Seventh International Conference on Multivariate Analysis
Barcelona Meeting, September, 21 - 24, 1992
in "Multivariate Analysis, Future Directions, C.Cuadras, C.R.Rao, Eds, North Holland, 1993, p 341-357.

(corrected version)

CORRESPONDENCE ANALYSIS AND CLASSIFICATION

LEBART L.,
Centre National de la Recherche Scientifique,
Ecole Nationale Supérieure des Télécommunications,
46, rue Barrault, 75013, Paris, France

and

MIRKIN B.G.
Department of Applied Statistics and Informatics
Central Economics-Mathematics Institute of Russian Academy of Sciences
(Currently at International Energy Agency,
2, rue André Pascal, 75775, Paris Cedex, 16, France)

1. Introduction

The present paper contains a survey of some of the most salient results about the links and the complementarity between clustering and correspondence analysis (CA) of contingency tables. It includes also a presentation of certain new contributions and domains of research.

The practitioners use to complement one approach with the other when a thorough exploration of data is needed, since the two points of view may provide quite different portrays of data. The involved processes are obviously distinct (projection onto a principal subspace on the one hand, grouping of similar categories on the other) but they could lead to identical results in specific situations. In more general cases, the parameters they produce are not independent. We will precisely focus on this interdependence and these specific situations below.

Two characteristics of CA are in favour of a reconciliation with classification : *the symmetry of the roles of rows and columns* in the process, and the *property of distributional equivalence* (Benzecri, 1973; Escofier, 1978; Gilula, 1986; Greenacre, 1988), allowing for a great stability of the results when agglomerating elements with similar profiles. Agglomerating the rows or the columns of a contingency table is "natural" in the sense that it is merely replacing classes by classes (instead of replacing individuals by groups, or variables by groups of variables...).

The questions of clustering in contingency data tables based on grouping of homogeneous items are discussed in Cazes (1986), Escoufier (1988), Greenacre (1988), Gilula (1986), Goodman (1981), Jambu (1978).

Seventh International Conference on Multivariate Analysis
Barcelona Meeting, September, 21 - 24, 1992
in "Multivariate Analysis, Future Directions, C.Cuadras, C.R.Rao, Eds, North Holland, 1993, p 341-357.

2. Some links between the two approaches

One can find a series of theoretical bridges between these approaches, exemplified by some particular models. We discuss below a set of such models having in mind the following purposes: to unify the previous developments and to propose certain new approaches. Let us illustrate this discussion with a numerical example of a symmetric 8 by 8 contingency table $K_{IJ} = (k_{ij})$ comprising k=640 cases (table 1). The marginals k_i and k_j are identical (equal to 80) in this particular example, but all the results concern as well the cases with unequal marginals.

Table 1
Contingency table K_{IJ}

	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8
LIG1	30	18	12	12	2	2	2	2
LIG2	18	30	12	12	2	2	2	2
LIG3	12	12	27	21	2	2	2	2
LIG4	12	12	21	27	2	2	2	2
LIG5	2	2	2	2	24	20	14	14
LIG6	2	2	2	2	20	24	14	14
LIG7	2	2	2	2	14	14	23	21
LIG8	2	2	2	2	14	14	21	23

The agglomerative clustering methodology based on chi-square distance using a generalized Ward criterion (see Benzecri, 1973; Greenacre, 1988; Jambu, 1978), agglomerates the elements pairwise, as it is shown on Figure 1.

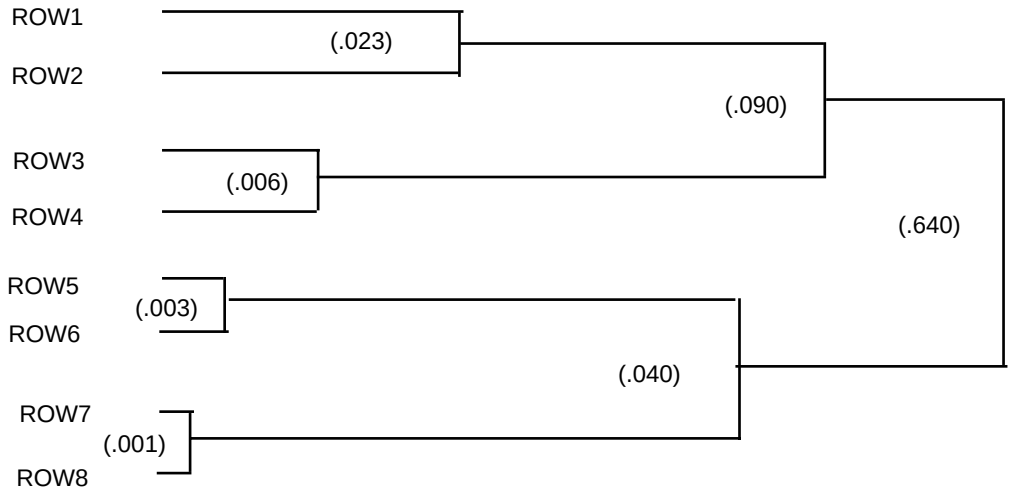

Figure 1
Sketched dendrogram issued from hierarchical clustering of the (8,8) table K_{IJ}

Moving centers (k-means) method based on chi-square distance gives similar results. For example, beginning with the two centers corresponding to the elements 1 and 8, we obtain easily the 2-class partition presented corresponding to the upper part of the dendrogram (cf. Figure1).

To express both the symmetry and distributional equivalence in a unified form let us consider for each $i \in I$ and $j \in J$ the value

$$q_{ij} = (k \infty k_{ij})/(k_i k_j) \; - \; 1 \qquad (i \in I, j \in J)$$

which expresses the relative increment (or decrement) RIP(i/j) of the probability of row i due to the knowledge of column j. Dual interpretation of q_{ij} as the relative increment RIP(j/i) of the probability of column j due to row i is straightforward. Relative increments for subsets are defined in analogous way using the total probabilities (or frequencies). The RIP values in table 1 are calculated by multiplying the entries by .1 and substracting 1 afterwards. Note that we have the two following relationships expressing the classical Chi-square X^2 as a function of the RIP coefficients :

$$X^2 = \Sigma_{ij} \; k_{ij} \; q_{ij} \quad = \quad (1/k) \; \Sigma_{ij} \; k_i \, k_j \; (q_{ij})^2$$

The RIP concept is useful in many aspects (Mirkin, 1985, 1992). In the present context we should point out that the RIP concept underlies the basic reconstruction formulas of CA :

$$q_{ij} \; = \; \Sigma_{h \in H} \, \mu_h F_h(i) G_h(j) \qquad (1)$$

where F_h, G_h are CA factors corresponding to singular value μ_h ($h \in H$).

2.1 A global approximation formulation

Using this concept, the distributional equivalence principle can be specified in a symmetric form as follows. In rough terms, the block structure of the coinciding RIP values in matrix $Q = \{q_{ij}\}$ reflects the CA presentation in such a way that the sub-arrays (boxes) of the equal RIP values correspond to the sets of the equal row or column-points in CA space. This can be expressed also in terms of the equalities (1) using Boolean vectors instead of CA factors. Explicitely, let the classes of some partitions $\{V_s : s \in S\}$ on I and $\{W_t : t \in T\}$ on J represent the sets of coinciding row and column-points in CA space. The formulas (1) express the principle if $H = S \times T$ and Boolean $F_h(i)$, $G_h(j)$ are defined for $h=(s,t)$ as follows: $F_h(i)=1$ iff $i \; \varpi \; V_s$ and $G_h(j)=1$ iff $j \in W_t$. This form of the principle allows us to formulate the partitioning problem of a contingency table K_{IJ} as an approximation problem: to find a pair of partitions, $\{V_s : s \in S\}$ on I and $\{W_t : t \in T\}$ on J,

and corresponding values μ_h for h=(s,t) to approximate the RIP matrix $Q=\{q_{ij}\}$, that is, to minimize the difference between the left and right parts of (1) (in the Boolean form) measured by the weighted least square criterion L^2 such that :

$$L^2 = \Sigma_{ij}\, k_i\, k_j \; [q_{ij} - \Sigma_h \mu_h F_h(i) G_h(j)]^2 \qquad (2)$$

(the weight of the entry (i,j) is to be equal to $k_i k_j$ (Carroll, Pruzansky, and Green, 1977; Escoufier, 1988). When the user wants to clusterize only one of the sets, I (or, J) the corresponding partition of J (or, of I) consists of the set of singletones.

Evidently, for F_h and G_h ($h \in H$) fixed, the optimal values μ_h are equal to the corresponding RIP values, that is, for each h = (s,t), the optimal value is such that $\mu_h = q_{st}$.

It is not difficult to prove also that the alternating algorithm for minimizing L^2 is equivalent to the chi-square distance moving centers method, and that an agglomerative suboptimal algorithm is equivalent to the chi-square distance based agglomerating clustering procedure using generalized Ward criterion (Mirkin, 1992). The value of the criterion can be expressed through the difference of the chi-square contingency coefficients for the initial and aggregated contingency tables: $L^2 = (X^2(I,J) - X^2(S,T))$. This approach can account for various results and findings derived in Benzecri et al.(1980), Cazes (1986), Moussaoui (1987), Jambu (1978).

2.2 Simultaneous clustering of rows and columns

This approximation clustering approach can be expanded to the problems of finding "mixed" clusters containing the rows and columns simultaneously (an approach dating back to Hartigan, 1972 ; Braverman et al., 1974; see also Govaert, 1977; Bock, 1979). The chi-square distance concept cannot help in this matter, since no satisfactory concept exists to measure distance between a row and a column! But we can consider the model (1) as a set of approximate equalities with arbitrary Boolean vectors F_h and G_h (and corresponding cluster boxes $B_h=\{(i,j) : F_h(i)=1 \text{ and } G_h(j)=1\}$) to find out.

A suboptimal algorithm to fit the model (1) for this case was developed in Mirkin (1992): the cluster boxes B_h are separated sequentially maximizing the accounted part of the general value of $X^2(I,J)$; the values μ_h are estimated by the RIP values of the cluster boxes obtained.

More explicitely, each iteration h (the index h is omitted below for convenience) aims at minimizing the following reduced form of criterion (2) :

$$L^2 = \Sigma_{ij}\, k_i\, k_j \; [q_{ij} - \mu F(i) G(j)]^2 \qquad (3)$$

μ, F(i) and G(j) are unknown, whereas the q_{ij} are the residuals computed after each iteration (for the first iteration, the q_{ij} are the initial RIP values).

The optimal μ for any fixed box VxW (defined as V = {i : F(i) = 1} and W = {j : G(j) = 1}) is determined by the weighted average of q_{ij} computed within the box :

$$\mu = \Sigma_{i \in V} \Sigma_{j \in W} k_i k_j q_{ij} / k_v k_W$$

which equals $q_{v W}$.

Substituting this value into (3) leads to the following equality :

$$L^2 = \Sigma_{i \in I} \Sigma_{j \in J} k_i k_j (q_{ij})^2 - \mu^2 k_v k_W$$

which shows that minimizing L^2 is equivalent to maximizing the following form of the criterion, depending on box VxW only :

$$g(V,W) = \mu^2 k_v k_W = (\Sigma_{i \in V} \Sigma_{j \in W} k_i k_j q_{ij})^2 / k_v k_W$$

To maximize this criterion, the following step-by-step procedure of box generation can be performed : each step adds to the box issued from the previous step only one element, a row or a column, to maximize the increment of the criterion due to the added element. At the first step, two elements are simultaneously selected : a row i and a column j , maximizing g({i},{j}) for all the pairs of singletones. The process stops when the maximal increment becomes negative. The suboptimal cluster box obtained through this algorithm has the following property (Mirkin, 1992) : For each row i or column j outside the cluster box, the absolute value of the relative increments $q_{Vj} = q_{jV}$ and $q_{iW} = q_{Wi}$ are at least twice smaller than the absolute value of the relative "internal" increment $q_{VW} = q_{WV}$.

The residual data in this sequential fitting procedure are obtained through substracting the solution provided by the h-th iteration from the residual data of the preceding iteration.

$$q_{ij,h+1} = q_{ij,h} - \mu_h F_h(i) G_h(j) \qquad (i \in I, j \in J).$$

For the first iteration, $q_{ij,1} = q_{ij}$ $\qquad (i \in I, j \in J).$

Even in the case of overlapping boxes, the initial Chi-square can be partitioned into components corresponding to these boxes in order to evaluate the contribution of each cluster, and to help fixing the number of cluster (by using traditional values of the

accumulated contributions, or testing the hypothesis of independence for the residual data).

The obtained boxes are shown to correspond to certain fragments of the CA space (maximally connected if $\mu_h>0$, or maximally disconnected if $\mu_h<0$). In our example, the algorithm separates, initially, the singletone boxes $\{1\}\times\{1\}$ and $\{2\}\times\{2\}$ (each having 2 as the RIP value and accounting 7.8% of $X^2(I,J)$), then the pair segments $\{3,4\}\times\{3,4\}$, $\{5,6\}\times\{5,6\}$, and $\{7,8\}\times\{7,8\}$ are obtained sequentially followed by the link boxes ($\{1\}\times\{2\}$ and $\{2\}\times\{1\}$) for the first two elements. The RIP value for each of these boxes is positive (evidently, the values are 1.4, 1.2, 1.2, .8, .8, respectively). Then boxes $\{1,2,3,4\}\times\{5,6,7,8\}$ and $\{5,6,7,8\}\times\{1,2,3,4\}$ appear having negative RIP value .8. All this structure accounts for 95.8% of the $X^2(I,J)$.

2.3 An example of coincidence between clustering and C.A.

Unfortunately, the Boolean form of decomposition (1) has no longer the weighted orthonormality properties of the CA factors. But for the symmetric matrices K_{II} (which is exactly the case of our example), Benzecri (1973, vol.2, ch.11) has pointed out a situation where the discrete orthonormal eigen-functions are relevant.

This author has derived a representation of a binary hierarchy H through a set of orthogonal functions allowing to build a symmetric contingency table (through the reconstruction formula) whose CA restitutes the initial hierarchy.

The preceding symmetric (8,8) contingency table $\mathbf{K_{IJ}}$ has thus the property of providing an exact coincidence between correspondence analysis and hierarchical clustering (using the Ward's criterion) in the following sense : each eigenvalue of the CA corresponds exactly to a node of the classification.

Table 2

Eigenvalues issued from the C.A. of $\mathbf{K_{IJ}}$

$\lambda 1$ =	.640	(80 % of the trace)
$\lambda 2$ =	.090	(11 %)
$\lambda 3$ =	.040	(5 %)
$\lambda 4$ =	.023	(3 %)
$\lambda 5$ =	.006	(.7 %)
$\lambda 6$ =	.003	(.4 %)
$\lambda 7$ =	.001	(.1 %)

The associated axis of the CA separates the two sets of elements constituting this node.

Correspondence analysis of table $\mathbf{K_{IJ}}$ leads to 7 clearly separated eigenvalues (see table 2).

The sequence of patterns that can be observed in the columns of table 3 (eigen-vectors) is typical of a hierarchical structure : the non-zero coordinates on each principal axis can take only two distinct values, opposing two groups of elements.

Table 3

Principal Coordinates issued from the C.A. of $\mathbf{K_{IJ}}$

Axes		1	2	3	4	5	6	
ROW1	*	-.80	.42	0.00	.30	0.00	.00	*
ROW2	*	-.80	.42	0.00	-.30	0.00	.00	*
ROW3	*	-.80	-.42	0.00	0.00	-.15	.00	*
ROW4	*	-.80	-.42	0.00	0.00	.15	.00	*
ROW5	*	.80	0.00	-.28	0.00	0.00	.10	*
ROW6	*	.80	0.00	-.28	0.00	0.00	-.10	*
ROW7	*	.80	0.00	.28	0.00	0.00	.00	*
ROW8	*	.80	0.00	.28	.00	0.00	.00	*

The first axis, for instance, opposes (ROW1 ... ROW4) to (ROW5 ... ROW8). The second axis, within the first group isolated by axis 1, opposes (ROW1, ROW2) to (ROW3, ROW4), etc.. Correspondence analysis performs in this case like a divisive algorithm, working iteratively from the upper to the lower level of a hierarchy.

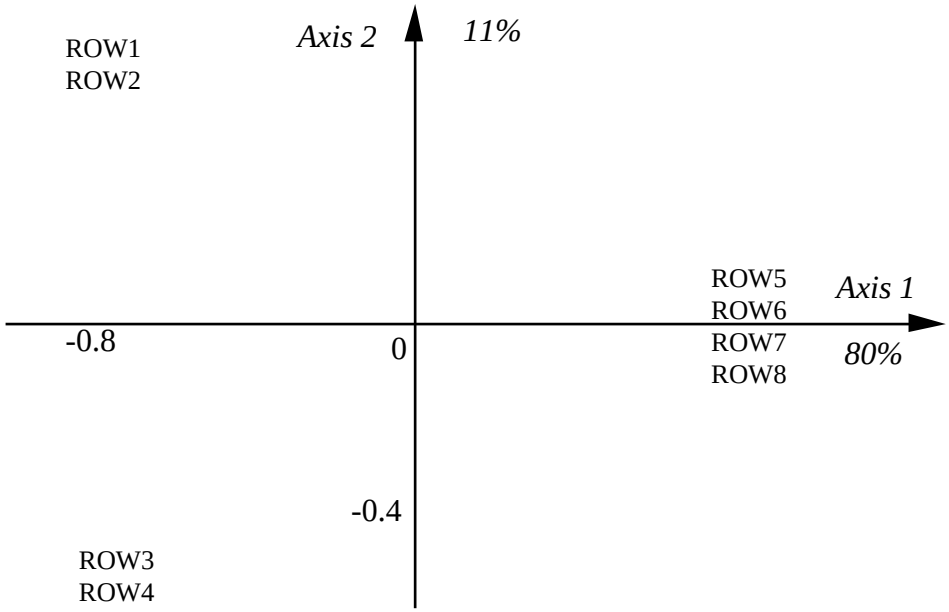

Figure 2

Planar display of table K_{IJ} through CA.

We notice that the configuration of points in the principal plane of CA (Figure 2) highlights only a limited part of the underlying structure, by comparison with the dendrogram (also a planar representation) of Figure 1.

Figure 2 gives neither pertinent information about the distances between ROW1 and ROW2 (the corresponding points are superimposed on the plane, suggesting a null distance), nor useful information about the distances between ROW4, ROW5, ROW6, also superimposed on the graphical display. This shrinkage of distances, easily explained by the geometrical properties of the initial swarm of points, should prompt the users to use simultaneously the two kinds of methods to obtain a reliable description of the data.

2.4 Properties of these "compatible" matrices

The above example concerns the case of a binary hierarchy H, whose each nonterminal element $h \in H$ can be partitioned in a unique way into two sets a(h) and b(h) belonging to H. The orthonormal set of "3-valued" functions f_h is defined as follows: $f_h(i)$ equals d_a for $i \in a(h)$, $-d_b$ for $i \in b(h)$, 0 for others elements i, where d_a, d_b are chosen in order to make the average of f_h equal to zero, and the norm equal to 1.
Evidently, $d_a = [\ (k \times k_{b(h)}) / (k_{a(h)} k_h)\ ]^{1/2}$, $d_b = [\ (k \times k_{a(h)}) / (k_{b(h)} k_h)\ ]^{1/2}$).
We say that a square symmetric contingency table is *compatible* if (1) holds for some binary hierarchy H with $F_h(i)=f_h(i)$, $G_h(j)=f_h(j)$ and some $\mu_h>0$ ($h \in H$). In general, a method to approximate the RIP values with those 3-valued eigen-function decomposition can be developed. The method fits model (1) sequentially, each iteration finding a bi-partition of current set h into two subsets, a(h) and b(h), to minimize the weighted least square criterion, or, equivalently, to maximize the "explained" part of the chi-square value which is shown to be equal to :

$$(\mu_h)^2 = (q_{a(h)a(h)} + q_{b(h)b(h)} - 2\, q_{a(h)b(h)})^2.$$

This divisive clustering procedure, in our example, leads to the hierarchy of Figure 1.

3. Eigenvalues and indices

3.1 Some inequalities

The largest eigenvalues issued from the CA of a contingency table are greater or equal to the largest index corresponding to the last node of a hierarchical clustering of the rows or

of the column of this contingency table (using the chi-square distance and the generalized Ward criterion to ensure a compatibility between the two techniques). The equality occurs for special tables such as the compatible matrices dealt with in the previous section. This upper bound for the indices could be derived easily from the above considerations since the indices and the eigenvalues appear to be solutions of the same optimization problem, with supplementary constraints for the indices. Benzecri and Cazes (1978) have more generally shown that the quantity $(\lambda_1 + \lambda_2 + .. \lambda_p)$ is greater or equal than the sum of the p indices corresponding to the p highest nodes of the associated hierarchy (a property which can be directly derived from the general criterion (2), where F and G are less constrained in the case of CA). Moreover, these authors have produced a counter-example showing that there exists no general lower bound for the index corresponding to the highest node : one can find distributions of density such that the largest index remains an arbitrarily small fraction of the largest eigenvalue.

3.2 The case of block-structured contingency tables

The limiting case of multiple eigenvalues $\lambda_i = 1$, (i=1,m) (for the CA of a rectangular contingency table) is particularly interesting since it is closely related to the classification of rows and columns (the trivial eigenvalue 1 corresponding to a constant eigenvector is supposed to be removed beforehand). It is straightforward that such multiple 1-eigenvalues exist iff there exist a block-structure of the contingency table into m+1 blocks. (i.e. : iff only m +1 diagonal blocks contains non-zero elements). Surprisingly enough, no similar property hold for the most usual agglomerative algorithms. Kharchaf and Rousseau (1988,1989) present some counter-examples of bloc-structures in contingency tables easily recognised by CA, although undetected by an agglomerative clustering technique.

3.3 Experiments about the joint distribution of indices and eigenvalues

We give in this section some empirical results about the joint behavior of the indices and the eigenvalues issued from the same random contingency table.
Under the hypothesis of independence (also called homogeneity in the case of contingency tables), a series of 1000 pseudo-random independent (8,8) contingency table with equal theoretical marginal are generated, according to a multinomial scheme. For each generated table, the total number of observations k is 1000.

Table 4
Mean values and standard deviations of the Eigenvalues and the Clustering indices.
(1000 independent random (8 , 8) contingency tables C . For each table C, k = 1000.)

Identifier	*Mean-value*	*Standard deviation*	*Standard deviation of the mean*
Eigenvalues			
EV1	.02130	.00560	.00018
EV2	.01282	.00353	.00011
EV3	.00772	.00234	.00007
EV4	.00442	.00156	.00005
EV5	.00214	.00100	.00003
EV6	.00070	.00050	.00002
EV7	.00010	.00014	.00000
Indices of rows (INRi) **and columns** (INCi)			
INR1	.01692	.00452	.00014
INR2	.01063	.00289	.00009
INR3	.00733	.00197	.00006
INR4	.00537	.00148	.00005
INR5	.00391	.00117	.00004
INR6	.00280	.00090	.00003
INR7	.00183	.00074	.00002
INC1	.01679	.00450	.00014
INC2	.01061	.00291	.00009
INC3	.00739	.00202	.00006
INC4	.00535	.00151	.00005
INC5	.00396	.00118	.00004
INC6	.00280	.00091	.00003
INC7	.00182	.00075	.00002

The 7 eigenvalues issued from the CA of each table as well as the 7 indices of the hierarchical classification of the rows and the columns (always using the generalized Ward criterion and the chi-square distances) of the same table are computed, enabling to estimates the means, variances and correlations relating to these 21 variates.
Table 4 summarizes the results concerning the means, the standard deviations of the initial variables and the standard deviations of the means.

The results concerning the eigenvalues are consistent with some previous approximations (Lebart, 1976), since their distribution is similar to the one of the eigenvalues of a Wishart matrix (n=7, p=7). The sum τ of the means of the different eigenvalues equals 0.0492 ; the statistics $k\tau$ has thus the value 49.2 (no significant difference with the expectation of a Chi-square with 7x7 degrees of freedom)

The indices corresponding to the clustering of the rows and of the columns are distinct for each simulated matrix. The statistical identity of their first and second order moments is a further indication of the consistency of the simulation process.
As expected, the largest indices INR1 and INC1 are smaller than the largest eigenvalue λ_1=EV1, whereas the smallest indices INR7 and INC7 are on the average much larger their counterpart.

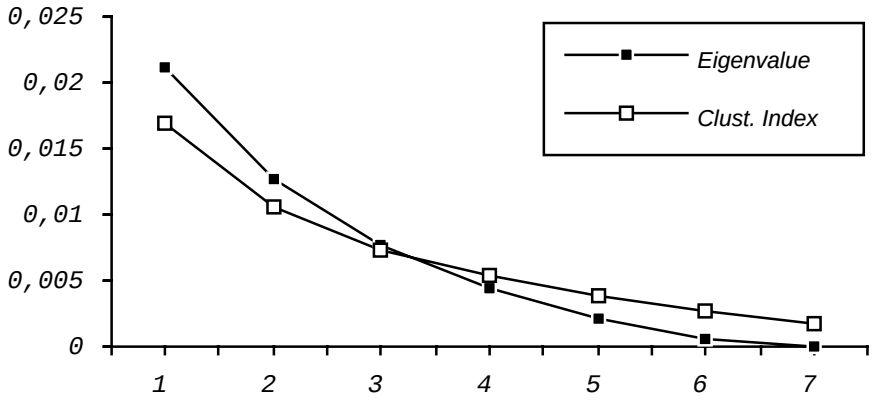

Figure 3. Sequences of eigenvalues and indices

Figure 3 shows the compared trajectories of these two quantities, highlighting the smaller range of variation of the indices.

Figure 4 below presents the scattering diagram of the joint distribution of the first eigenvalue λ_1 = EV1 and the first row-clustering index INR1 both issued from the same pseudo-random matrix. The correlation coefficient between λ_1 and INR1 is 0.91. (The same value is obtained for the correlation coefficient between λ_1 and INC1). The theoretical constraint INR1$\leq \lambda_1$ clearly defines the upper left boundary of the swarm of points.

Seventh International Conference on Multivariate Analysis
Barcelona Meeting, September, 21 - 24, 1992
in "Multivariate Analysis, Future Directions, C.Cuadras, C.R.Rao, Eds, North Holland, 1993, p 341-357.

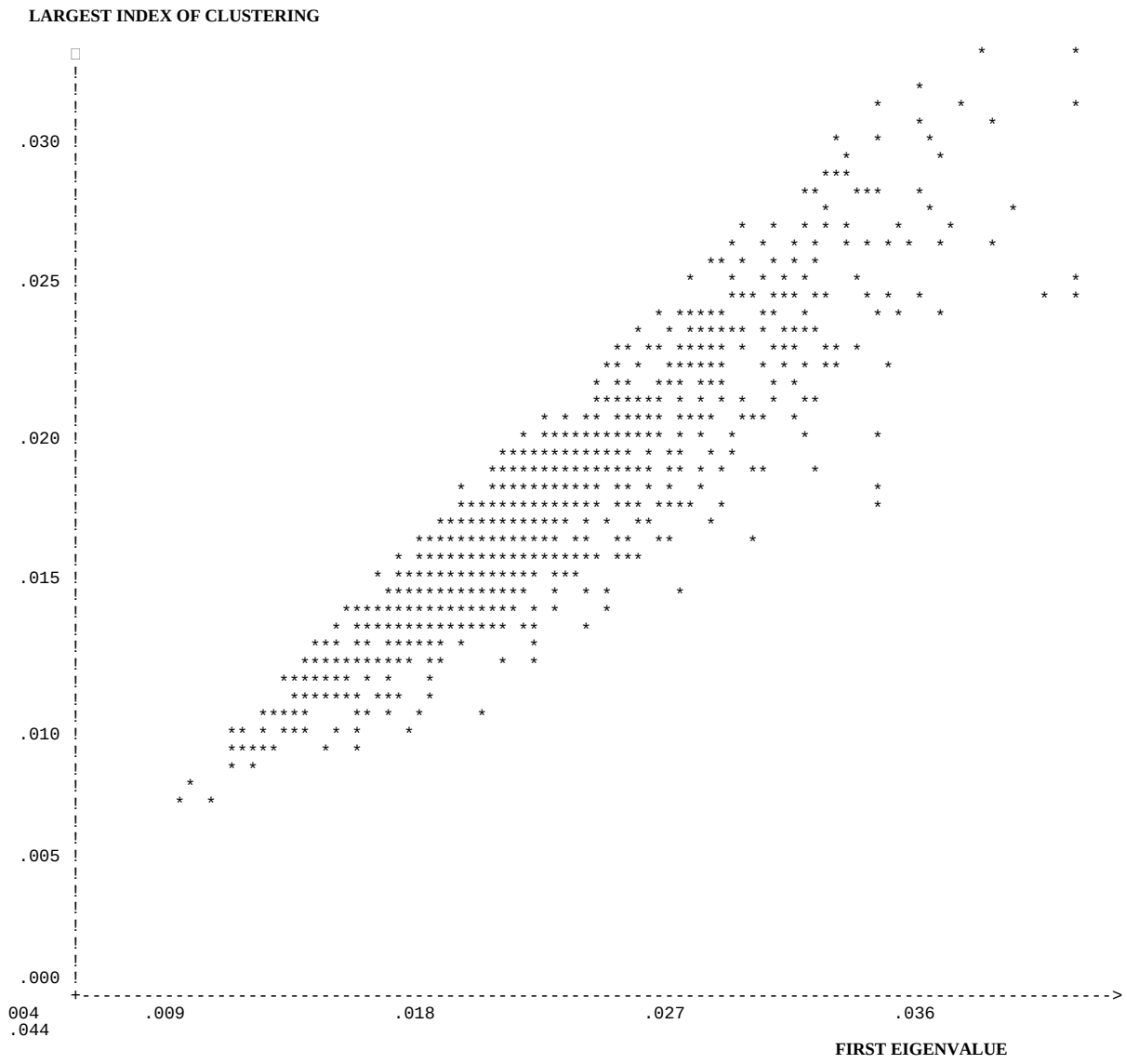

Figure 4
Correlation between $\lambda_1 = EV_1$ and the first clustering index INR1

To study the complex system of relationships between the various indices and the eigenvalues, we will visualize the corresponding correlation matrix through a principal component analysis (PCA), which will summarize the main observable patterns.

Figure 5 shows the principal plane of a PCA whose the active elements are the eigenvalues and the illustrative elements are the indices. A classical size effect (all the

coordinates on the first axis are positive) corresponds to the fact that all the involved correlation coefficients are positive.

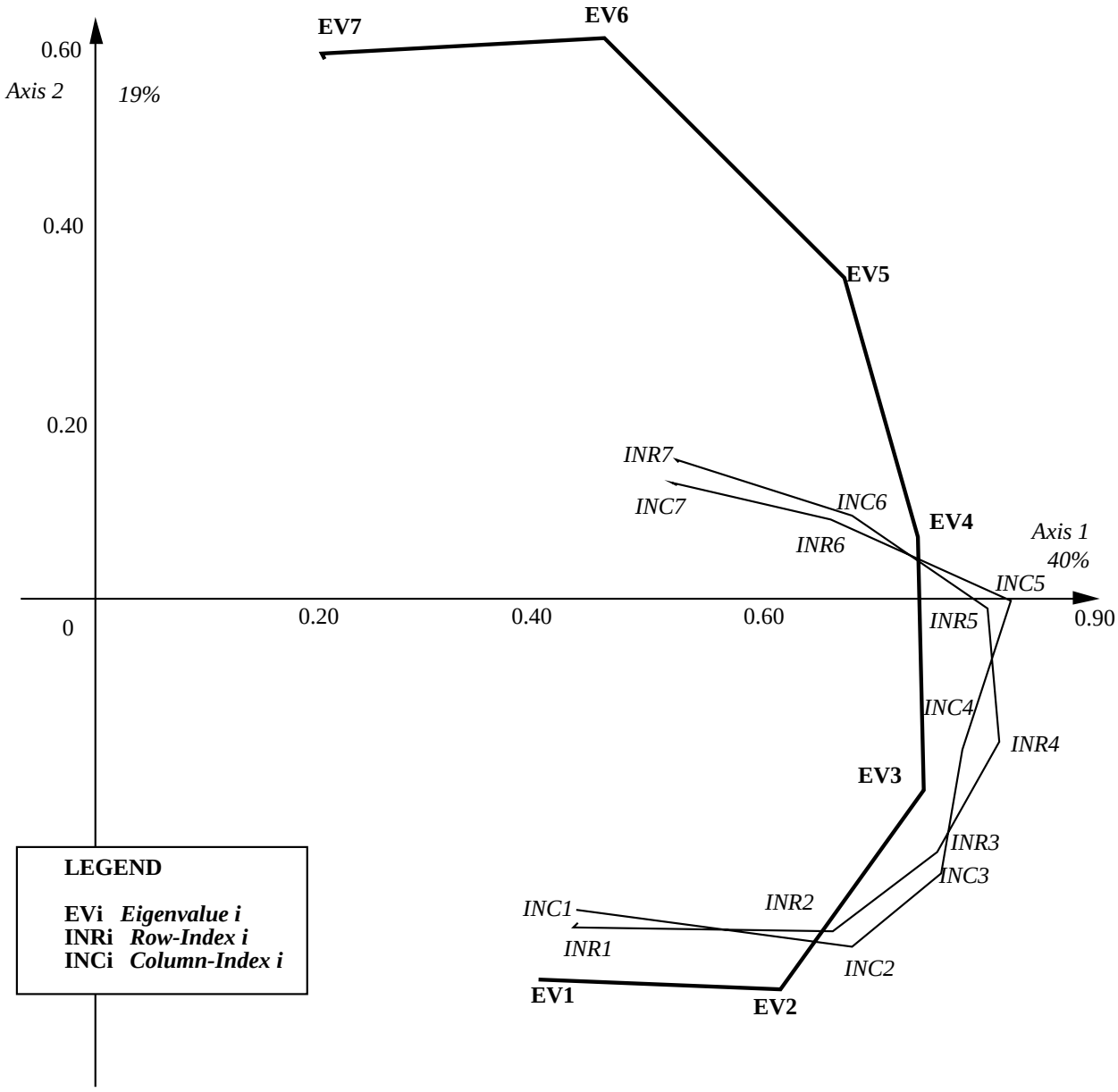

Figure 5. Structure of the correlation between eigenvalues and indices

(Principal plane of a Principal Component Analysis of the (1000,7) matrix containing the 1000 observations of the 7 eigenvalues EV1,...EV7.)
[Note that the 7 row-indices INR1, ...INR7 and the 7 column indices INC1,...INC7 have been projected afterwards as supplementary elements onto this principal plane]

The first indices are clearly correlated with the first eigenvalues. As mentioned previously, the two correlation coefficients between each of the largest indices (INR1 and

INC1) and the first eigenvalue take the value 0.91 (the correlation between INR1 and INR2 is only 0.80, but these relatively small differences are not visible on the display).
The positive autocorrelations beween successive eigenvalues or indices entail regular trajectories on the plane spanned by the two first principal components, but these trajectories diverge for the smallest eigenvalues and indices.
This pattern established from pseudo-random matrices is an assessment of the intuitive experience of the practitioners : on the one hand the upper part of the dendrogram provides the user with about the same results than the first axes ; on the other hand the lower part of the dendrogram often pinpoints some interesting local properties of the data, while the smallest eigenvalues correspond to some unidentifiable noise.

4. Some hybrid methods

Two series of works involving simultaneously at different level both CA (or other principal axes method) and clustering are briefly mentioned below.

4.1 Clustering involving optimal coding

In the case of individuals described by several categorical variables (these variables could be measured on nominal, ordinal or interval scales), van Buuren and Heiser (1989) propose an algorithm achieving simultaneoulsly a coding of the variable and a clustering of the individuals. An alternative least square algorithm is used, starting from a multiple correspondence analysis of the data table.

4.2 Principal axes method for displaying or discovering clusters.

Some techniques related to projection pursuit and discrimination can be considered also as an intermediate step between the two approaches.
Let us consider n objects described by p variables (y_{ij} is the value of variable j for object i). Furthermore, these objects are also the vertices of a symmetric graph G, whose associated matrix is **M** ($m_{ii'}$ = 1 if nodes i and i' are joined by an edge, $m_{ii'}$ = 0 otherwise). Such situation occurs when objects are time-points, geographic areas, or if they are assigned to a priori classes. *Contiguity Analysis* simultaneously uses the local covariance matrix **C** (such that $c_{jj'} = (1/2m) \Sigma_{i,i'} m_{ii'}(y_{ij} - y_{i'j})(y_{ij'} - y_{i'j'})$) , and the global covariance matrix **V**. If the graph is made of k disjoined complete subgraphs, **V** is very similar to the classical "within covariance matrix" used in linear discriminant analysis, and coincides with it when the graph is regular (i.e. each vertex is provided with the same number of edges). The

minimization of the ratio: u'**C**u / u'**V**u (u being a p-vector) provides then a generalization of linear discriminant analysis in the case of overlapping clusters (see for instance Aluja and Lebart., 1984).

Using more general similarity indices in place of the binary quantity $m_{ii'}$ allows to define a series of indices analogous to those used in Projection Pursuit (see Caussinus, 1992).

It is easy to derive a contiguity matrix from the basic data array itself: any threshold applied to the set of n(n-1) distances or similarities between observations allows to define a binary relationship which can be described by a symmetric graph. Similarly, a contiguity matrix can be derived, from the k nearest neighbours of each observation.

The contiguity analysis applied to such matrices (Burtschy and Lebart, 1991) is closely related to the techniques proposed by Gnanadesikan *et al.* (1982), Art *et al.* (1982). It produces planar (or low dimensional) representations which can be viewed as compromises between the outcomes of principal axis techniques (CA or PCA) and those of clustering techniques.

5. Complementarity from a practical point of view

Various authors have insisted upon the complementarity between principal axes techniques and classification, which concerns the comprehension of the data structure as well as the interpretation of the results. Gower and Ross (1969), for example, have shown how the drawing of a minimum spanning tree onto a principal plane issued from a principal component analysis could enrich the interpretation of the represented distances between points. Benzecri *et al*. (1980) have developed a thorough methodology for the conjoint use of CA and hierarchical clustering, comprising various parameters which describes the mutual links between axes and nodes.

CA , like PCA, could entail shrinkages and distorsions due to both the projection onto the principal dimensions and the possible lack of robustness of the global fit (sensitivity to outliers). It is then advisable to complement it with a classification performed in the whole space. The clusters are not only used to mark out the factorial planes by a sample of well described areas. Being derived in a much higher dimensional space, they can supply elements of information that could have been hidden by the projection onto a low dimensional subspace.

A practical issue reinforces this need for both approaches : it is much easier to describe a set of clusters than a continuous space. The most significant categories or variables for each cluster could be automatically selected, therefore producing a computer aided description of the classes, and hence, of the whole space. A series of statistical tests

allow to select and to sort (according to the computed levels of significance) the most characteristics items for each cluster (see for instance Lebart *et al*., 1984). .

From a purely computational point of view, when dealing with very large data sets such as those provided by survey data files, it may prove efficient to perform a classification using a limited number of factors issued from CA to increase the performances of the techniques (Morineau and Lebart., 1986).

Finally, the user may wish to discover some unexpected latent factors or some hidden existing groups within the data. Although the theoretical models underlying CA and classification are seldom referred to by exploratory data analysts, it is clear that each tool has its own vocation and idiosyncrasies. Even if the history of statistical applications abounds in examples of groups discovered through eigen-analyses as well as latent factors discovered through clustering, it seems wiser to systematically use both techniques.

References

Aluja Banet T., L. Lebart (1984). Local and Partial Principal Component Analysis and Correspondence Analysis, *COMPSTAT Proceedings*, 113-118, Physica Verlag, Vienna.

Art D., Gnanadesikan R, Kettenring J.R.(1982). Data Based Metrics for Cluster Analysis, *Utilitas Mathematica*, 21 A, 75-99.

Benzécri J.P. (1973) *Analyse des Données*.-Paris: Dunod.

Benzécri, J.P. (1983) Analyse d'inertie intraclasse par l'analyse d'un tableau de correspondance, *Les Cahiers d'Analyse des Données*, 8, no.3, 351-358.

Benzécri J.P., Cazes P. (1978) Problème sur la classification. *Les Cahiers d'Analyse des Données*, 3, no.1, 95-101.

Benzécri J.P., Jambu M. (1976) Agrégation suivant le saut minimum et arbre de longueur minimum. *Les Cahiers d'Analyse des Données*, 1, no.4, 441-452.

Benzécri, J.P., Lebeaux M.O., and Jambu M. (1980) Aides a l'interpretation en classification automatique, *Les Cahiers de l'Analyse des Données*, vol.V, n.1, 101-123.

Bock H. H. (1979) Simultaneous clustering of objects and variables. in Analyse des donnees et informatique, European C.C. Courses, INRIA, p 187-203.

Braverman E.M., Kiseleva N.E., Muchnik I.B., and Novikov, S.G. (1974) Linguistic approach to the problem of processing large bodies of data, *Automation and Remote Control*, 35, no.11, part 1, 1768-1788.

Burtschy B., and Lebart L. (1991) Contiguity analysis and projection pursuit, 117-128. in *Applied Stochastic Models and Data Analysis*, World Scientific , Singapore.

van Buuren S., and Heiser W.J. (1989) Clustering N objects into k groups under optimal scaling of variables, *Psychometrika*, 54, no.4, 699-706.

Carrol J.D., Pruzansky S., and Green P.F. (1977) Estimation of the parameters of Lazarsfeld's Latent Class Model by application of canonical decomposition CANDECOMP to multi-way contingency tables, *AT&T Bell Laboratories*, unpublished paper, 18 p.

Cazes P. (1986) Correspondance entre deux ensembles et partition de ces deux ensembles, *Les Cahiers de l'Analyse des Données*, vol.XI, no.3, 335-340.

Cazes P., and Moreau J. (1991) Contingency table in which the rows and columns have a graph structure, in E.Diday, Y.Lechevallier (Eds) *Symbolic-Numeric Data Analysis and Learning*, Nova Science Publishers: New York, 271-280.

Caussinus H.(1992). Projections Revelatrices in *Modèles pour l'Analyse des Données Multidimensionnelles*, J.J. Droesbeke, B. Fichet, P.Tassi, eds, Economica, Paris.

Escofier B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statist. Appl.* vol. 26, n°4, p 29-37.

Escoufier Y. (1988) Beyond correspondence analysis. In: H.H.Bock (Ed.) *Classification and Related Methods of Data Analysis*. Elsevier Sc.P.

Gilula Z. (1986) Grouping and association in contingency tables: an exploratory canonical correlation approach, *Journal of American Statistical Association*, vol.81, no.395, 773-779.

Gnanadesikan R., Kettenring J.R., Landwehr J.M. (1982). Projection Plots for Displaying Clusters, *in Statistics and Probability, Essays in Honor of C.R. Rao*, G. Kallianpur, P.R. Krishnaiah, J.K.Ghosh, eds, North-Holland.

Goodman L.A. (1991) Measures, models, and graphical displays in the analysis of cross-classified data (with Discussion), *Journal of American Statistical Association*, vol.86, No.416, 1085-1138.

Goodman L.A.(1981) Criteria for determining whether certain categories in a cross-classification table should be combined with special reference to occupational categories in an occupational mobility table, *American Journal of Sociology*, 87, 612-650.

Govaert G. (1977) Algorithme de classification d'un tableau de contingence. In: *"Premières Journées Internationales Analyse des Données et Informatique (Versailles 1977)*" INRIA, p. 487-500.

Gower J.C, Ross G. (1969) Minimum spanning tree and single linkage cluster analysis. *Appl.Statistics*, vol 18, p 54-64.

Greenacre M.J. (1988) Clustering the rows and columns of a contingency table, *Journal of Classification*, 5, 39-51.

Hartigan J.A. (1972) Direct clustering of a data matrix, *Journal of American Statistical Association*, vol.67, p. 123-129.

Jambu M. (1978) *Classification Automatique pour l'Analyse des Données, I- Méthodes et Algorithms.* Paris:Dunod.

Kharchaf I., Rousseau R. (1988, 1989) Reconnaissance de la structure de blocs d'un tableau de correspondance par la classification ascendante hiérarchique: parts 1 and 2, *Les Cahiers de l'Analyse des Données*, vol.XIII, n.4, 439-443; vol.XIV, n.3, 257-266.

Lebart L. (1976) The significance of eigenvalues issued from correspondence analysis. *Proceedings in Comp. Stat., COMPSTAT*, Physica verlag, Wien, p 38-45.

Lebart L., Morineau A., Warwick K. (1984) - *Multivariate Descriptive Statistical Analysis*, J.Wiley, New-York.

Marcotorchino F. (1987) Block seriation problems: a unified approach, *Journal of Applied Stochastical Models and Data Analysis*, vol.3, no.3, 73-93.

Mirkin B.G. (1985) *Grouping in SocioEconomic Studies*. Finansy i Statistika Publishers, Moscow (in Russian).

Mirkin B.G. (1992) Correspondence-wise clustering for contingency tables, *submitted for publication*.

Morineau A., Lebart L. (1986) Specific Clustering Algorithms for Large data sets and Implementation in SPAD Software. in *Classification as a Tool of Research*, Gaul W., Schader M., Eds, North Holland, 1986.

Moussaoui A.E. (1987) Sur la reconstruction approchée d'un tableau de correspondance a partir du tableau cumulé par blocs suivant deux partitions des ensembles I et J, *Les Cahiers de l'Analyse des Données*, vol.XII, n.3, 365-370.

Key-words :
Correspondence Analysis, Clustering techniques, Classification, Hybrid approaches in Data Analysis, Contingency tables.