Control and Cybernetics

vol. **24** (1995) No. 4

Simultaneous clustering of rows and columns

by

Gérard Govaert

Université de technologie de Compiègne,
URA CNRS 817,
BP 649 F-60206 Compiègne France

Like factor analysis methods but different to most clustering methods, simultaneous clustering aims to deal simultaneously with two sets which are related together in a data table.

In this paper we present the basic principle of a new approach. It includes different algorithms which are suited to different kinds of data tables (contingency tables, binary tables, quantitative data, questionnaires, ...). Then we examine closer some of these algorithms with the help of some examples.

Key words: clustering, simultaneous clustering, contingency table, binary table continuous data, questionnaire,

1. Introduction

Although data analysis methods are many and various, there underlies nearly all of them a desire to summarize, to simplify and, eventually to explain the data.

Factor analysis methods as principal components analysis (Anderson 1958), correspondence analysis (Benzecri 1973, Greenacre 1984) or multiple correspondence analysis (Lebart, Morineau and Warwick 1984) aim to simplify a rectangular data table defined on two sets I and J. One of the distinctive feature of these methods is that they obtain results simultaneously on the two sets.

When analyzing data by means of clustering procedures, it seems relevant to have the same approach. But most of the existing clustering methods are concerned only with one of the two sets. The object of this work is to present a general methodology to define methods which aim to provide simultaneously a partition of both sets.

This approach is not a new one. Fisher (1969) sets the problem of finding two simultaneous partitions by means of matrix computation; he defines a criterion to optimize, without proposing any method for solving this problem. Among the existing clustering problems Anderberg (1973) cites the selection of the

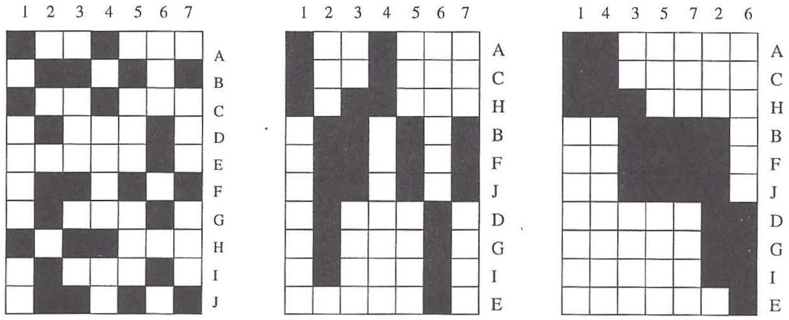

Figure 1. Reorganization by permutation of rows and columns

set to classify. He considers the clustering of the variables as sensible as the clustering of the objects. This leads him to propose an iterative approach where objects and variables are alternatively classified until both partition become "mutually harmonious". Toledano and Brousse (1977) and Greenacre (1988) set a similar problem for hierarchical clustering. Bock (1979) shows the interest of simultaneous clustering, gives several examples where these particular approach of clustering provides good solutions and proposes some clustering methods based on statistical models.

Simultaneous clustering can be related to direct clustering (Hartigan, 1972 and 1975), also called block modeling or block clustering. Direct clustering organizes directly the initial table and simplifies the table by permuting rows and columns (see figure 1).

A wide variety of procedures have been proposed for finding patterns in data matrices. These procedures differ in the pattern they seek, the types of data to which they apply, and the assumption on which they rely. Among these procedures, we can cite the bond energy algorithm (McCormick, Schweitzer, and White 1972, Arabie and Hubert 1990), the GPM algorithm of Garcia and Proth (1986), the block seriation approach of Marcotorchino (1987) and the permutation-based algorithm of Duffy and Quiroz (1991). We can also cite the interesting work of Bertin (1977) which have designed a device to organize manually the initial table by permuting rows and columns.

In this paper, we are only interested in the seek of partitions simultaneously on two sets which are related together in a data table. In the terminology of Tucker (1964), our work can be characterized as seeking partition both modes of a two-mode matrix.

Taking place in the context of exploratory analysis, we have made it a rule to respect the following conditions :

- to propose a simple summary easily understood for the user;
- to be able to use jointly factorial methods and clustering methods;
- to be able to have at one's disposal close methods for main data tables;
- to obtain classical clustering methods when we only try to cluster one the

two sets.

To achieve this aim, we propose a methodology which generalizes the two algorithms *Crobin* and *Croki2* that we have developed for the contingency tables (Govaert 1977) and the binary tables (Govaert 1984).

After introducing the general principle of this methodology in Section 2, we will illustrate in Sections 3 and 4 this process by describing the algorithms *Crobin* and *Croki2*. Moreover, for contingency tables, we will show the very close links of our approach with correspondence analysis. In Section 5, we will examine how our approach can be applied to quantitative data and we will show the links which exist with principal component analysis. Section 6 is devoted to the presentation of an illustrative example. Finally, we will finish by a concluding section.

2. Basic principle

2.1. Data

Our approach is concerned with data that consists of a matrix $X(I, J)$ defined on two sets I and J (two-mode two-way matrix in the Fisher terminology).

2.2. Summary matrix

The main idea of our approach is to try to resume the initial matrix $X(I, J)$ by a matrix $X(P, Q)$, much smaller, simply defined from a couple of partitions P and Q of I and J, and having the same structure than the initial table. For instance, we will resume a 1000x200 binary matrix of by a 10x5 binary matrix or a 200x100 contingency table by a 8x5 contingency table.

The justification of this process is to allow a simple use of results since they have going to be presented in the same form than the initial data. Besides, the same structure of the two matrices $X(I, J)$ and $X(P, Q)$ will allow us to define the objective function, which has to be optimized, more easily.

2.3. Objective function

Depending on the types of data, two approaches can be used to define an objective function :

- It exists an information measure I such that every summary means loss of information :
  $$I[X(I, J)] \leq I[X(P, Q)].$$
  Then, the clustering problem consists in minimizing the loss of information when passing from the initial table to the summary table.
- We use a function Δ which measures the difference between the two matrices $X(I, J)$ and $X(P, Q)$. Then, the purpose is to find the matrix $X(P, Q)$ which minimizes $\Delta(X(I, J), X(P, Q))$.

After this formulation, it remains to find the pair of partitions which optimize the objective function.

2.4. Algorithm

Several algorithms are able to seek a local optimum of the objective function. For solving this problem we propose an algorithm which defines a sequence (P^n, Q^n) of partition pairs. Starting from an initial partition pair (P^0, Q^0) the following procedure is applied: one of these partitions is fixed and a better partition of the other set is searched for. Then the resulting partition is fixed and a better partition of the first set is searched for. These two steps are repeated until convergence. Thus an iterative algorithm is obtained. For finding the better partition at each step, we use dynamic cluster analysis (Diday et al. 1980, Celeux et al. 1989). This method is also iterative. Hence the simultaneous clustering algorithms have two levels of iterations.

The properties of this type of algorithms are simplicity, speed of convergence and the possibility to process big data tables. The drawbacks are due to the fact that they only provide local optimum.

2.5. Remarks

In this work, we have not put the stress on the algorithm but rather on the form of the searched result (the summary table) and on the criterion to be optimized. The problem of the research of an optimal algorithm still exist but for the proposed criteria which are extension of criteria of k-means types, the problem will have to be solved for the k-means before considering optimal solutions to our problem. The solution could be, for instance, stochastic variants of the EM algorithm as SEM (Celeux, Diebolt 1985), or simulated annealing variants as CAEM and SAEM (Celeux, Govaert 1992). Of course, an algorithm providing the global optimal solution would be an great improvement of this approach.

As with all methods converging towards a local optimum, the results obtained depend on the initial partitions. The algorithm is consequently applied several times, starting with random initial partition. Then many possibilities may be considered. For example, it is possible to find a strong agreement in the solutions (Celeux et al. 1989) indicating groups of elements which are stable for all random starts. Here, we have chosen to keep to the initial goal. This means optimizing the objective function. Thus the program keeps the best pair of partitions after having performed several initial random drawings.

Another problem that we have not considered in this work is the choice of the number of clusters that must be fixed by the user. Let us notice that the problem is not yet solved in a satisfactory way. We can quote the recent works of Celeux and Soromenho (1995) done in the Gaussian mixture approach of this problem.

3. Binary data

To illustrate our approach, we study in this section how it can be applied to a binary table.

3.1. The problem

Before describing the problem, let us make our purpose precise with a simple example. Let us consider data in figure 2. The rows correspond to a set of ten micro-computers and the rows to ten properties that these computer may (value 1) or may not have (value 0).

	1	2	3	4	5	6	7	8	9	10
a	1	0	1	0	1	0	0	1	0	1
b	0	1	0	1	0	1	1	0	1	0
c	1	0	0	0	0	0	0	1	1	0
d	1	0	1	0	0	0	0	1	0	0
e	0	1	0	1	0	1	1	0	1	0
f	0	1	0	0	0	1	1	0	1	0
g	0	1	0	0	0	0	0	1	0	1
h	1	0	1	0	1	1	0	1	1	1
i	1	0	0	1	0	0	0	0	0	1
j	0	1	0	1	0	0	1	0	0	0

Figure 2. Example of binary data.

In order to justify and detail the search for simultaneous partitions of a binary table, let us cite some ideas proposed by Lerman about the notion of polythetic cluster (Lerman 1981). He first recalls the notion of polythetic cluster: "A polythetic cluster G of a natural clustering refers to a subset B of attributes in such a way that:

1. each element of the cluster has an important proportion of attributes from B;
2. each attribute of B is present in an important proportion;
3. an attribute is not necessarily shared by all elements of G."

Lerman generalizes this notion: "In the more general situation of a good clustering on E with a good clustering on A , each cluster E_i of the partition (E_1, ..., E_l) corresponds to the union B of clusters of the partition (A_1, ..., A_k). Conversely, each cluster A_j corresponds to the union G of clusters of the partition (E_1, ..., E_l)."

This situation may be represented by a binary table reorganized according to the partitions (figure 3).

Let assume that the shaded regions correspond to regions with high one density and that the unshaded regions correspond to regions with high zero density.

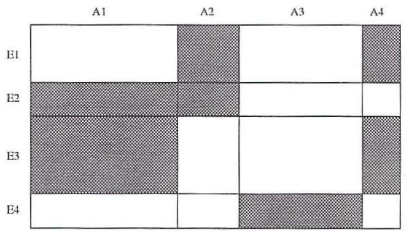

Figure 3. Binary table reorganized according to row and column partitions

3.2. Summary matrix

With the previous example, if the partitions P and Q are respectively $\{\{a, d, h\}, \{b, e, f, j\}, \{c, g, i\}\}$ and $\{1, 3, 5, 8, 10\}, \{2, 4, 6, 7, 9\}\}$, we get table of the figure 4 by reorganizing rows and columns according these two partitions and the initial binary matrix can be summarized by the binary matrix of figure 5 when crossing the two partitions.

		1					2				
		1	3	5	8	10	2	4	6	7	9
A	a	1	1	1	1	1	0	0	0	0	0
	d	1	1	0	1	0	0	0	0	0	0
	h	1	1	1	1	1	0	0	1	0	1
B	b	0	0	0	0	0	1	1	1	1	1
	e	0	0	0	0	0	1	1	1	1	1
	f	0	0	0	0	0	1	0	1	1	1
	j	0	0	0	0	0	1	1	0	1	0
C	c	1	0	0	1	0	0	0	0	0	1
	g	0	0	0	1	1	1	0	0	0	0
	i	1	0	0	0	1	0	1	0	0	0

Figure 4. Representation of the simultaneous clustering.

	1	2
A	1	0
B	0	1
C	0	0

Figure 5. Summary of the data.

A, B and C correspond to the three clusters of computers, 1 and 2 to the two clusters of properties. Each table division is associated with 0 or 1 according a majority vote. This table allows to see that computers of cluster A usually have properties 1 but not 2, computers of cluster B have properties 2 but not 1 and that the one of cluster C have no properties. Thus a table with six values summarizes a table of a hundred values.

The idea of summary table appears in this example. This table has the same structure as the initial table. This time, the summary table is a binary table (with each pair of clusters an ideal value is associated) and the criterion will measure the deviation between this ideal table and the initial data. It is easy to see, that this transformation comes to obtain homogeneous blocks of 0 or 1 by reorganizing rows and columns of the initial table.

3.3. The objective function

We aim to find homogeneous blocks, which are full of zeros or full of ones. Each pair of clusters (k, m) is associated with an ideal binary value (1 or 0). Thus we get a binary table which we will call "kernel". We want to minimize the number of times where the value associated with a pair (i, j) is different from the ideal value associated to the clusters pair to which (i, j) belongs. This quantity represents the difference between the initial table and the ideal table.

If we write :

- (x_i^j) the initial binary table defined on the two sets I and J with sizes n and p,
- P and Q the partitions into K clusters and M clusters of the two sets I and J,
- $\mathbf{L}_{K,M}$ the set of binary table with K rows and M columns; $\mathbf{L}_{K,M}$ represents the kernel set:
  $\lambda \in \mathbf{L}_{K,M} \iff \lambda = (a_k^m)$ where $a_k^m \in \{0,1\}$ $\forall k = 1, \dots, K$ and $\forall m = 1, \dots, M$,

the objective function we would like to minimize can be defined by

$$W(P, Q, \lambda) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q_m} |x_i^j - a_k^m|.$$

When we only deal with the research of one partition, we obtain the well-known maximal predictive classification criterion proposed by Gower (1974). Let us mention that we also find this criterion when we use a classification maximum likelihood criterion (Celeux and Govaert 1991) or an entropy criterion (Gyllenberg, Koski and Verlaan 1994) on the Bernoulli mixture model.

If we constrain the summary to have a diagonal structure (same number of clusters for the two partitions, 1 on the diagonal and 0 everywhere else), we find the criterion proposed by Garcia and Proth (1986). Let us notice that introducing such a constraint in our algorithm is very easy.

3.4. Algorithm

As we have seen in the Section 2, it remains to define more precisely the two steps of the algorithm. If P and Q is a pair of partitions and λ a kernel, for Q

fixed, we try to improve partition P and kernel λ. Thus, we seek for a partition P' and a kernel λ' such that:

$$W(P,Q,\lambda) \geq W(P',Q,\lambda').$$

We may write

$$W(P,Q,\lambda) = \sum_k \sum_{i \in P_k} \sum_m \sum_{j \in Q_m} |x_i^j - a_k^m|).$$

If we note $y_i^m = \sum_{j \in Q_m} x_i^j$ and $q_m = \# Q_m$, we have

$$A = \begin{cases} y_i^m & \text{if} \quad a_k^m = 0 \\ q_m - y_i^m & \text{if} \quad a_k^m = 1 \end{cases}$$

which can be rewritten

$$A = |y_i^m - q_m a_k^m|$$

and we obtain

$$W(P,Q,\lambda) = \sum_k \sum_{i \in P_k} \sum_m |y_i^m - q_m a_k^m|.$$

It is easily shown that the dynamic cluster algorithm provides a solution to our problem. The algorithm must be defined on a set of n elements and M variables associates with table $(y_i^m)_{i \in I, m=1,M}$. This table is supplied with distance L^1 with kernel of the form $(q_1 a_k^1, ..., q_M a_k^M)$ where $a_k^m \in \{0,1\}$ (each component of the kernel can be exclusively the minimum or the maximum reached by regrouping the columns of the initial table). We may obviously start with a fixed partition P and improve the partition Q and the kernel λ.

4. Contingency table

4.1. Introduction

In this section, we first describe the *Croki2* algorithm which performs simultaneous clustering on contingency tables defined with two sets I and J, or more generally on tables which have the same properties. Then, we will show the very close links between this method and correspondence analysis.

4.2. Notations

$X(I,J) = (n_{ij})$ will denote the initial contingency table defined on the two sets I and J of sizes n and p. The usual terminology is used:

- s is the sum of the table elements $(\sum_{i \in I} \sum_{j \in J} n_{ij})$,
- F is the frequency table $(f_{ij} = \frac{n_{ij}}{s}, i \in I, j \in J)$, ($f_{ij}$ is an estimation of probability that an object has simultaneously the category i and the category j),

- $f_{i.}$ et $f_{.j}$ are the marginal frequencies : $\forall i \in I, f_{i.} = \sum_{j \in J} f_{ij}$ and $\forall j \in J, f_{.j} = \sum_{i \in I} f_{ij}$
- $f_I = (f_{1.}, ...f_{i.}, \dots, f_{n\cdot})$ and $f_J = (f_{.1}, \dots, f_{.j}, \dots, f_{.p})$ are the marginal laws defined on I and J,
- $f_j^i = \frac{f_{ij}}{f_{i.}}$ and $f_i^j = \frac{f_{ij}}{f_{.j}}$ are the conditional frequencies,
- $f_J^i = (f_1^i, \dots, f_j^i, \dots, f_p^i)$ and $f_I^j = (f_1^j, \dots, f_i^j, \dots, f_n^j)$ are the conditional laws, also named row and column profiles.
- $D_{\frac{1}{f_I}}$ is the diagonal matrix with diagonal terms $\frac{1}{f_1^i}, \dots, \frac{1}{f_j^i}, \dots, \frac{1}{f_p^i}$,
- $D_{\frac{1}{f_J}}$ is the diagonal matrix with diagonal terms $\frac{1}{f_1^j}, \dots, \frac{1}{f_i^j}, \dots, \frac{1}{f_n^j}$.

4.3. The summary table

Following the principle stated in the introduction, the summary table associated with the two partitions must also be a contingency table. It is obtained by regrouping the rows and columns according the partitions P and Q in the following manner: If $P = (P_1, ..., P_K)$ is a partition of I into K clusters and $Q = (Q_1, ..., Q_M)$ a partition of J into M clusters, it becomes possible to define a new contingency table by summing the elements of the initial contingency table corresponding to each pair of clusters (P_k, Q_m). This table, denoted $T(P, Q)$, is defined by:

$$T(k, m) = \sum_{i \in P_k} \sum_{j \in Q_m} n_{ij} \qquad \forall k = 1, \dots, K \text{ and } \forall m = 1, \dots, M.$$

4.4. The objective function

4.4.1. Definition

The chosen information measure we would like to preserve is the χ^2 of contingency (or Pearson chi-square statistic) which measures the dependence between I and J by the contingency χ^2:

$$\chi^2(I, J) = s \sum_{i \in I} \sum_{j \in J} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}.$$

This measure usually provides statistical evidence of a significant association, or dependence between rows columns of the table. This quantity represents the deviation between the theoretical frequencies $f_{i.}f_{.j}$, that we would have if I and J were independent, and the observed frequencies f_{ij}. If I and J are independent the χ^2 will be zero and if there is a strong relationship between I and J, the χ^2 will be high. So, a significant chi-square indicates a departure from row or column homogeneity and can be used as a measure of heterogeneity. The chi-square can be used as a measure of the information brought by a contingency table.

As for the initial table, it is possible to measure the information brought by the new contingency table $X(P,Q)$. This is measured by the associated contingency χ^2 noted $\chi^2(P,Q)$. We chose this quantity to measure the quality of the two partitions P and Q of sets I and J.

The following relationship is easily demonstrated:

$$\chi^2(I,J) > \chi^2(P,Q). \tag{1}$$

Regrouping the elements of each cluster leads to a loss in χ^2. Minimizing this loss is equivalent to searching for the partitions P and Q which maximize the contingency χ^2 of their associated table. This leads to maximizing the dependence between partition P and partition Q. The problem we like to solve is the simultaneous search of two partitions P and Q which maximize the contingency χ^2 of the associated table.

In order to show the pertinence of this problem, let us consider an example presented by Benzecri (1973). In this example a contingency table is defined on a set I of towns and a set J of professions. t_{ij} represents the number of people practicing profession j in town i:

"A town classification in terms of a partition P will be good if knowledge of a town class give us more information about the repartition of the professions in this town. In the same way, a classification of the professions in terms of partition Q will be good if knowledge of a profession class gives us more information about the repartition of this profession in the towns."

Notice that in the ideal case where the row profiles and the column profiles are equal inside each cluster of P and Q, there is no loss of information.

4.4.2. Justification

The Pearson chi-square usually provides statistical evidence of a significant association, or dependence, between rows and columns of the contingency table. Various methods (Goodman 1985) have been proposed for investigating this association. Some of them are graphical approaches and the best known is Correspondence analysis (Benzecri 1973, Greenacre 1984, Lebart, Morineau and Warwick 1984). This technique represents the rows and the columns of the table in high-dimensional space and then projects them onto a best-fitting subspace or lower dimensionality for ease of interpretation. Here, we propose to analyze the χ^2 by means of clustering.

Remarks that this function is very closed to the Goodman RxC association function

$$\sum_{i,j} f_{ij} \ln(\frac{f_{ij}}{f_{i.}f_{.j}})$$

and gives similar results.

Among the numerous properties of this approach, we can cite the "principle of distributional equivalence" (Benzecri 1973) : If two row profiles are identical then the corresponding two rows of the original data matrix may be replaced by their summation (a single row) without affecting the geometry of the.columns profiles.

4.5. Search of a partition optimizing the χ^2

In order to define more precisely our algorithm, it is necessary to recall that a variant of the dynamic cluster algorithm is able to optimize the χ^2 criterion:

From a contingency table defined on I and J, the k-means algorithm applied to the set $N(I)$ of profiles f_J^i with weights $f_{i.}$ and considering the χ^2 metric which is the Euclidean metric defined by the matrix $D_{\frac{1}{f_J}}$, allows to partition I into K clusters optimizing the contingency χ^2.

Indeed, the optimized criterion may be written:

$$W(P) = \sum_{k=1,K} \sum_{i \in P_k} f_{i.} d^2(i, G(P_k))$$

where $P = (P_1, ..., P_K)$, d is the $D_{\frac{1}{f_J}}$ distance and $G(P_k)$ is the center of gravity of P_k. The following relationship is easily demonstrated:

$$\chi^2(I, J) = sW(P) + \chi^2(P, J). \tag{2}$$

$sW(P)$ represents the information lost in regrouping the elements according the partition P, and $\chi^2(P, J)$ corresponds to the preserved information. As the quantity $\chi^2(I, J)$ does not depend on the partition P, the search of the partition minimizing the criterion $W(P)$ is consequently equivalent to the search for the partition P maximizing $\chi^2(P, J)$. The dynamic cluster method maximizes the contingency χ^2 of the table (P, J). Notice that relation 1 from the preceding paragraph is easily demonstrated when starting with relation 2.

4.6. The *Croki2* algorithm

A sequence (P^n, Q^n) is computed from any initial pair (P^0, Q^0) so that the associated series of χ^2 values is increasing.

4.6.1. Computation of P^{n+1} from (P^n, Q^n)

Let $X(I, Q^n)$ be a contingency table defined by

$$X(i, k) = \sum_{j \in Q_k^n} n_{ij}$$

where

$$Q^n = (Q_1^n, \ldots, Q_M^n).$$

The partition P^{n+1} is obtained by applying the preceding algorithm to the table $X(I, Q^n)$ while considering that the objects to classify are the elements of I, the variables are the clusters of Q^n and taking P^n as initial partition.

4.6.2. Computation of Q^{n+1} from (P^{n+1}, Q^n)

The principle is the same. But the table taken into consideration is $X(P^{n+1}, J)$ defined by:

$$X(k, j) = \sum_{i \in P_k^{n+1}} n_{ij}.$$

The objects to classify are the elements of J and the variables are the K clusters of P^{n+1}. The algorithm starts from the partition Q^n to get the partition Q^{n+1}.

From the convergence properties of the dynamic cluster method, it is possible to show equivalent convergence properties for the *Croki2* algorithm.

4.6.3. Relationship with correspondence analysis

We have just seen that the *Croki2* algorithm is applied to the same kind of data and uses the same measure of information (the contingency χ^2) as in correspondence analysis. It is possible to go further in this comparison and show that the two problems are close: simultaneous clustering may be considered as a constrained correspondence analysis.

5. Continuous data

5.1. Introduction

In this section, we suppose that we have observations of p continuous variables on each of n individuals. I will be the set of individuals and J the set of variables The data can be arranged in a $n \times p$ matrix $X = (x_i^j, i \in I, j \in J)$. Let us notice that, in this case, data matrix present a dissymmetry: rows and columns, which correspond to entities of different types, will not be handled in the same way, contrary to what we have done in the two precedent examples.

To analyze this type of data, we can use principal component analysis (PCA) which summarizes data by means of new axes. Here, we propose to resume the data by means of clusters of the individuals and of the variables.

As for the PCA, we associate to the set of individuals the weights $\mu(I) = (\mu_1, \ldots, \mu_n)$ which verify

$$\forall i \in I \quad \mu_i > 0 \text{ and } \sum_{i \in I} \mu_i = 1$$

and we associate to the set of variables the weights $\nu(J) = (\nu_1, \ldots, \nu_p)$ which verify

$$\forall j \in I \quad \nu_j > 0.$$

For instance, we can use

$$\forall i \in I \quad \mu_i = \frac{1}{n} \quad \text{and} \quad \forall j \in J \quad \nu_j = 1.$$

Moreover, we suppose than X is "column-centered", that is the means of the columns of X are equal to 0 :

$$\forall j \in J \qquad \sum_{i \in I} \mu_i x_i^j = 0.$$

Remarks that if it is not true, it is easy to modify the initial table to obtain this property.

Finally, the initial data, which are are defined by a "column-centered" matrix $(X(I,J)$ and two vectors $\mu(I)$ and $\nu(J)$, will be noted $(X(I,J), \mu(I), \nu(J))$.

5.2. Geometrical representation

We can associate to the data two geometrical representations :

- a geometrical representation of the individuals by a set of n points of $I\!R^p$: the coordinate of the n points are the rows of X, the μ_i are the weights of the point and the ν_j can be used to define an Euclidean metric :
$$d^2(i,i') = \sum_{j=1}^{p} \nu_j |x_i^j - x_{i'}^j|^2.$$
- a geometrical representation of the variables by a set of p points of $I\!R^n$: the coordinate of the n points are the columns of X, the ν_j are the weights of the point and the p_i can be used to define an Euclidean metric :
$$d^2(j,j') = \sum_{i=1}^{n} \mu_i |x_i^j - x_i^{j'}|^2.$$

5.3. The summary

Following the principles stated in the introduction, we must associate with the two partitions a summary table $X(P,Q)$ and two sets of weights $\mu(P)$ and $\nu(Q)$ which have the same structure that the initial data $(X(I,J), \mu(I), \nu(J))$. For this, if $P = (P_1, ..., P_K)$ is a partition of I into K clusters and $Q = (Q_1, ..., Q_M)$ a partition of J into M clusters, we define :

$$\forall k = [1, \ldots, K], \forall m = [1, \ldots, M], \qquad x_k^m(P,Q) = \frac{\sum_{i \in P_k} \sum_{j \in Q_m} \mu_i \nu_j x_i^j}{\sum_{i \in P_k} \sum_{j \in Q_m} \mu_i \nu_j},$$

$$\forall k = [1, \ldots, K], \qquad \mu_k = \sum_{i \in P_k} \mu_i$$

and

$$\forall m = [1, \ldots, M], \qquad \nu_m = \sum_{i \in Q_m} \nu_i.$$

This new structure will be noted $(X(P,Q), \mu(P), \nu(Q))$. It is easy to verify that this structure has the same property than $(X(I,J), \mu(I), \nu(J))$:

the means of the columns of $X(P,Q)$ are equal to 0,

$$\forall m = [1,\dots,M] \quad \sum_{k=1}^{K} \mu_k x_k^m(P,Q) = 0,$$

$$\forall k = [1,\dots,K] \quad \mu_k > 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1,$$

$$\forall m \in [1,\dots,M] \quad \nu_m > 0.$$

In the following, we note $(X(I,Q),\mu(I),\nu(Q))$ the structure obtained when, in the partition P of I, each element of I is a cluster. Similarly, we note $(X(P,J),\mu(P),\nu(J))$ the structure obtained when, in the partition Q of J, each element of J is a cluster.

5.3.1. Inertia

The chosen information measure we would like to preserve is the following :

$$I(X(I,J),\mu(I),\nu(J)) = \sum_{i\in I}\sum_{j\in J} \mu_i \nu_j (x_i^j)^2.$$

With the geometrical representations, this information represents the inertia of the set I in $\mathbb{R}^p$ or the inertia of the set J in $\mathbb{R}^n$. Let us notice that this information measure is the measure used by PCA.

5.4. The objective function

5.4.1. Definition

As the summary $(X(P,Q),\mu(P),\nu(Q))$ has the same structure that initial data, we can define the information measure $I(X(P,Q),\mu(P),\nu(Q))$. Moreover, it is possible to prove that this information is less than the information associated to the initial data, that is to say that the grouping in two partitions P and Q leads to a loss of information. In our method, we propose to minimize this loss or, equivalently, to maximize the criterion $I(X(P,Q),\mu(P),\nu(Q))$.

5.4.2. Link with the Fisher criterion

We can show that in the simple case where every weights p_i et q_j are equal, the loss of information is exactly the Fisher criterion (1969). We can illustrate it on the following example. If data are

$$X = \begin{pmatrix} 1 & 2 & 8 \\ 2 & 1 & 7 \\ 3 & 4 & 7 \\ 4 & 3 & 6 \end{pmatrix},$$

with weights

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \frac{1}{4} \text{ and } \nu_1 = \nu_2 = \nu_3 = 1$$

and if partitions are

$$P = (\{1,2\}, \{3,4\}) \text{ and } Q = (\{1,2\}, \{3\}),$$

then, summary table is

$$X(P,Q) = \begin{pmatrix} 1.5 & 7.5 \\ 3.5 & 6.5 \end{pmatrix},$$

and with weights

$$\mu_1 = \mu_2 = \frac{1}{2} \text{ and } \nu_1 = 2, \nu_2 = 1.$$

The Fisher criterion can be expressed as trace$(X-Y)'(X-Y)$ where

$$Y = \begin{pmatrix} 1.5 & 1.5 & 3.5 & 3.5 \\ 1.5 & 1.5 & 3.5 & 3.5 \\ 7.5 & 7.5 & 6.5 & 6.5 \end{pmatrix}.$$

then, it is easy to show that

$$\text{trace}(X-Y)'(X-Y) = I(X(I,J), \mu(I), \nu(J)) - I(X(P,Q), \mu(P), \nu(Q)).$$

5.4.3. K-means

If we restrict to the sought of the partition P, our criteria can be expressed as the loss of information due to the partition. By using the Huygens theorem, we can show

$$I(X(I,J), \mu(I), \nu(J)) - I(X(P,J), \mu(P), \nu(J)) = W(P/J)$$

where

$$W(P/J) = \sum_{k=1}^{K} \sum_{i \in P_k} \mu_i \sum_{j=1}^{p} \nu_j (x_i^j - x_k^j)^2.$$

Thus, we find the intra-class inertia function minimized by the classical k-means algorithm and our criterion generalizes the k-means criterion to the simultaneous search for partitions.

5.5. Algorithm

Let us remind that, by applying the general principle, we have to compute for a fixed partition of J the best partition of I, and, for a fixed partition of J the best partition of I. For this, it is sufficient to notice that

$$I(X(I,Q),\mu(I),\nu(Q)) - I(X(P,Q),\mu(P),\nu(Q)) = W(P/Q)$$

where $W(P/Q)$ is the intra-class inertia criterium for the partition P of I when data are $(X(I,Q),\mu(I),\nu(Q))$ and

$$I(X(P,J),\mu(P),\nu(J)) - I(X(P,Q),\mu(P),\nu(Q)) = W(Q/P)$$

where $W(Q/P)$ is the intra-class inertia criterium for the partition Q of J when data are $(X(P,J),\mu(P),\nu(J))$.

Thus, using K-means algorithm alternatively on $(X(I,Q)$, $\mu(I)$, $\nu(Q))$ and $(X(P,J)$, $\mu(P)$, $\nu(J))$, we obtain an algorithm optimizing the criterion

$$I(X(P,Q),\mu(P),\nu(Q)).$$

6. Illustrative example

To illustrate the family of methods developed in this work, we have chosen to present some results obtained with the *Croki2* algorithm on a small contingency table.

6.1. The data

The data concerns the comparison between time-budgets (Jambu 1976). We have a table $X(I,J) = n_{ij}, i \in I, j \in J$ where n_{ij} represents the number of hours spent to practise the activity j by the population i during a certain time period. The set I is made of 28 types of population characterized by the sex, the country, the professional activity and the marriage. In a row identifier, the letter meaning is the following: h: man, f: woman, a: working, na: not working, m: married, c: unmarried, us or u: USA, we or w: west country, es or e: east country, yo or y: Yugoslavia. The set J is made of 10 activity clusters: prof: Occupational work, tran: Activity related to occupational work, mena : Home work, enfa: Activity related to the child work, cour: Shopping, toil: Washing and personal care, repa: Mealtime, somm: Sleep, tele: Television, lois: Other leisure. We have reported the data in figure 6.

6.2. Results

Here we present the best result obtained among ten initial drawings when the numbers of clusters of the row and column partitions are respectively 5 and 3. The initial χ^2 value was 9658 and the resultant χ^2 value is 8048. The percentage of χ^2 explained by the partition pair is very good with this small

	prof	tran	mena	enfa	cour	toil	repa	somm	tele	lois
haus	610	140	60	10	120	95	115	760	175	315
faus	475	90	250	30	140	120	100	775	115	305
fnau	10	0	495	110	170	110	130	785	160	430
hmus	615	141	65	10	115	90	115	765	180	305
fmus	179	29	421	87	161	112	119	776	143	373
hcus	585	115	50	0	150	105	100	760	150	385
fcus	482	94	196	18	141	130	96	775	132	336
hawe	652	100	95	7	57	85	150	807	115	330
fawe	510	70	307	30	80	95	142	815	87	262
fnaw	20	7	567	87	112	90	180	842	125	367
hmwe	655	97	97	10	52	85	152	807	122	320
fmwe	168	22	529	69	102	83	174	825	119	392
hcwe	642	105	72	0	62	77	140	812	100	387
fcwe	389	34	262	14	92	97	147	848	84	392
hayo	650	140	120	15	85	90	105	760	70	365
fayo	560	105	375	45	90	90	95	745	60	235
fnay	10	10	710	55	145	85	130	815	60	380
hmyo	650	145	112	15	85	90	105	760	80	357
fmyo	260	52	576	59	116	85	117	775	65	295
hcyo	615	125	95	0	115	90	85	760	40	475
fcyo	413	89	318	23	112	96	102	774	45	409
haes	650	142	122	22	76	94	100	764	96	334
faes	578	106	338	42	106	94	52	752	64	228
fnac	24	8	594	72	158	92	128	840	86	398
hmes	652	133	134	22	68	94	102	762	122	310
fmes	434	77	431	60	117	88	105	770	73	229
hces	627	148	68	0	88	92	86	770	58	463
fces	433	86	296	21	128	102	94	758	58	379

Figure 6. The data

example: more than 83% of the χ^2 has been preserved. The partition of the rows are the following: (fmus fmwe fmyo), (fayo faes fmes), (faus fcus fawe fcwe fcyo fces), (fnau fnaw fnay fnae), (haus hmus hcus hawe hmwe hcwe hayo hmyo hcyo haes hmes hces) and the partition of the columns are the following: (prof, tran), (mena, enfa), (cour, toil, repa, somm, tele, lois).

The contingency table $X(P, Q)$ obtained by regrouping rows and columns according to the two partitions P and Q (see figure 7) summarizes the two partitions. Let us recall that the program aims to maximize the χ^2 preserved in the summary table. A more interesting output is the array of figure 8, defined by the values $\frac{f_{km}}{f_{k.}f_{.m}}$. This table allows us to characterize the two partitions.

	1	2	3
1	1741	710	4832
2	1291	1860	3993
3	1765	3165	9363
4	2690	89	6818
5	1201	9134	18456

Figure 7. X(P,Q)

	1	2	3
1	1846	437	1024
2	1395	1168	863
3	953	993	1011
4	2165	41	1096
5	322	1423	989

Figure 8. $\frac{f_{km}}{f_{k.}f_{.m}}$

The most interesting values are the values far from the mean 1000. These values are brought to the fore by an underline. Notice that row cluster 3 and column cluster 3 are always around the average. Thus these clusters are not characteristic of anything. On the contrary, the other clusters are very characteristic.

6.3. Relationship with correspondence analysis

To illustrate the relationship between the correspondence analysis and simultaneous clustering, we have applied correspondence analysis to the preceding example and reported in the figures 9 and 10 the representation of I and J on the two first axes which explain 84% of the χ^2. We have also reported the clusters obtained by our simultaneous clustering method.

With this simple example, we obtain the same conclusions as with simultaneous clustering. For the rows, clusters 4 and 5 are strongly opposed, cluster 1

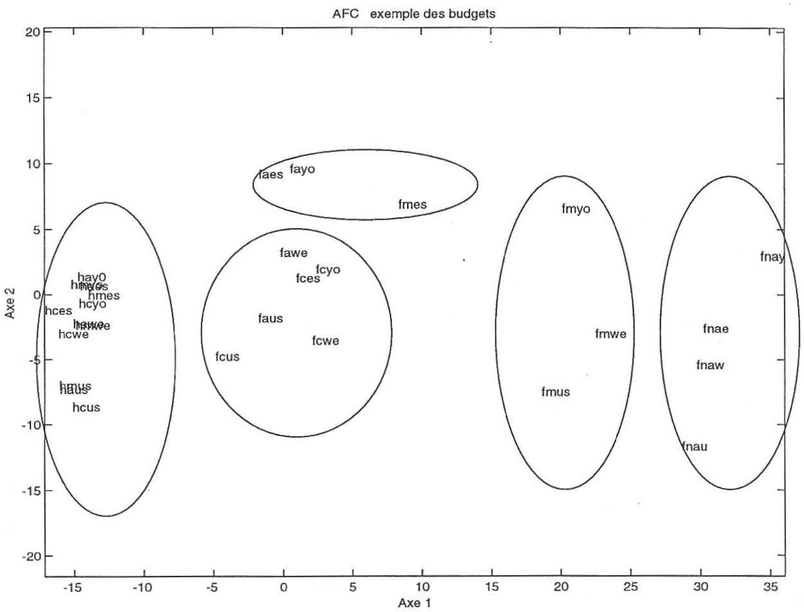

Figure 9. Representation of I on the two first axes

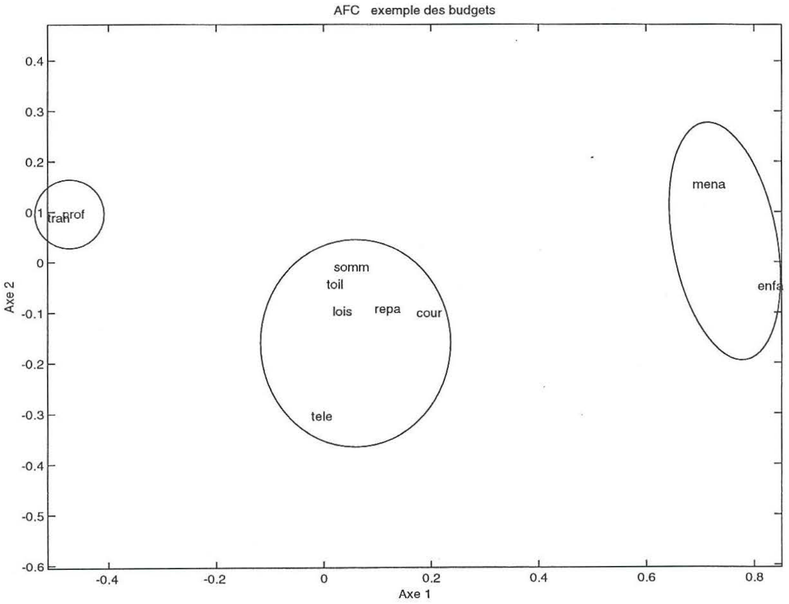

Figure 10. Representation of J on the two first axes

has the same tendency as cluster 4, cluster 3 is the·middle cluster and cluster 2 is apart from the others. For the columns, clusters 1 and 2 are strongly opposed and cluster 3 is the middle cluster.

7. Concluding remarks

In this paper, we have proposed a general approach to simultaneously analyze two sets which are related together in a data table by means of clustering. This approach has been applied with success to 3 kinds of data and has provided algorithms which used the most adapted information measures. It remains to apply this approach to other tables such as qualitative data or to other clustering structures such as hierarchical clustering.

The extension of this approach to tables which have more than 2 dimensions does not set up any problem, provided that the structure of the table allows it (binary table or contingency table).

The clustering methods of only one set could have been modeled by the mixture approach for quantitative data (Celeux and Govaert 1992, 1993) or binary data (Govaert 1990, Celeux and Govaert 1991, Gyllenberg, Koski and Verlaan 1994). Then, it would interesting to extend this modeling to simultaneously clustering methods defined in this work.

References

ANDERBERG M., (1973) Cluster Analysis for Applications, New York, Academic Press.

ANDERSON T. W., (1958) An Introduction to Multivariate Statistical Analysis, New York, Wiley.

ARABIE P. AND HUBERT L. J., (1990) The Bond Energy Algorithm Revisited, IEEE Transactions on Systems, Man, and Cybernetics, **20**, 1, 268-274.

BENZECRI J.-P., (1973) L'Analyse des Données, tome 1, Paris, Dunod.

BERTIN J., (1977) La Graphique et le Traitement Graphique de l'Information, Paris, Flammarion.

BOCK H., (1979) Simultaneous Clustering of Objects and Variables, in Analyse des données et Informatique, E. Diday et al. eds, Le Chesnay, INRIA, 187-203.

CELEUX G., DIDAY E., GOVAERT G., LECHEVALLIER Y., RALAMBONDRAINY H., (1989) Classification Automatique des Données: Environnement Statistique et Informatique, Paris, Dunod.

CELEUX G., DIEBOLT J., (1985) The SEM Algorithm: A Probabilistic Teacher Algorithm derived from the EM Algorithm for the Mixture Problem, Computational Statistics Quarterly, **2**, 73-82.

CELEUX G., GOVAERT G., (1991) Clustering Criteria for Discrete Data and Latent Class Models, Journal of Classification, **8**, 157-176.

CELEUX G., GOVAERT G., (1992) A Classification EM Algorithm for Clustering and two Stochastic Versions, Computational Statistics and Data Analysis, **14**, 315-332.

CELEUX G., GOVAERT G., (1993) Comparison of the Mixture and the Classification Maximum Likelihood in Cluster Analysis, J. Statis. Comput. Simul., **47**, 127-146.

CELEUX G., SOROMENHO G. (1995) An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model, Journal of Classification, to appear.

DIDAY E. AND COLL., (1980) Optimisation en Classification Automatique, Le Chesnay, INRIA.

DUFFY D. E., QUIROZ A. J., (1991) A Permutation-Based Algorithm for Block Clustering, Journal of Classification, **8**, 65-91.

FISHER W., (1969) Clustering and Aggregation in Economics, Baltimore, The Johns Hopkins Press.

GARCIA H., PROTH J. M., (1986) A new cross-decomposition algorithm: The GPM comparison with the bond energy method, Contr. Cybern., **15**, 155-165.

GOODMAN L. A., (1985) The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories: Associated Models, Correlation Models, and Asymmetry Models for Contingency Tables with or without Missing Entries, Annals of Statistics,**13**,10-69.

GOVAERT G., (1977) Algorithme de Classification d'un Tableau de Contingence, Premières journées internationales Analyse de Données et Informatique, Le Chesnay, INRIA, 487-500.

GOVAERT G., (1984) Algorithme de classification d'un tableau de contingence, in Data analysis and informatics 3, E. Diday et al., eds, Amsterdam, North-Holland, 223-236.

GOVAERT G., (1990) Classification Binaire et Modèles, Revue de Statistique Appliquée, **38**, 1, 67-81.

GOWER J. C., (1974) Maximal Predictive Classification, Biometrics, **30**, 643-654.

GREENACRE M., (1984) Theory and Applications of Correspondence Analysis, London, Academic Press.

GREENACRE M., (1988) Clustering the Rows and Columns of a Contingency Tables, Journal of Classification, **5**, 39-51.

GYLLENBERG M., KOSKI T., VERLAAN M., (1994) Classification of Binary Vectors by Stochastic Complexity, University of Turku (Finland), Institute for Applied Mathematics, Research report A5.

HARTIGAN J., (1972) Direct Clustering of a Data Matrix, JASA, **67**, 337, 123-129.

HARTIGAN J., (1975) Clustering Algorithms, New York, John Wiley & Sons.

JAMBU M., (1976) Sur l'Interprétation Naturelle d'une Classification Hiérarchique et d'une Analyse des Correspondances, Revue de Statistiques Appliquées **24**, 2, 45-72.

LEBART L., MORINEAU A., WARWICK K., (1984) Multivariate Descriptive Statistical Analysis, New York, Wiley.

LERMAN I.C., (1981) Classification et Analyse Ordinale des Données, Paris, Dunod.

MARCHOTORCHINO F., (1987) Block seriation problems: A unified approach, Appl. Stochastic Models and Data Anal., **3**, 73-91.

MC CORMICK W. T., SCHEITZER P. J., WHITE T. W., (1972) Problem decomposition and data reorganization by a clustering technique, Oper. Res., **20**, 993-1009.

TOLEDANO J., BROUSSE J., (1977) Une Méthode de Classification Simultanée des Lignes et des Colonnes, Premières journées internationales Analyse de Données et Informatique, Le Chesnay, INRIA, 105-108.

TUCKER L. R., (1964) The extension of factor analysis to three-dimensional matrices, in Contribution to Mathematical Psychology, N. Frederiksen and H. Gulliksen, eds., New York, Holt, Rinehart and Winston, 109-127.