Control
and Cybernetics
VOL. 15 (1986) No. 2

Canonical analysis from the automatic classification point of view

by

EDWIN DIDAY

INRIA
Rocquencourt, Le Chesnay
France

One of the major objectives of data analysis is to discover relations between variables describing a given population. This problem is considered beginning with the canonical analysis, and passing via the factor analysis of correspondences and the discriminant analysis. An algorithm solving the problem is proposed as well for the centered as for the non-centered case and its convergence is proved. Its numerical aspects are discussed.

1. Introduction

One of the major objectives of the various data analysis methods is to discover relations between the variables which characterize a given population of objects. Thus, in canonical analysis, the correlations between the variables of the predefined groups are to be found: in the factor analysis of correspondances, (see Benzecri (1973) and Lébart and Fenelon (1975)), we look for the most characteristic contingencies between variables. In discriminant analysis (see Caillez and Pages (1976)), we look for a linear combination of variables, having a good separating power for the a priori given clusters characterizing the variable to predict. None of these methods takes into account the fact that the discovered relations may differ according to the type of the subpopulation which is being considered.

Thus, in a survey on the towns of a country, it is clear that the behaviour of variables will be different, if the towns which are considered are suburbian towns, surrounding large cities, on one hand, or if they are located in the country, far from any city, on the other hand. The problem that we consider in this paper, is the following: how to exhibit clusters of objects which have a homogeneous behaviour with respect to characteristic relations between the variables? Let us now recall the classical case of two uncorrelated variables: consider a set of objects, characterized by the two variables x and y and suppose that the set of the values of (x, y) for each object, varies according to the curve as in Fig. 1.

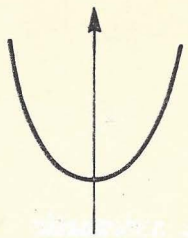

Fig. 1

In the canonical analysis case, the problem can be formulated as: given two groups of variables X and Y, characterizing a set E of objects, find a partition of E, such that, for each one of the constituting clusters there exists a linear combination ξ of the variables of X and a linear combination η of the variables of Y, such that their squares have the maximum correlation. If we are in the case when all the variables are qualitative, the problem is formulated in terms of the factor analysis of the correspondences and reduces to looking for the clusters of objects, which induce the largest χ^2 on the contingency between the variables (all qualitative).

Finally, from the discriminant analysis point of view, the population has to be partitioned into several parts P_i, and a linear combination L_i of the variables, having a large separating power of the a priori given clusters C_i, has to be associated to each part P_i (see Fig. 2).

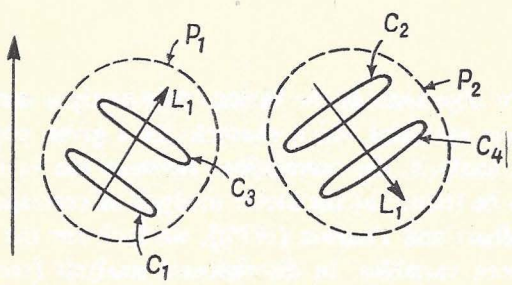

Fig. 2

After this brief review of the classical properties of canonical analysis, we formulate the canonical problem from the automatic classification point of view, first in the centered case (that is, when the variables are centered in each cluster), then, in the non centered case. Next, we give expressions for the criteria to optimize, and, in each case, algorithms to find the local optima, with a study of their convergence properties. The particular cases of the factor analysis of correspondences and of the factor discriminant analysis are then analyzed. Finally, we attempt to see the various aspects of the generalized canonical analysis (more then two sets of variables).

Among the various interests of all these techniques, the following three are particularly important:

1 — For very large data arrays (more than 50 000 individuals, for example in mail order sales), it does not seem very realistic to assume that the relations between variables or the decision rules are independant of the subpopulations.
2 — All these methods lead to more accurate local visual representations, since the local axes obtained have a larger inertia than that of the corresponding axes, for the whole population.
3 — Because of the increasing development of micro-computers, the partitioning of large data arrays into small blocks of similar behaviour with respect to the variables, can be implemented leading to easier interactive procedures on displays, and thus, giving a better assistance in the decision making processes.

2. Canonical Typological Analysis

2.1. Data:

Let $E=\{z_1, ..., z_n\}$ be the set of objects; a weight m_i is associated with each object. Each object is characterized by two groups of parameters: $\{x^1, x^2, ..., x^p\}$ and $\{y^1, ..., y^q\}$. Let X be the matrix having the parameters x^j, $j=1, ..., p$, as rows and the objects $z_1, ..., z_n$ as columns; similarly, Y is the matrix having the parameters $y^1, ..., y^q$ as rows and the objects $z_1, ..., z_n$ as columns.

Denote by x_{ij} (respectively y_{ij}) the value taken by the object z_i for the parameter x^j (respectively y^j). In other words, $X=(x_{ij})$ and $Y=(y_{ij})$.

2.2. The criterion to optimize

Let us first recall, for the case of only one cluster of elements of E, the few classical properties that will be used in canonical typological analysis.
Let A_1 (respectively A_2) be the linear projection operator into the subspace W_1 (respectively W_2), spanned by the parameters $\{x^1, ..., x^p\}$ (respectively $\{y^1, ..., y^q\}$).
Denote by $M=\begin{pmatrix} m_1 & & 0 \\ & \ddots & \\ 0 & & m_n \end{pmatrix}$ the diagonal matrix of the weights associated with each object. We have

PROPOSITION 1. The linear combinations $\xi=X^1 a$ and $\eta=Y^1 b$ such that the value of $\dfrac{\langle \xi, \eta \rangle_M}{\|\xi\|_M \|\eta\|_M}$ is maximum, are given by ξ (respectively η) associated with the largest eigenvalue λ of $A_1 \circ A_2$ (respectively $A_2 \circ A_1$).
Moreover

$$\frac{\|\xi-\hat{\xi}\|_M^2}{\|\xi\|_M^2}=1-\frac{\langle \xi, \eta \rangle_M^2}{\|\xi\|_M^2 \|\eta\|_M^2}=1-\frac{\|\hat{\xi}\|_M^2}{\|\xi\|_M^2}=1-\lambda \qquad (1)$$

where $\hat{\xi}=A_2 \xi=\sqrt{\lambda}\,\eta$.

For the proof these properties, see Caillez and Pages (1973). From Proposition 1, it follows that maximization of the correlation between ξ and η is equivalent to minimization of

$$\frac{\|\xi-\hat{\xi}\|_M}{\|\xi\|_M}$$

On the other hand, $\frac{\|\xi-\hat{\xi}\|_M}{\|\xi\|_M}$ can be expressed as a function of a; indeed, if we set: $V_{11}=XMX'$, $V_{22}=YMY'$ and $V_{21}=XMX'$, we easily get

$$\frac{\|\xi-\hat{\xi}\|_M}{\|\xi\|_M}=\frac{\|X'a-\sqrt{\lambda}\,Y'b\|_M}{\|a\|_{V_{11}}}=\frac{\|X'a-Y'V_{22}^{-1}V_{21}a\|_M}{\|a\|_{V_{11}}} \tag{2}$$

because a and b chosen to optimize (2) are related by:

$$b=\frac{1}{\sqrt{\lambda}}V_{22}^{-1}V_{21}a. \tag{3}$$

It is this last quantity which will be used to define the criterion for the canonical typological analysis.

Knowing that $A_1=X'\circ(X\circ M\circ X')^{-1}\circ X\circ M$, $A_2=Y'\circ(Y\circ M\circ Y')^{-1}\circ Y\circ M$ and that X' is one-to-one, we conclude that the eigenvalues of $A_1\circ A_2$ are the same as those of $V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}$. ($Y'$ denotes the transpose of Y). Thus, we have the following corollary of Proposition 1:

COROLLARY 1. *The vector which minimizes*

$$R(a)=\frac{\|X'a-Y'Y_{22}^{-1}V_{21}a\|_M}{\|a\|_{V_{11}}} \tag{4}$$

is the eigenvector corresponding to the largest eigenvalue of $V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}$.

In the canonical typological analysis case, notation necessitates use of an index i for each cluster; denote by $P=P_1,\ldots,P_k$ a partition of the population E into k clusters; n_i be the number of objects of cluster P_i; X_i and Y_i be the matrices deduced from X and Y by taking into account only the objects of the cluster i; $V_{i11}=X_iM_iX_i'$, $V_{i22}=Y_iM_iY_i'$ and $V_{i21}=Y_iM_iX_i'$.
With these notations, the criterion to optimize in canonical typological analysis is the mapping W: $\mathbb{R}^{pk}\times\mathbb{P}^k\rightarrow\mathbb{R}^+$ whose values are obtained via

$$W(a,P)=\sum_{i=1}^{k}\frac{\|X_i'a_i-Y_i'V_{i22}^{-1}V_{i21}a_i\|_{M_i}^2}{\|a_i\|_{V_{i11}}^2} \tag{5}$$

where $\mathbb{P}_k$ is the set of all partitions into k clusters. The problem amounts to finding $a=(a_1,\ldots,a_k)\in\mathbb{R}^{pk}$ and $P\in\mathbb{P}_k$ minimizing W.

We first restrict ourselves to the case where M_i is the diagonal matrix of weights associated with each object of the cluster P_i. We shall analyse in 2.5 a more general criterion which will allow, in particular, the use of several factors instead of considering only the one associated with the largest eigenvalue. The case where $M_i = \frac{1}{n_i} I_i$, i.e. in which M_i depends on the size of the clusters will also be considered.

2.3. The centered and the non-centered cases

In the centered case, the variables x^l and y^l are centered. In the non-centered case, they are not. The centered case corresponds to the classical canonical analysis, and the noncentered case applies to the factor analysis of correspondances. Finally, the hybrid case, in which the variables x^l are centered while the y^j's are not, applies to the discriminant analysis.

2.3.1. The non-centered case

The criterion to optimize will be used under the following form obtained from (5) by substitution according to (3):

$$W(a, P) = \sum_{i=1}^{k} \frac{\|X_i' a_i - \sqrt{\lambda}\, Y_i' b_i\|^2_{M_i}}{\|a_i\|^2_{V_{i11}}}$$

The basic idea of the algorithm is simple: Starting from a random initial guess $a^0 = (a_1^0, \ldots, a_k^0)$ and P^0, we improve, by an iterative procedure the choice of such a and P that the criterion $W(a, P)$ decreases. Before giving a complete description of the algorithm as a whole (see Section 2.3.3) we discuss some of its aspects.

Representation of a cluster: Denote by $a_i \in \mathbb{R}^p$, the kernel of the cluster i. Given a partition P, denote by g the mapping: $\mathbb{R}_k \to \mathbb{R}^{pk}$ such that $g(P) = a$, where $a = (a_1, \ldots, a_k)$ and a_i is given by the eigenvector associated with the largest eigenvalue of $V_{i11}^{-1} V_{i12} V_{i22}^{-1} V_{i21}$.
Having a, we construct a partition P with the help of a measure D of similarity of an object $z = \begin{pmatrix} x \\ y \end{pmatrix} \in R^{p+q}$ to a kernel a_i of the cluster P_i, in the following manner:

$$D(z, a_i, P_i) = \frac{(x' a_i - \sqrt{\lambda_i}\, y' b_i)^2 - \delta_i(z) \langle a_i, x\, x'\, a_i \rangle}{\frac{1}{p(z)} \|a_i\|^2_{\hat{V}_{i11}} + \langle a_i, x\, x'\, a' \rangle} \tag{6}$$

where $p(z)$ is the weight of the object, and

$$\delta_i(z) = \frac{\|\hat{X}_i' a_i - \sqrt{\lambda_i}\, \hat{Y}_i' b_i\|^2_{M_i}}{\|a_i\|^2_{\hat{V}_{i11}}} \tag{7}$$

where $\begin{pmatrix}\hat{X}_i\\ \hat{Y}_i\end{pmatrix}$ is the matrix of the coordinates of the objects of cluster i, such that z is omitted if it belongs to cluster i. Similarly, $\hat{V}_{i11}=\{V_{i11}$, without z if it belongs to the cluster $i\}$. We remark that D can take negative values (the lower the value of $D(z, a_i, P_i)$ the closer z to the cluster i).

The allocation:

Let IP be the power set of E. Denote by F the mapping from $IP_k \times IP \times IR^{pk}$ into P_k which associates with a partition a subset of E and a k-tuple of kernels $a=(a_1, ..., a_k)$, a new partition into k, at most, clusters. F is constructed like in the MDNS, see Diday et al. (1978). It allows to work on blocks of objects $H_1, ..., H_r$, where $H_i \in IP$ and $H=(H_1, ..., H_r)$ forms a partition of E.
Thus, we can avoid having the whole population in the CPU memory.
Let us now give a more precise definition of F:
$F(Q, H, a)=P$, where P is constructed using a sequence $\{\pi_n\}$ of IP_k defined as follows: Take $\pi_l=(\pi_{l1}, ..., \pi_{lk})$, $H=\{z_1, ..., z_n\}$, where $z_l \in E$, and r — the number of elements in H. We begin with $\pi_0=Q$. Let us now assume that π_{l-1} is already obtained and consider the conditions

$$D(z_l, a_i, \pi_{l-1,i}) < D(z_l, a_j, \pi_{l-1,j}), \tag{8}$$

$$D(z_l, a_i, \pi_{l-1,i}) = D(z_l, a_j, \pi_{l-1,j}) \tag{9}$$

where $z_l \in \pi_{l-1,j}$. If neither (8) is satisfied for any i nor (9) holds for each $i>j$, then we take $\pi_l=\pi_{l-1}$. Otherwise, i.e. if either there exists i such that (8) is fulfilled or (9) is true for some $i<j$, then we choose the index i by the formula

$$D(z_l, a_i, \pi_{l-1,i}) = \operatorname{Min}_r D(z_l, a_r, \pi_{l-1,r}),$$

and define $\pi_{l,i}=\pi_{l-1,i} \cup \{z_l\}$, $\pi_{l,j}=\pi_{l-1,j}-\{z_l\}$ and $\pi_{lp}=\pi_{l-1,p}$, $p \in \{1, 2, ..., k\}$ such that $p \neq i$, $p \neq j$.

Denote by $H^{(0)}, H^{(1)}, ..., H^{(N)}$ partitions of E and define the sequence $\{H^{(m)}\}$ for $m \in IN$ (m is an integer which may be larger than N) by taking $H^{(m)}=H^{(r)}$, where r is the remainder of the division of m by $N+1$.

The algorithm is based on the following sequence $\{v_m\}$:

$$v_0=(a^{(0)}, P^{(0)}), \text{ and } v_{m+1}=(a^{(m+1)}, P^{(m+1)})$$

is computed on the basis of $v_m=(a^{(m)}, P^{(m)})$ by taking

$$a^{(m+1)}=g(P^{(m)}) \text{ and } P^{(m+1)}=F(P^{(m)}, H^{(m)}, a^{(m+1)}).$$

Let $u_m=W(V_m)$, W being defined by (5) and (3).
If we add an object z to the cluster P_i, the matrices X_i', Y_i', V_{ilm}, M_i are modified, and they are denoted X_i^+, Y_i^+, V_{ilm}^+, M_i^+. Let

$$R_i^+(a_i)=\frac{\|X_i'^+ a_i-\sqrt{\lambda_i}\ Y_i'^+ b_i\|^2_{M_i^+}}{\|a_i\|^2_{V_{i11}^+}}$$

i.e. the modified counterpart of $R_i(a_i)$ defined analogously as in (4), using (3):

$$R_i(a_i)=\frac{\|X_i' a_i-\sqrt{\lambda_i}\ Y_i' b_i\|^2_{M_i}}{\|a_i\|^2_{V_{i11}}} \tag{10}$$

With the above, we have the following result:

LEMMA 1. *If z does not belong to cluster P_i, then*

$$R_i^+(a_i)=R_i(a_i)+D(z, a_i, P_i)$$

Proof. Evidently

$$\|a_i\|^2_{V^+_{i11}}=\langle a_i, V^+_{i11} a_i\rangle=\langle a_i, (V_{i11}+p(z)\,xx')\,a_i\rangle=$$

$$=\|a_i\|^2_{V_{i11}}+p(z)\,\|a_i\|^2_{xx'} \tag{11}$$

On the other hand:

$$\|X_i^+ a_i-\sqrt{\lambda_i}\ Y_i^+ b_i\|^2_{M_i}=\|X_i' a_i-\sqrt{\lambda_i}\ Y_i' b_i\|^2_{M_i}+(x' a_i-\sqrt{\lambda_i}\ y' b_i)^2\,p(z) \tag{12}$$

where $p(z)$ is the weight of the object z,
Thus, using (11), (12) and (10), we get

$$R_i^+(a_i)=\frac{\|X_i' a_i-\sqrt{\lambda_i}\ Y_i' b_i\|^2_{M_i}+(x' a_i-\sqrt{\lambda_i}\ y_i' b_i)^2\,p(z)}{\|a_i\|^2_{V_{i11}}+p(z)\,\|a_i\|}=R_i(a_i)+$$

$$+\frac{p(z)\,(x' a_i-\sqrt{\lambda_i}\ y_i' b_i)^2\,\|a_i\|^2_{V_{i11}}-\|X_i' a_i-\sqrt{\lambda_i}\ Y_i' b_i\|^2_{M_i}\,p(z)\,\|a_i\|^2_{xx'}}{\|a_i\|^2_{V_{i11}}(\|a_i\|^2_{V_{i11}}+p(z)\,\|a_i\|^2_{xx'})}$$

which gives the result required according to (6) and (7). ■
Therefrom

$$R_i^+(a_i)=R_i(a_i)+\frac{p(z)\,(x' a_i-\sqrt{\lambda_i}\ y_i' b_i)^2-p(z)\,\|a_i\|^2_{x'x}}{\|a_i\|^2_{V_{i11}}+p(z)\,\|a_i\|^2_{xx'}}$$

Now, we formulate

PROPOSITION 2. The sequence $\{u_m\}$ is monotonically decreasing and convergent.

Proof. Let us show that $u_m<W(a^{(m)}, p^{(m-1)})\leqslant u_{m-i}$
We have: $W(a^{(m)}, P^{(m-1)})\leqslant u_{m-1}=W(a^{(m-1)}, P^{(m-1)})$ for

$$R_i(a_i^{(m)})=\frac{\|X_i' a_i^{(m)}-Y_i' V_{i22}^{-1} a_i^{(m)}\|^2_{M_i}}{\|a_i^{(m)}\|^2_{V_{i11}}}\leqslant$$

$$\leqslant\frac{\|X_i' a_i^{(m-1)}-Y_i' V_{i22}^{-1} V_{i21} a_i^{(m-1)}\|^2_{M_i}}{\|a_i^{(m-1)}\|^2_{V_{i11}}}=R_i(a_i^{(m-1)})$$

because $a_i^{(m)}$ is an eigenvector of $V_{i11}^{-1} V_{i12} V_{i22}^{-1} V_{i21}$, by definition, and due to Corollary 1, it minimizes $R_i(a)$. Thus, it remains to show that $u_m=W(a^{(m)}, P^{(m)})\leqslant$

$\leqslant W(a^{(m)}, P^{(m-1)})$. Two cases can occur: either both of the objects remain in their respective cluster, that is $P^{(m)}=P^{(m-1)}$, in which case $W(a^{(m)}, P^{(m)})=W(a^{(m)}, P^{(m-1)})$, or at least one object z moves to another cluster. In other words, there exist two indices i and l such that (8) or (9) holds for a replaced by $a^{(m)}$ (with respective subscripts, of course), because we known that we go from $P^{(m-1)}$ to $P^{(m)}$ by means of a sequence $\{\pi_m\}$.

In this case, the object z_l moves to another cluster and two terms are modified in the criterion W. Namely, $R_j(a_j^{(m)})$ which previously contained z_l and $R_l(a_i^{(m)})$ which now contains it, since z_l is transferred to the ith from the jth component. The contribution of z_l in each of these terms is, from Lemma 1, precisely:

$$D(z_l, a_l, \pi_{l-1,l}) \text{ and } D(z_l, a_j, \pi_{l-1,j}),$$

respectively.

Therefore using again (8) and (9), we get that the sequence $\{u_m^0\}$ is monotonically decreasing. Its convergence follows from the fact that W is positive. ■

2.3.2. The centered case

Denote by $\bar{X}_l^+$ (respectively $\bar{X}_l^-$) the matrix obtained after centering the rows of X_l^+ (respectively X_l^-). If the object $z=\begin{pmatrix} x \\ y \end{pmatrix}$ is added to the cluster P_l, we have:

$$\left.\begin{aligned} \bar{X}_l^+ &= \{X_l^+ \text{ centered}\} = \overbrace{(X_l\, x)}^{\text{centered rows}} \\ \bar{X}_l^- &= \{X_l^- \text{ centered}\} = \overbrace{(X_l\, x \text{ omitted})}^{\text{centered rows}} \end{aligned}\right\} \text{similar definitions for } \bar{Y}_l^+ \text{ and } \bar{Y}_l^-$$

Let

$$\bar{V}_{i11}^+ = \bar{X}_i^+ M_i^+ \bar{X}_i^+, \text{ where } M_i^+ = \begin{pmatrix} M_l & 0 \\ 0 & p(z) \end{pmatrix}$$

and

$$\bar{R}_l^+(a_l) = \frac{\|\bar{X}_l'^+ a_l - \sqrt{\lambda_l}\ \bar{Y}_l'^+ b_l\|_{M_l^+}^2}{\|a_l\|_{\bar{V}_{i11}^+}^2}$$

$\bar{R}_l^-$ is defined similarly, by replacing everywhere "+" by "−", and (10) with bars above X, Y and V gives $\bar{R}_l(a_l)$. The representation function g is here the same as in the non centered case. However, as far as the allocation function F is concerned, we must modify the definition of D. Let $\bar{D}$ be the new allocation function such that:

$$\bar{D}(z, a_l, P_l) = (\bar{R}_l^+(a_l) - \bar{R}_l(a_l)) + (\bar{R}_j^-(a_j) - \bar{R}_j(a_j)),$$

where j is the index of the cluster to which z belongs.

The function F is defined as in the non centered case, with the difference that one object z belonging to cluster j is transferred to cluster i if and only if $\bar{D}(z, a_i, P_i)$ is minimum and negative.

If we define the sequences $\{\bar{u}_n\}$ and $\{\bar{v}_n\}$ similarly as in the non-centered case, we can easily show that $\{\bar{u}_n\}$ is monotonically decreasing and convergent, since

$$W(a, P)=\sum_{\substack{l=1\\ l\neq i\\ l\neq j}}^{k} \bar{R}_l(a_l)+\bar{R}_i(a_i)+\bar{R}_j(a_l)$$

If we denote by Q the partition P modified after z has been transferred from cluster j to cluster i, we have:

$$W(a, Q)=\sum_{\substack{l=1\\ l\neq i\\ l\neq j}}^{k} \bar{R}_l(a_l)+\bar{R}_i^{+}(a_i)+\bar{R}_j^{-}(a_j)$$

It follows that $D(z, a_l, P_l)=W(a, Q)-W(a, P)$. Therefore, when an object goes from one cluster to another, the criterion decreases.

2.3.3. The algorithm

We can summarize the algorithm in both the centered and the non-centered cases by the four following steps. First decompose the population into groups $\{H^{(0)}, H^{(1)}, ..., H^{(N)}\}$, and let $n=0$.

1. We start from the partition $P^{(n)}=(P_1^{(n)}, ..., P_k^{(n)})$.
(It can be estimated or randomly chosen).
2. With the help of k canonical analyses on the clusters $P_i^{(n)}$, we obtain k canonical factors $a^{(n)}=(a_1^{(n)}, ..., a_k^{(n)})$.
3. We consider each individual z of the group $H^{(n)}\subset E$. In the non-centered case, we assign it to cluster i if:

$$D(z, a_i^{(n)}, P_i^{(n)})=\operatorname{Min}_l D(z, a_l^{(n)}, P_l^{(n)})$$

In the centered case, we assign it to cluster i if

$$\bar{D}(z, a_l, P_i)=\operatorname{Min}_l \bar{D}(z, a_l, P_l)<0$$

This way a new partition $P^{(n+1)}$ is obtained,
4. We set $n=n+1$ and we start again from 2, until $P^{(n)}=P^{(n+1)}$ is reached.

REMARKS.
a) If $n>N$, we start again from $H^{(0)}$, that is $H^{(n+1)}=H^{(0)}$, $H^{(n+2)}=H^{(r)}$, ... • etc
b) If the size of the data is not too large (everything can be stored in the memory

of CPU), we take $H^{(0)}=E$ and we iterate on the whole population (i.e. $H^{(n)}=E$, $\forall n$) until convergence.

c) If the data array is too large, and thus it requires a buffer memory, it is better to avoid reducing each group $H^{(n)}$ to a single element of E, since otherwise we have too many diagonalizations.

d) Whether we take $H^{(n)}=E$ or not, the result obtained depends on the order of the objects of E.

2.4. Comparison between the centered and the non-centered case

In the centered case, the canonical axes defined by the vectors ξ_{ij} and η_{ij} (associated with the cluster P_i of the partition we are looking for), are centered. In the non-centered case, they are not. To avoid this problem, two classical techniques can be used:

a) We can assign an additional parameter to each one of the two parameter groups. This new parameter is an $n\times 1$ vector with all coordinates equal to 1.

b) If the parameters are sufficiently homogeneous, we can divide each coordinate of a given parameter by the sum of the values taken by this parameter over all the objects.

In both cases, $W_1 \cap W_2$ contains the vector $\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}$ $\uparrow$ n $\downarrow$

Thus $\xi_{i1}=\eta_{i1}=\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}$, and since ξ_{ij} (respectively η_{ij}) is M-orthogonal to ξ_{i1} (respectively η_{i1}), the ξ_{ij}'s and the η_{ij}'s are centered.

REMARK. Centering the canonical factors does not imply that the recoded parameters $a_{ijl}\,X^l$ (a_{ijl} is the l'th coordinate of the factor j associated with the cluster ij) will be centered. However, this technique enables us to center the mean values $a_{ijl}\,X^l$.

Indeed,

$$\xi_{ij}=\sum_{l=1}^{p} a_{ijl}\,X^l \text{ implies } \xi_{ij}=\sum_{l=1}^{p} a_{ijl}\,\bar{x}^l=0$$

because ξ_{ij} being centered, so is its mean value.

c) In the non centered case, when the covariances of the parameters of each cluster are computed, greater weights should be given to the coordinates which have large absolute values.

d)* In the case where M is the weighting matrix of the objects, we have:

$$V_{i11}=\begin{pmatrix} X_i \\ \bar{1} \end{pmatrix} M (X_i' \bar{1}') \text{ where } \bar{1}'=(\underbrace{1, 1, ..., 1}_{n \text{ times}})$$

Therefore:

$$V_{i11}=\begin{pmatrix} X_i M X_i' & \tilde{m} \\ \tilde{m}^T & 1 \end{pmatrix} \text{ where } \tilde{m}=\begin{pmatrix} \tilde{m}^1 \\ \vdots \\ \tilde{m}^p \end{pmatrix}$$

with $\tilde{m}^j=\sum_{i=1}^{n} p_i x_i^j$

where p_i is the weight associated with the ith object. If we denote by V_{i11}^c the matrix corresponding to V_{i11} in the centered case, we can easily show that:

$$V_{i11}^c=X_i M X_i'-\tilde{m}\tilde{m}'$$

But the inverse of a matrix of the form $\begin{pmatrix} A & B \\ B' & C \end{pmatrix}$ is $\begin{pmatrix} (A-BC^{-1}B')^{-1} & \cdot \\ \cdot & (C-BA^{-1}B^{-1})^{-1} \end{pmatrix}$, where only the diagonal terms are shown, for simplicity.

Consequently: V_{i11}^{-1} and $(V_{i11}^c)^{-1}$ coincide on their upper diagonal block.

2.5. Determination of several canonical factors in each cluster

We generalize the formulation of the criterion of the algorithm by introducing an additional index, associated with the factors we are looking for. Denote by a_{ij} the factor j, associated with cluster i and $a_i=(a_{i1}, a_{i2}, ..., a_{im})$, where m is the number of desired factors $(m<Min(p, q))$, which must be fixed and remain the same for all the clusters. Hence, a_{ij} is a $p\times 1$ vector, so a_i is a $p\times m$ matrix (p rows, m columns). Let $a=(a_1, ..., a_k)$ be the matrix (p rows and $k\times m$ columns) representing the factors associated with the k clusters. Let λ_{ij} denote the jth largest eigenvalue of $V_{i11}^{-1} V_{i12} V_{i22}^{-1} V_{i21}$.

Similarly as in (3) we get

$$b_{ij}=\frac{1}{\sqrt{\lambda_{ij}}} V_{i22}^{-1} V_{i21} a_{ij} \tag{13}$$

Finally, the criterion to minimize can be expressed under the more general following form:

$$W_m(a, P)=\sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\|X_i' a_{ij}-Y_i' V_{i22}^{-1} V_{i21} a_{ij}\|_{M_i}^2}{\|a_{ij}\|_{V_{i11}}^2} \tag{14}$$

It remains to define the mappings D, F and g.

Non centered case

If we now apply the previous proofs, but this time taking the index j into account, we obtain:

* I would like to thank M. Baudoin who brought this remark to my attention.

$$D_m(z, a_t, P_t) = \sum_{j=1}^{m} \frac{p(z)(x' a_{tj} - \sqrt{\lambda_{tj}}\, y' b_{tj})^2 - \delta_{tj}(z)\langle a_{tj}, x M x' a_{tj}\rangle}{\|a_{tj}\|^2_{\hat{V}_{t11}} + \langle a_{tj}, x M x' a_{tj}\rangle}$$

where

$$\delta_{tj}(z) = \frac{\|\hat{X}'_t a_{tj} - \sqrt{\lambda_{tj}}\, \hat{Y}'_t b_{tj}\|^2_M}{\|a_{tj}\|^2_{\hat{V}_{t11}}} \tag{15}$$

and the circumflex has the same meaning as described before, (7). We recall that M is the diagonal matrix of the weights of the objects ($p(z)$ dor the object z).

REMARK. If we want the weights of the objects to be dependent on the size of the clusters, we set $p(z) = \frac{1}{n_j}$ if $z \in P_j$.

In this case, D_m can be written as:

$$D_m(z, a_t, P_t) = \sum_{j=1}^{m} \frac{(x' a_{tj} - \sqrt{\lambda_{tj}}\, y' b_{tj})^2 - \delta_{tj}(z)\langle a_{tj}, xx' a_{tj}\rangle}{\|a_{tj}\|^2_{\hat{V}_{t11}} + \langle a_{tj}, xx' a_{tj}\rangle} \tag{16}$$

where $\hat{V}_{t11} = X_t X'_t$ with x omitted if it belongs to cluster i.

Centered case

The decision function is defined by:

$$\bar{D}(z, a_t, P_t) = (\bar{R}^+_t(a_t) - \bar{R}_t(a_t)) + (\bar{R}^-_j(a_j) - \bar{R}_j(a_j))$$

where $z \in P_j$, with:

$$\bar{R}^+_t(a_t) = \sum_{j=1}^{m} \frac{\|\bar{X}'^+_t a_{tj} - \sqrt{\lambda_{tj}}\, \bar{Y}'^+_{tj} b_{tj}\|^2_{M^+_t}}{\|a_{tj}\|^2_{V^+_{t11}}}$$

$R^-_t(a_t)$ is defined in the same way, by replacing everywhere + by −, and

$$\bar{R}_t(a_t) = \sum_{j=1}^{m} \frac{\|\bar{X}'_t a_{tj} - \sqrt{\lambda_{tj}}\, \bar{Y}'_{tj} b_{tj}\|^2_{M_t}}{\|a_{tj}\|^2_{V_{t11}}}$$

The mapping F is defined as in 2.3.1 and 2.3.2 for the centered and the non centered cases, respectively.

The mapping G: $IP_k \to IR^{npk}$ with values $g(P) = a = (a_1, \ldots, a_k)$, where $a_i = (a_{i1}, \ldots, a_{im})$ represents the m eigenvectors associated with the largest eigenvalues of $V^{-1}_{t11} V_{t12} V^{-1}_{t22} V_{t21}$.

2.6. Some numerical aspects

2.6.1. Problem of non symmetry of $V^{-1}_{t11} V_{t12} V^{-1}_{t22} V_{t21}$

We set $CC' = V^{-1}_{t11}$ and $V_{t12} V^{-1}_{t22} V_{t21} = Z$. It follows that $CC'Z = \lambda a$ (in this section, we shall omit the indices i and j for λ and a, for simplicity).

Let $a=C\alpha$. Hence, $CC'ZC\alpha=\lambda C\alpha$. This equation has the same solutions as $CC'Z=\lambda a$. So we have to compute the eigenvectors of the symmetric matrix $C'ZC$. Having α, we easily determine a by the equality $a=C\alpha$.

Orthonormalization:

We easily verify that if α_1 and α_2 are two distinct eigenvectors of $C'ZC$ and $a_{i1}=C\alpha_1$, $a_{i2}=C\alpha_2$, then $\langle a_{i1}, a_{i2}\rangle_{V_{i11}}=0$.

Indeed $\langle a_{i1}, a_{i2}\rangle_{V_{i11}}=\langle C\alpha_1, C\alpha_2\rangle_{V_{i11}}=\langle \alpha_1, C'C\alpha_2\rangle_{V_{i11}}=\langle \alpha_1, V_{i11}^{-1}\alpha_2\rangle_{V_{i11}}=\langle \alpha_1, \alpha_2\rangle=0$ for α_1 and α_2 are two distinct eigenvectors of the same symmetric matrix.

Normalisation: Let $\hat{\alpha}=\dfrac{\alpha}{\|\alpha\|}$ and $a=C\hat{\alpha}$, then

$$\|\xi\|_{D_p}=\langle a, a\rangle_{V_{i11}}=\|a\|^2_{V_{i11}}=\langle C\hat{\alpha}, C\hat{\alpha}\rangle_{V_{i11}}=\langle \hat{\alpha}, (C)'\, C\hat{\alpha}\rangle_{V_{i11}}=$$

$$=\langle \hat{\alpha}, (CC')\,\hat{\alpha}\rangle_{V_{i11}}=\langle \hat{\alpha}, T^{-1}\,\hat{\alpha}\rangle_{V_{i11}}=\langle \hat{\alpha}, \hat{\alpha}\rangle=\frac{1}{\|\alpha\|^2}\langle \alpha, \alpha\rangle=1$$

3. Factor analysis of the local correspondances

3.1. Statement of the problem

We now consider the non centered case, and X and Y represent two qualitative variables, $(x^1, ..., x^p)$ on one hand and $(y^1, ..., y^q)$ on the other hand are their respective characteristic variables. We will show that the criterion is a simple function of the sum of k χ^2's with $(p-1)(q-1)$ degrees of freedom, which must be maximized.

3.2. Expression of the criterion to maximize

We assign the weight $\dfrac{1}{n}$ to each object (n is the total number of objects in the whole population).

Denote by M_i the diagonal matrix of the weights of the objects of cluster i, and set $M_i=\dfrac{1}{n}I_i$, where I_i is the unit matrix of dimensions card $P_i\times$card P_i. Take $V_{i11}=X_i M_i X_i'$ and $V_{i22}=Y_i M_i Y_i'$, where the X_i's and the Y_i's are not centered. If we denote by D_{X_i} (respectively D_{Y_i}) the diagonal matrices of the modality frequencies in cluster i and by S_i the cross-contingency matrix of X_i and Y_i, we get

$$V_{i11}=D_{X_i}=\begin{pmatrix}\dfrac{n_1^i}{n_i} & & 0\\ & \ddots & \\ 0 & & \dfrac{n_p^i}{n_i}\end{pmatrix} \tag{17}$$

$$V_{i22}=D_{Y_i}=\begin{pmatrix}\dfrac{n_1^i}{n_i} & & 0\\ & \ddots & \\ 0 & & \dfrac{n_q^i}{n_i}\end{pmatrix} \tag{18}$$

where n_i=card P_i, n_l^i is the number of times the modality l of the variable X is attained in cluster i; n_{lm}^i is the number of times the modality l of the variable X and the modality m of the variable Y are simultaneously attained in cluster i. Then

$$V_{i12}=X_i M_i Y_i'=S_i=\left(\frac{n_{lm}^i}{n_i}\right)_{\substack{l=1,\ldots,p\\ m=1,\ldots,q}} \tag{19}$$

$$V_{i21}=S_i' \tag{20}$$

Therefore, we obtain:

$$V_{i11}^{-1} V_{i12} V_{i22}^{-1} v_{i21}=D_{X_i}^{-1} S_i D_{Y_i}^{-1} S_i' \tag{21}$$

i.e.

$$D_{X_i}^{-1} S_i D_{Y_i}^{-1} S_i' a_{ij}=\lambda_{ij} a_{ij}$$

and, by (13)

$$b_{ij}=\frac{1}{\sqrt{\lambda_{ij}}} V_{i22}^{-1} V_{i21} a_{ij}=\frac{1}{\sqrt{\lambda_{ij}}} D_{Y_i}^{-1} S_i' a_{ij}.$$

The criterion to minimize is, see (14),

$$W_m(a,P)=\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\|X_i' a_{ij}-Y' V_{i22}^{-1} V_{i21} a_{ij}\|_{M_i}^2}{\|a_{ij}\|_{V_{j11}}^2}=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\|X_i' a_{ij}-Y_i' D_{Y_i}^{-1} S_i' a_{ij}\|_{M_i}^2}{\|a_{ij}\|_{D_{X_i}}^2}$$

Using (17)–(21), we get

$$\|X_i' a_{ij}-Y_i' D_{Y_i}^{-1} S_i' a_{ij}\|_{M_i}^2=$$

$$=\langle (X_i'-Y_i' D_{Y_i}^{-1} S_i') a_{ij}, M_i (X_i'-Y_i' D_{Y_i}^{-1} S_i') a_{ij}\rangle=$$

$$=\langle a_{ij}, (X_i-S_i D_{Y_i}^{-1} Y_i) M_i (X_i'-Y_i' D_{Y_i}^{-1} S_i') a_{ij}\rangle=$$

$$=\langle a_{ij}, (X_i M_i X_i'-X_i M_i Y_i' D_{Y_i}^{-1} S_i'-S_i D_{Y_i}^{-1} Y_i M_i X_i'+$$

$$+S_i D_{Y_i}^{-1} Y_i M_i Y_i' D_{Y_i}^{-1} S_i') a_{ij}\rangle=$$

$$=\langle a_{ij}, (D_{X_i}-S_i D_{Y_i}^{-1} S_i'-S_i D_{Y_i}^{-1} S_i'+S_i D_{Y_i}^{-1} S_i') a_{ij}\rangle=$$

$$=\langle a_{ij}, (D_{X_i}-S_i D_{Y_i}^{-1} S_i') a_{ij}\rangle=\|a_{ij}\|_{D_{X_i}}^2-\langle a_{ij}, S_i D_{Y_i}^{-1} S_i' a_{ij}\rangle$$

That result combined with (14) gives

$$W_m(a,P)=mk-\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{1}{\|a_{ij}\|_{D_{X_i}}}\langle a_{ij}, S_i D_{Y_i}^{-1} S_i' a_{ij}\rangle$$

3.3. Relation between the criterion to minimize and the contingency χ^2

From (1), (2), Corollary 1 and Proposition 1, it follows that (14) can be rewritten as

$$W_m(a, P)=km-\sum_{i=1}^{k}\sum_{j=1}^{m}\lambda_{ij}=km-k-\sum_{i=1}^{k}\sum_{j=2}^{m}\lambda_{ij} \tag{22}$$

for $\lambda_{i1}=1\ \forall i$. Therefore

$$W_m(a, P)=k(m-1)-\sum_{i=1}^{k}\sum_{j=2}^{m}\lambda_{ij}.$$

But the trace of the matrix to diagonalize, i.e.

$$D_{X_i}^{-1}S_i D_{Y_i}^{-1}S_i', \text{ is equal to } \sum_{j=1}^{p}\lambda_{ij} \tag{23}$$

which can also be written as

$$\sum_{j=1}^{p}\sum_{l=1}^{q}\frac{p_{jl}^{i2}}{p_{j.}^{i}p_{.l}^{i}}, \text{ if we set } p_{jl}^{i}=\frac{n_{jl}^{i}}{n^{i}}, p_{j.}^{i}=\sum_{l=1}^{q}p_{jl}^{i},$$

and $p_{.l}^{i}=\sum_{j=1}^{p}p_{jl}^{i}$.

On the other hand, the term $\Phi^2=\frac{\chi^2}{n}$ of a contingency array of generator p_{ij} is:

$$\frac{\chi_i^2}{n^i}=\Phi_i^2=\sum_{j=1}^{p}\sum_{l=1}^{q}\frac{(p_{jl}^{i}-p_{j.}^{i}p_{.l}^{i})^2}{p_{j.}^{i}p_{.l}^{i}}=\sum_{j=1}^{p}\sum_{l=1}^{q}\frac{(p_{jl}^{i})^2}{p_{j.}^{i}p_{.l}^{i}}-1=\sum_{j=2}^{p}\lambda_{ij}$$

by (23). Finally, we have

$$W_m(a, P)=k(p-1)-\sum_{i=1}^{k}\Phi_i^2$$

3.4. The algorithm

The algorithm is the same as that indicated in 2.3.3., with the important difference (from the programmer's point of view) that only the spaces of frequency XMY' of the array intervene $\left(M=\frac{1}{n}I, \text{ where } I \text{ is the } n\times n \text{ unit matrix}\right)$. Namely, when the allocation is made, we use (16) and (15), where $\hat{V}_{i11}$ is obtained from (17). Moreover, knowing the position of the space we immediately deduce the value of the vectors x and y.

EXAMPLE: Suppose that $p=4$ and $q=5$. Then XMY' is an 4×5 array of frequency. The objects associated with the space corresponding to the 2nd row and the 3rd column of this array have the coordinates $z=\begin{pmatrix}x\\y\end{pmatrix}$ where

$$x=\begin{pmatrix}0\\1\\0\\0\end{pmatrix} \quad \text{and} \quad y=\begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix}, \text{ for example.}$$

To compute the factors a_i, we must diagonalize the matrices $D_{X_i}^{-1} S_i D_{Y_i}^{-1} S_i'$ which depend only on the spaces of the frequency array XMY' defined by the initial data, since all the objects associated with the same space of this array have the same coordinates and therefore belong to the same cluster. In fact, we now see that the problem consists in starting from a frequency matrix, then decomposing it into a sum of matrices such that the sums of the Φ^2's associated with these matrices are maximum.

REMARK. The matrix to diagonalize, $D_{X_i}^{-1} S_i D_{Y_i}^{-1} S_i'$, is not symmetric, in general. In order to use the results of the symmetric case, we just have to set: $a_{ij}=D_{X_i}^{-\frac{1}{2}} \alpha_{ij}$. It follows that:

$$D_{X_i}^{-1} S_i D_{Y_i}^{-1} S_i' D_{X_i}^{-\frac{1}{2}} \alpha_{ij}=\lambda_i D_{X_i}^{-\frac{1}{2}} \alpha_{ij}, \text{ i.e. } D_{X_i}^{-\frac{1}{2}} S_i D_{Y_i}^{-1} S' D_{X_i}^{-\frac{1}{2}} \alpha_{ij}=\lambda_i \alpha_{ij}$$

Since $S_i D_{Y_i}^{-1} S'$ is symmetric, α_{ij}'s are the eigenvectors of a symmetric matrix. In order to orthonormalize the a_{ij}'s ($j=1, \ldots, p$), we just have to set:

$$a_{ij}=D_{X_i}^{-\frac{1}{2}} \hat{\alpha}_{ij}, \quad \text{with } \hat{\alpha}_{ij}=\frac{\alpha_{ij}}{\|\alpha_{ij}\|}$$

Thus, we have:

$$\langle a_{ij}, a_{il}\rangle_{D_{X_i}}=\langle D_{X_i}^{-\frac{1}{2}} \hat{\alpha}_{ij}, D_{X_i}^{-\frac{1}{2}} \hat{\alpha}_{il}\rangle_{D_{X_i}}=\langle \hat{\alpha}_{ij}, D_{X_i}^{-1} \hat{\alpha}_{il}\rangle_{D_{X_i}}=\langle \hat{\alpha}_{ij}, \hat{\alpha}_{il}\rangle=0$$

since $\hat{\alpha}_{ij}$ is an eigenvector of a symmetric matrix. Moreover,

$$\langle a_{ij}, a_{ij}\rangle_{D_{X_i}}=\langle \hat{\alpha}_{ij}, \hat{\alpha}_{ij}\rangle=1.$$

3.5. Maximization of the criterion

We have to decompose the initial matrix into a sum of matrices such that their rows and their columns are as independent as possible. In some cases, this problem has a significant importance. To solve it we can use the same algorithm, inverting the inequalities, when the assignment of the space is made and taking the m eigenvectors of smaller eigenvalue, when the computation of the factors is made.

3.6. Another possible algorithm

We can decompose the initial contingency matrix into a sum of k matrices chosen randomly. Then, the algorithm consists in transferring the spaces which are full from one matrix to the other, as long as the criterion improves. It remains to define a good heuristic procedure for these transfers, because the number of possibilities is very large.

4. Local discriminant analysis

4.1. The problem

We have to apply the typological canonical analysis to the particular case where the second group of variables Y represents a qualitative variable. More precisely, $Y=(y^1, ..., y^q)$ where the y^j are the characteristics of Y.

EXAMPLE. The qualitative variable $\begin{pmatrix}1\\2\\1\\3\\3\\2\end{pmatrix}$ with three modalities can be written as:

$$Y^1=\begin{pmatrix}1&0&0\\0&1&0\\1&0&0\\0&0&1\\0&0&1\\0&1&0\end{pmatrix}=(y^1, y^2, y^3)$$

with

$$y^2=\begin{pmatrix}0\\1\\0\\0\\0\\1\end{pmatrix}$$

We will show that minimization of the criterion amounts to the search for local discriminant axes.

EXAMPLE. If $K=2$, we have to find the clusters P_1 and P_2 and the axes A_1 and A_2 which discriminate in the best possible way the 4 clusters defined by the variable Y.

4.2. Expression of the criterion to minimize

Let $M_l=\frac{1}{n} I_l$, where I_l is the card $P_l \times$ card P_l unit matrix. $V_{l11}=X_l M_l X_l'=V_l$ is the covariance matrix of the set of the objects of cluster P_l, since X_l is centered. $V_{l22}=Y_l M_l Y_l'=D_{Yl}=\begin{pmatrix} \frac{n_1^l}{n} & & 0 \\ & \ddots & \\ 0 & & \frac{n_q^l}{n} \end{pmatrix}$ is the diagonal matrix of the modality frequencies of the variable Y in the cluster l (Y_l is not centered). $V_{l12}=X_l M_l Y_l'=$ $=G_l D_{Y_l}$, and $G_l=(g_l^1, \ldots, g_l^q)$ with $g_l^j=\begin{Bmatrix} \frac{1}{n_j^l} \sum_{i \in [P_l] \cap I_j} x_{i1} \\ \vdots \\ \frac{1}{n_j^l} \sum_{i \in [P_l] \cap I_j} x_{ip} \end{Bmatrix}$ where $[P_l]$ is the set of indices of the objects of cluster P_l, and I_j is the set of indices of the objects which take the modality j of the variable Y. The criterion to minimize has the form

$$W_m(a, P)=\sum_{l=1}^{k} \sum_{j=1}^{m} \frac{\|X_l' a_{lj}-Y_l' V_{l22}^{-1} V_{l21} a_{lj}\|_{M_l}^2}{\|a_{lj}\|_{V_{l11}}^2}$$

Replacing the V_{llj}'s by their values, we obtain:

$$W_m(a, P)=\sum_{l=1}^{k} \sum_{j=1}^{m} \frac{\|X_l' a_{lj}-Y_l' D_{Y_l}^{-1} D_{Y_l} G_l' a_{lj}\|_{M_l}^2}{\|a_{lj}\|_{V_{l11}}^2}$$

But

$$\|X_l' a_{lj}-Y_l' G_l' a_{lj}\|_{M_l}^2=\langle (X_l'-Y_l' G_l') a_{lj},\ M_l (X_l'-Y_l' G_l') a_{lj}\rangle=$$

$$=\langle a_{lj}, (X_l-G_l Y_l) M_l (X_l'-Y_l' G_l') a_{lj}\rangle=$$

$$=\langle a_{lj}, (X_l M_l X_l'-X_l M_l Y_l' G_l'-G_l Y_l M_l X_l'+G_l Y_l M_l Y_l' G_l') a_{lj}\rangle=$$

$$=\langle a_{lj}, (V_l-G_l D_{Y_l} G_l'-G_l D_{Y_l} G_l'+G_l D_{Y_l} G_l') a_{lj}\rangle=$$

$$=\langle a_{lj}, (V_l-G_l D_{Y_l} G_l') a_{lj}\rangle=\|a_{lj}\|_{V_l}^2-\|a_{lj}\|_{G_l D_{Y_l} G_l'}^2$$

$B_l=G_l D_{Y_l} G_l'$ is the covariance matrix of the centers of gravity, in other words, the interclass covariance matrix. Finally:

$$W_m(a, P)=\sum_{l=1}^{k} \sum_{j=1}^{m} \frac{\|a_{lj}\|_{V_l}^2-\|a_{lj}\|_{B_l}^2}{\|a_{lj}\|_{V_l}^2}$$

that is,

$$W_m(a, P)=\mathrm{km}-\sum_{l=1}^{k} \sum_{j=1}^{m} \frac{\|a_{lj}\|_{B_l}^2}{\|a_{lj}\|_{V_l}^2}$$

REMARKS. In order to compute the factors a_{lj} we must diagonalize the matrix $V_{l11}^{-1} V_{l12} V_{l22}^{-1} V_{l21} = V_l^{-1} G_l D_{Y_l} D_{Y_l}^{-1} D_{Y_l} G_l' = V_l^{-1} B_l$. Thus: we have

$$V_l^{-1} B_l a_{lj} = \lambda_{lj} a_{lj}, \text{ i.e. } \frac{\|a_{lj}\|_{B_l}^2}{\|a_{lj}\|_{V_l}^2} = \lambda_{lj}.$$

Therefore, $W_m(a, P) = \text{km} - \sum_{l=1}^{k} \sum_{j=1}^{m} \lambda_{lj}$, and $\lambda_{lj} = \frac{\|a_{lj}\|_{B_l}^2}{\|a_{lj}\|_{V_l}^2} < 1$. So, minimizing the critetion $W_m(a, P)$ is equivalent to maximizing the sum of the λ_{lj}'s. Therefore, we have to find the linear combinations $\xi_{lj} = X_l a_{lj}$ which maximize the interclass variance and minimize the total variance, or equivalently, from Huygens' theorem, which minimize the intra-class variance.

4.3. The algorithm

We have to use the algorithm which was defined for the non centered case. Here, the allocation function is:

$$\hat{D}(z, a_l, P_l) = (\hat{R}_l^+(a_l) - \hat{R}_l(a_l)) + (\hat{R}_s^-(a_s) - \hat{R}_s(a_s)).$$

where s belongs to cluster P_s, and

$$\hat{R}_l(a_l) = \sum_{j=1}^{m} \frac{\|\bar{X}_l' a_{lj} - \sqrt{\lambda_{lj}}\, Y_l' b_{lj}\|_{M_l}^2}{\|a_{lj}\|_{\bar{V}_{l11}}^2}$$

where $b_{lj} = \frac{1}{\sqrt{\lambda_{lj}}} V_{l22}^{-1} \hat{V}_{l21} a_{lj} = \frac{1}{\sqrt{\lambda_{lj}}} \bar{G}_l' a_{lj}$, with $\hat{V}_{l21} = Y_l M_l \bar{X}_l'$ and $\bar{G}_l$ defined by $X_l M_l Y_l' = \bar{G}_l D_l$.

Using the same type of computations as in 4.2, we obtain:

$$\hat{R}_l(a_l) = \sum_{j=1}^{m} \frac{\|\bar{X}_l' a_{lj} - Y_l' \bar{G}_l' a_{lj}\|_{\bar{B}_l}^2}{\|a_{lj}\|_{\bar{V}_l}^2} =$$

$$= \sum_{j=1}^{m} \frac{\|a_{lj}\|_{\bar{V}_l}^2 - \|a_{lj}\|_{\bar{B}_l}^2}{\|a_{lj}\|_{\bar{V}_l}^2} = m - \sum_{j=1}^{m} \frac{\|a_{lj}\|_{\bar{B}_l}^2}{\|a_{lj}\|_{\bar{V}_l}^2},$$

where $\bar{B}_l = \bar{G}_l D_{Y_l} \bar{G}_l'$ and $V_l = \bar{X}_l M_l \bar{X}_l'$.

Similarly

$$\hat{R}_l^+(a_l) = m - \sum_{j=1}^{m} \frac{\|a_{lj}\|_{\bar{B}_l^+}^2}{\|a_{lj}\|_{\bar{V}_l^+}^2}, \quad \text{where } \bar{B}_l^+ = \bar{G}_l^+ D_{Y_l}^+ \bar{G}_l^{+\prime},$$

$D_{Y_l}^+ = Y_l^+ M_l^+ Y_l^{+\prime}$, $\bar{G}_l^+$ is given by $X_l^+ M_l^+ Y_l^{+\prime} = \bar{G}_l^+ D_{Y_l}^+$, and finally $\bar{V}_l^+ = \bar{X}_l^+ M_l^+ \bar{X}_l^{+\prime}$. $\hat{R}_s^-(a_s)$ is defined analogously as $\hat{R}_l^+(a_l)$, with each superscript + + replaced by $-$. Therefore,

$$\hat{D}(z, a_l, P_l) = \sum_{j=1}^{m} \left[\frac{\|a_{lj}\|_{\bar{B}_l}^2}{\|a_{lj}\|_{\bar{V}_l}^2} - \frac{\|a_{lj}\|_{\bar{B}_l}^2}{\|a_{lj}\|_{\bar{V}_l^+}^2} + \frac{\|a_{sj}\|_{\bar{B}_s}^2}{\|a_{sj}\|_{\bar{V}_s}^2} - \frac{\|a_{sj}\|_{\bar{B}_s}^2}{\|a_{sj}\|_{\bar{V}_s^-}^2} \right]$$

The mapping F is not modified. It is defined as in the non centered case (see 2.3.2). The mapping $g: IP_k \to IR^{mpt}$ is such that $g(P)=a=(a_1, ..., a_k)$, where $a_l=(a_{l1}, ..., a_{lm})$ represents the m eigenvectors of the largest eigenvalues of $V_{l11}^{-1} V_{l12} V_{l22}^{-1} V_{l21}= =V_l^{-1} B_l$.

REMARK. $V_l^{-1} B_l$ is generally not symmetric. To obtain again the symmetric case we set: $V_l^{-1}=C_l C_l'$ and $a_{lj}=C_l \alpha_{lj}$. Thus $V_l^{-1} B_l a_{lj}=\lambda_{lj} a_{lj}$, which yields $C_l C_l' B_l C_l \alpha_{lj}=\lambda_{lj} C_l \alpha_{lj}$. The α_{lj}'s which satisfy this last equality are also the eigenvectors of $C_l' B_l C_l$, which is symmetric.

In order to orthonormalize a_{lj}, we just have to set $\hat{\alpha}_{lj}=\dfrac{\alpha_{lj}}{\|\alpha_{lj}\|}$ and $a_{lj}=C_l' \hat{\alpha}_{lj}$.
Thus $\|a_{lj}\|_{V_l}^2=\langle a_{lj}, a_{lj}\rangle_{V_l}=\langle \hat{\alpha}_{lj}, V_l C_l C_l' \hat{\alpha}_{lj}\rangle$, i.e.

$$\|a_{lj}\|_{V_l}^2=\langle \hat{\alpha}_{lj}, V_l C_l C_l' \hat{\alpha}_{lj}\rangle=\left\langle \frac{\alpha_{lj}}{\|\alpha_{lj}\|}, \frac{\alpha_{lj}}{\|\alpha_{lj}\|} \right\rangle=1$$

On the other hand, $\langle a_{lj}, a_{li}\rangle=\dfrac{1}{\|\alpha_{lj}\| \|\alpha_{li}\|}\langle \alpha_{lj}, \alpha_{li}\rangle=0$, since the α_{lj}'s (for $j=1, ..., p$) are the eigenvectors of a symmetric matrix.

4.4. Choice of the initial partition

We can use, for example, the following technique: Construct an array of dimension $q \times q$, where q is the number of modalities of the variable Y to predict. Each modality i of Y induces a cluster Q_i of objects. In the entry (i, j) of this array, we put the number of objects of the cluster Q_j which belong to the l closest neighbors of the objects (taken one by one) of the cluster Q_i. We set to 0 all the clusters which contain less objects than a given threshold. Thus, initial partition will be the set of the connected parts of the graph induced by the matrix defined above.

4.5. Assistance in the decision making processes

Having obtained say, 2 discriminant factors for each cluster of the partition we are trying to determine, we have at our disposal a good local discriminant factorial representation. When a new object comes in, we compute its (locally centered) distance to each one of the factor planes thus defined and we assign it to the closest. In this plane, we can display it visually as an additional element. We can also define a simple decision function, using the plane to which it is assigned. To do it, we compute the distance of the projection of the new object to the centers of gravity of the clusters, represented in the chosen plane by means of the coordinates on the correspondant local axes, and we assign it to the cluster of the closest center.

5. Canonical typological analysis in the case of more than two groups of variables

5.1. Notations

The data are given in the array as shown in Fig. 3, where k denotes the number of clusters of the partition to determine, p is the number of groups of variables, n is the number of canonical components to determine in each cluster and X_i^j is the $m_i \times n_j$ matrix associated with the i-th group of variables and with the j-th cluster of the partition P.

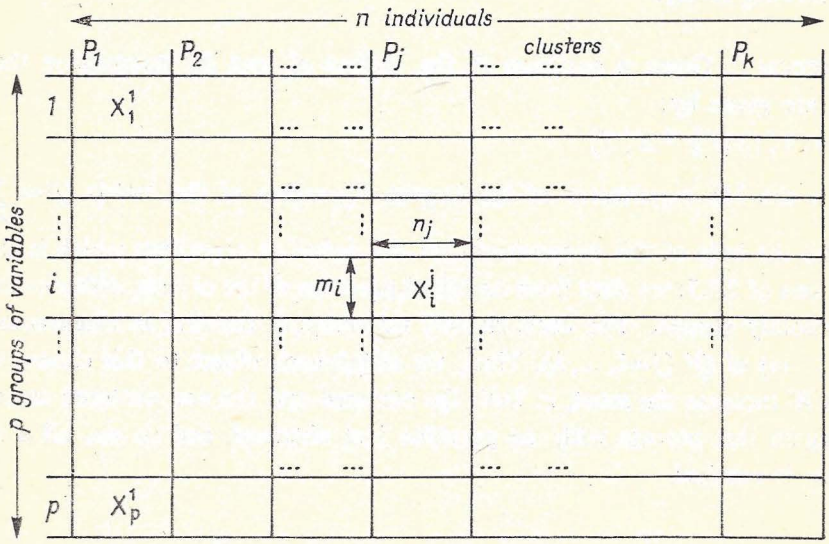

Fig. 3

Let W_i^j be the vector space linearly spanned by the variables of the i-th group. Denote by A_i^j the linear projection operator into W_i^j. In other words, we have:

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

5.2. The problem

The idea consists in looking for a partition $P=(P_1, ..., P_k)$, a vector Z_j in $IR^{\text{card } I_j}$ for each cluster P_j and p linear combinations of the variables ξ_j^i $(i=1, ..., p, j=1, ..., k)$ of each group such that the sum of the squares of the correlations between these combinations and Z_j is maximum for all the classes.

In the general case, where we are looking for m canonical components, we denote by ξ_{ij}^l the l-th canonical component associated with the linear subspace W_i^j.

It follows that: $\xi_{il}^j=X_i^j\,a_{il}^j$, where a_{il}^j is the l-th canonical factor associated with W_i^j.

Therefore, the mathematical formulation of the problem is:

$$\text{Maximize } W(P, \xi, Z) = \sum_{j=1}^{k} \sum_{i=1}^{p} \sum_{l=1}^{m} \frac{\langle \xi_{il}^j, Z_l^j \rangle_{M_j}^2}{\|\xi_{il}^j\|_{M_j}^2 \|Z_l^j\|_{M_j}^2},$$

where M_j is the matrix of weights associated with the objects.

5.3. The algorithm

In the centered as well as in the non centered cases, the algorithm is based on the following result:

PROPOSITION 3. Given a partition P, the vectors a_{il}^j and Z_l^j maximizing the criterion are given by:

a) $a_{il}^j = (X_i^j (X_i^j)^1)^{-1} X_i^j Z_l^j$

b) Z_l^j is the l-th eigenvector of the largest eigenvalue of the matrix $Q^j = \sum_{i=1}^{p} A_i^j$.

With the help of this proposition, we can define an algorithm which is similar to the one of 2.3.3: we start from an initial partition of the objects, either estimated or randomly chosen. For each cluster, we compute the first m eigenvectors Z_l^j ($l=1, ..., m$) of Q^j ($j=1, ..., k$). Then, we assign each object to this cluster which makes W increase the most, in both the centered and the non centered cases. We start again this process with the partition just obtained, and so on, till a stable position is reached.

Conclusion

For large data arrays, where we are interested in finding predicting linear combinations, it is legitimate to assume that these linear combination may not work everywhere because of the possibility of non homogeneous data. Adopting this approach, we have presented several algorithms, directly applicable to computer programming, in both the cases when the data are centered or not. One of the main interests of these methods is to allow a better visual display of the data, with the help of the local factors obtained.

As far as further research problems in this direction are concerned, it would be interesting to generalize to the case of multiple arrays (generated by qualitative variables, for example) what has already been done in factor analysis of correspondances and in discriminant analysis. It would also be fruitful to study more precisely and to use the relation between the dimensions of the subspaces generated by the W_i's and the number of clusters and canonical components to determine.

Final remark

Automatic classification is often considered to be a mixture of algorithms whose main virtue is to yield classes! In fact, at closer examination it turns out that it is quite rare to find algorithms of classification which address the same problems and optimise exactly the same criteria. That is why in this paper the algorithmic aspect was somewhat left in a shadow in order to make better appear a certain number of problems currently met in classification and optimisation criteria corresponding to them. So as to well distinguish the two domains (problems and algorithms) suffices to see how in classical statistical analysis classification problems are arrived at, with no algorithms being given for their solution.

Automatic classification is a young scientific discipline with multiple perspectives, theoretical as well as practical, and it makes well evident the three principles formulated by J.P. Benzécri (1973):

1. Statistics is not probability theory,
2. A model should follow the data, and not the other way round,
3. Consider the impact of computer technology on statistics.

References

[1] Benzecri J. P. L'analyse des données. Paris, Dunod, 1973.
[2] Caillez F., Pages J. P. Introduction à l'analyse des données. Paris, Seuil, 1976.
[3] Lebert L., Fenelon J. P. Statistique et informatique appliquées. Paris, Dunod, 1975.

Analiza kanoniczna z punktu widzenia klasyfikacji automatycznej

Jednym z głównych celów analizy danych jest wykrycie relacji zachodzących między zmiennymi opisującymi daną populację. Problem ten jest dyskutowany począwszy od analizy kanonicznej, a następnie dla analizy czynnikowej i analizy dyskryminacyjnej. Zaproponowano algorytm rozwiązujący ten problem zarówno w przypadku centrowania jak i danych niecentrowanych i udowodniono jego zbieżność. Rozważono aspekty numeryczne algorytmu.

Канонический анализ с точки зрения автоматической классификации

Одной из основных целей анализа данных является проявление соотношении между переменными описывающими данную совокупность. Здесь осуждается эту задачу сперва в рамках канонического анализа, а затем в рамках факторного и дискриминантного анализа. Предложено алгоритм решения этой задачи так в случае центрированных, как и нецентрированных данных. Доказана сходимость и обсуждены численные аспекты работы алгоритма.