 - 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 $\xi$ of a set of $p$ standardized variables
$X_1, X_2, \ldots, X_p$ maximizes:

$$\sum_{k=1}^p r^2 (\xi; x_k)$$

(P1)

and that the other components correspond 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 $\xi_1, \xi_2, \ldots, \xi_p$ may be derived by regressing an auxiliary variable $\bar{z}$ defined by:

$$\max_\bar{z} \sum_{k=1}^p R^2 (\bar{z}; X_k)$$

(P2)

$$\xi_k = X_k (X_k'X_k)^{-1} X_k' \bar{z}$$

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 $z = \sum_{k=1}^{p} x_k a_k$, the component loadings $a_k$ are given by:

\[
\begin{bmatrix}
V_{11} & \ldots & V_{1p} \\
V_{p1} & \ldots & V_{pp}
\end{bmatrix}
\begin{bmatrix}
V_{22}^{-1}
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
a_p
\end{bmatrix}
= \lambda
\begin{bmatrix}
a_1 \\
a_2 \\
a_p
\end{bmatrix}
\]

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

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

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

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

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

\[
\sum_{k} (z' A_k z)^{1/2} - \frac{1}{2} \lambda z' z
\]

Differentiating in $z$ leads to:

\[
\frac{A_k z}{(z' A_k z)^{1/2}} = \lambda z . \quad \text{Let} \quad \lambda_{ij} = r(A_i z; A_j z) = \frac{z' A_i A_j z}{(z' A_i z)^{1/2} (z' A_j z)^{1/2}} .
\]

By straightforward computation we get:

\[
\sum_{i \neq j} \lambda_{ij} = \lambda^2
\]

Since $\lambda$ is to be maximized, the $A_i z$ are the canonical variables defined by Horst [6] and $z$ 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 $z$ from a set of $p$ categorical variables $x_1, x_2, \ldots, x_p$ with respective number of categories $m_1, m_2, \ldots, m_p$. 

Maximal Association Criteria

m.c.a. is equivalent to p.c.a. in that respect that the numerical scales \( z \) are stationary solutions of:

\[
\max \sum_{k=1}^{p} \eta^2 (z ; \chi_k) \tag{P4}
\]

where \( \eta^2 \) is the square correlation ratio (or variance ratio) between the numerical variable \( z \) and the categorical variable \( \chi_k \).

Let \( X_k \) be the indicator matrix of \( \chi_k \), since \( \eta^2 (z ; \chi_k) = R^2 (z ; X_k) = \max \gamma^2 (z ; X_k a_k) \).

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

The scales \( z \) are eigenvectors of the matrix \( Q \) of size \( n \) defined by:

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

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.:

\[
\max \left( \sum r^2 (\xi ; \chi_j) + \sum \eta^2 (\xi ; \chi_k) \right) \tag{P5}
\]

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

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

One finds that \( \bar{a} \) is solution of:

\[
V_{11}^1 V_{12} V_{21} V_{22} = \lambda \bar{a} \quad \text{with} \quad \lambda \quad \text{max}.
\]

When the variables of \( X_2 \) have unit-variances, the \( \xi_i' \) are the linear combinations of variables of \( X_1 \), maximally correlated with the variable of the second set according to the criterium:

\[
\max \sum_{k=1}^{p_2} r^2 (\xi_k ; \chi_k) \tag{P6}
\]

The \( \xi_k \) are principal component of the data matrix \( X_1 V_{11}^{1} V_{12} \), i.e. the matrix of the least-squares approximations of \( \chi_{k1}^{(2)} \), ..., \( \chi_{k2}^{(2)} \) 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:

\[
\int_0^T r^2 (z ; X_t) \, dt \tag{P7}
\]
and qualitative harmonic analysis in finding the maximum of
\[ \int_0^T \eta^2 (z; \chi_t) \, dt \] (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 \( \tau \).

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 Marcorchino 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 criterion 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:
\[ \max \sum_{k=1}^{p} \tau^2 (Y; X_k) \] (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 criterion for a partition is based upon the maximization of Trace \( B \) where \( B \) is the between-class covariance matrix.

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

\[
\max_Y \sum_{j=1}^{p} \eta^2 (x_j ; Y) \tag{P13}
\]

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

\[
\sum_{j=1}^{p} \phi (Y ; \mathcal{X}_j) \tag{P14}
\]

where \( \phi \) 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:

\[
\max_Y \sum_{j=1}^{p} R (Y ; \mathcal{X}_j) \tag{P14}
\]

Where \( R \) is the Rand's measure of association :

\[
R = 1 + \frac{2 \sum \sum \eta_{uv}^2 - \sum \eta_u^2 - \sum \eta_v^2}{n^2} \]

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

\[
\max_Y \sum_{j=1}^{p} \chi^2 (Y ; \mathcal{X}_j) \tag{P15}
\]

However unlike \( \text{(P14)} \), \( \text{(P15)} \) 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 \( n \times n \) matrices giving the similarities between pairs of observations according to the measurement level of the underlying variable.

It is well known, that \( r, r_s, \tau \) 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:
\[ a_{ij} = x_i \cdot x_j \quad \text{for } r \]
\[ a_{ij} = \text{rank } x_i \cdot \text{rank } x_j \quad \text{for } r_s \]
\[ a_{ij} = \frac{x_i - x_j}{|x_i - x_j|} \quad \text{for } \tau \]

When the matrices \( A \) and \( B \) 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 (14) and the scalar product between \( A \) and \( B \) leads to the RV coefficient.

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

The sum of similarity matrices for \( p \) centered numerical variables is thus equal to the matrix \( W \)
of scalar products: \( W = XX' \).

For a single categorical variable \( X \) with \( m \) categories we may use one of the three following similarities:

\[
\begin{align*}
(1) \quad b_{ij} & = \begin{cases} 
1 & \text{if } i \text{ and } j \text{ belong to the same category} \\
0 & \text{else}
\end{cases} \\
\text{Or} \quad (2) \quad t_{ij} & = 2 b_{ij} - 1 \\
(3) \quad u_{ij} & = \frac{2 b_{ij}}{b_i + b_j} = \frac{b_{ij}}{\frac{b_i b_j}{\sqrt{b_i b_j}}} \quad \text{which is the inverse of the frequency}
\end{align*}
\]

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 \( \text{P1 (p.c.a.)} \) is equivalent to:

\[
\max_{\xi} \sum_{j=1}^{p} \text{Trace } (\xi \xi' x_j x_j') = \max_{\xi} \text{Trace } (\xi \xi' W)
\]
Maximal Association Criteria

Problem P4 (m.c.a.) is equivalent to:
\[
\max_{\mathcal{C}} \sum_{j=1}^{n} \text{Trace} \left( \mathcal{C} \mathcal{C}' X_j (X_j' X_j)^{-1} X_j' \right) = \max_{\mathcal{C}} \text{Trace} \left( \mathcal{C} \mathcal{C}' Q \right)
\]

Problem P6 is equivalent to:
\[
\max_{\mathcal{C}} \sum_{j=1}^{n} \text{Trace} \left( \mathcal{C} \mathcal{C}' \left( P_1 X_j^{(2)} \right) \left( P_1 X_j^{(2)} \right)' \right) = \text{Trace} \left( \mathcal{C} \mathcal{C}' P_1 W_2 \right) \text{ where } P_1 = X_1 (X_1' X_1)^{-1} X_1'
\]

Problem P13 is equivalent to:
\[
\max_{\mathcal{Y}} \sum_{j} \text{Trace} \left( \mathcal{Y} (\mathcal{Y}' \mathcal{Y})^{-1} \mathcal{Y}' X_j X_j' \right) = \max_{\mathcal{Y}} \text{Trace} \left( \mathcal{Y}(\mathcal{Y}' \mathcal{Y})^{-1} \mathcal{Y}' W \right)
\]

Problem P14 is equivalent to:
\[
\max_{\mathcal{Y}} \sum_{j} \text{Trace} \left( (2 \mathcal{Y} \mathcal{Y}' - \mathbf{1} \mathbf{1}') (2 X_j X_j' - \mathbf{1} \mathbf{1}') \right) = \max_{\mathcal{Y}} \text{Trace} \left( (2 \mathcal{Y} \mathcal{Y}' - \mathbf{1} \mathbf{1}') \mathbf{T} \right)
\]

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 \( \mathcal{Y} \mathcal{Y}' = \mathcal{A} \) instead of the binary indicator matrix \( \mathcal{Y}_{nk} \) where \( k \) is generally unknown.

The elements of \( \mathcal{A} \) must verify the constraints
\[
\begin{align*}
    a_{ij} &= a_{ji} & \text{(reflexivity)} \\
    a_{ij} + a_{jk} - a_{ik} &\leq 1 & \text{(transitivity)}
\end{align*}
\]

(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:

\[
P13 : \max a_{ij} \sum_i \sum_j \frac{a_{ij}}{a_{ij}} w_{ij}
\]

\[
P15 : \max a_{ij} \sum_i \sum_j \frac{a_{ij}}{a_{ij}} q_{ij}
\]

unlike the Condorcet problem:

\[
P14 : \max a_{ij} \sum_i \sum_j a_{ij} t_{ij}
\]
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:

$$\max \text{ Trace } (Y Y' W)$$

(P16)

Alternative formulations of this criterium are:

$$\max \sum_{i} \sum_{j} a_{ij} w_{ij}$$

(P16')

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

$$Y' X = \begin{bmatrix} n_1 g_1 \\ n_2 g_2 \\ \vdots \\ n_k g_k \end{bmatrix}$$

$$\max \sum_{i=1}^{k} n_i^2 \|g_i\|^2$$

(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 centroids to the overall mean.

At first sight, it seems that this criterium will tend to create clusters far from the centroid, for the criterium is based on the scalar products "through the origin". This is confirmed by the analysis of few examples.

Exemple 1 is a $24 \times 11$ 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.

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

\[ \begin{array}{ccc} 
\times & \times & \times \\
\times & \times & \\
\times & \\
\times & \times & \times \\
\times & \times 
\end{array} \]

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


Univ. Toronto Press, Toronto.

applied research, Sankhya, A, 26, 329-357.

analysis. European Meeting of the Psychometric Society, Groningen, Netherlands.


correspondence analysis, optimal scaling, dual scaling, etc.. Psychometrika 50, 91 - 119.


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