Multidimensional data analysis for categorical variables

Gilbert Saporta

MULTIDIMENSIONAL DATA ANALYSIS FOR CATEGORICAL VARIABLES

Gilbert Saporta
Conservatoire National
des Arts et des Métiers
292, rue Saint-Martin
75141 Paris, France

1. INTRODUCTION

Before fitting a model to empirical data, an exploratory phase seems essential: a data analytic step where graphics play a major part in understanding the structure of the data.

Until the last decade, not much attention has been paid to the case of categorical variables, even though these occur very frequently in social sciences. Apart from cross-tabulations, unidimensional pie-charts and frequency diagrams, one could not find any other exploratory or descriptive technique for categorical variables in the standard statistical software such as SPSS, BMDP or SAS. (However, Gestalt allows correspondence analysis). It is worth noticing that even the recent revival of exploratory techniques tends to ignore multivariate categorical data (e.g., Tukey, 1977, Gnanadesikan, 1977) and is focused on metric variables.

However, statistical tools allowing the description of categorical variables similar to principal component analysis for metric data have existed for a long time in the psychometric literature. Rediscovered under the name of correspondence analysis (Benzecri, 1973, Hill, 1974, see for an historical survey, Nishisato, 1980) this method is an efficient way of describing multidimensional categorical data and may be associated successfully with other techniques, such as cluster analysis.

Section 2 is devoted to the bivariate case which is of theoretical and practical importance since many data sets are presented under the form of two-dimensional contingency tables, especially in official statistics. An example concerning French agricultural land use illustrates the use of correspondence and cluster analysis. In Section 3 we show that the analysis of a data set consisting of n individuals described by p categorical variables may be performed by a variant of correspondence analysis which is identical to the principal components of scales proposed by Guttman (1941). An example is given concerning expenditures of consumers in 17 European countries.

2. DESCRIPTION OF A CONTINGENCY TABLE

Let N be a two-dimensional contingency table with elements n_{ij}, resulting from the cross-classification of two categorical variables X_1 and X_2 with m_1 and m_2 categories respectively; D_1 = diag (n_i) and D_2 = diag (n_j) are the diagonal matrices of row and column totals.

As an illustration, consider the 22x9 array giving the distribution of agricultural land use in 1979 for 22 French regions according to 9 categories of land use of holdings. The items represented at the top of the table have the following meaning: CERE: cereals, AGRG: other general agriculture, VINE: vineyards, FRUIT: fruit, MILK: milk production (mainly), MEAT: beef and veal meat production (mainly), SHEEP: sheep, AGSH: general agriculture with sheep.

	CERE	AGRG	VINE	FRUIT	MILK	MEAT	MIX	SHEEP	AGSH
ILDF ILE-DE FRANCE	389.2	173.0	0.0	4.7	1.3	2.1	0.8	3.1	12.7
CHAM CHAMPAGNE-ARDENNE	328.3	514.6	32.1	1.2	49.1	29.5	75.0	27.2	142.6
PICA PICARDIE	188.8	690.1	2.5	3.3	47.1	12.7	21.8	13.5	162.4
HNOR HAUTE NORMANDIE	78.0	200.0	0.0	2.8	50.9	35.9	87.6	31.7	95.5
CENT CENTRE	1155.1	378.1	14.1	15.4	14.1	84.4	37.8	103.3	256.0
BNOR VASSE NORMANDIE	26.3	76.8	0.0	2.3	587.1	94.6	137.7	58.6	53.9
BOUR BOURGOGNE	325.6	135.2	34.3	5.5	19.0	453.8	129.6	139.6	192.9
NORD NORD-PAS DE CALAIS	25.7	333.2	0.0	1.1	53.3	9.0	14.1	13.8	127.6
LORR LORRAINE	83.7	50.0	0.0	2.3	131.0	31.3	103.0	47.4	134.3
ALSA ALSACE	44.5	31.2	17.1	1.7	21.2	3.6	14.2	8.5	45.2
FRCO FRANCE-COMTE	21.8	16.5	1.1	0.8	279.2	19.7	40.3	43.0	29.9
LOIR PAYS DE LA LOIRE	69.9	48.9	30.1	20.2	337.6	266.3	477.9	129.3	92.7
BRET BRETAGNE	22.6	94.1	0.0	3.3	726.5	54.4	61.1	37.8	54.4
POIT POITOU-CHARENTES	178.6	105.6	0.9	4.9	37.1	95.8	84.7	311.9	156.2
AQUI AQUITAINE	166.5	134.7	134.3	27.7	94.2	81.5	32.8	130.3	135.9
MIDI MIDI-PYRENEES	262.5	227.0	6.3	30.0	165.7	237.0	26.7	421.7	170.2
LIMO LIMOUSIN	1.7	2.9	0.0	3.7	43.5	406.9	72.2	281.3	8.4
RHON RHONE-ALPES	84.0	87.5	51.2	47.2	341.5	77.8	96.5	248.1	77.5
AUVE AUVERGNE	57.1	38.4	0.2	2.1	455.3	294.4	115.9	257.0	42.2
LANG LANGUEDOC-ROUSILLON	25.9	35.1	136.9	53.8	56.7	51.9	14.3	249.7	9.8
PROV PROVENCE-APLES-COTE D'AZUR	40.2	46.8	61.2	62.8	7.6	8.9	11.3	157.4	15.9
CORS CORSE	0.3	0.6	3.8	4.6	0.1	29.6	0.0	49.6	0.2

Table 1. Agricultural land use for 22 French regions (in 000 ha.)

2.1. Simultaneous Graphical Display of Rows and Columns

One of the simplest ways of presenting correspondence analysis is the following method of "reciprocal averaging" or "dual scaling" (Hill, 1974; Nishisato, 1980). Suppose that the 9 categories of land use are displayed as points with coordinates b_j (j=1,..9) over an axis. It then seems natural to represent the ith region by a point a_i which is the weighted mean of the b_j, where the weights are the conditional frequencies n_{ij}/n_i.

$$a_i = \sum_j \frac{n_{ij}}{n_{i\bullet}} b_j \qquad (1)$$

In vector notation with $\underline{a} = (a_1 \ldots a_{m_1})'$ and $b = (b_1 \ldots b_{m_1})'$ We have

$$\underline{a} = D_1^{-1} N \underline{b} \qquad (2)$$

Of course the b_j are arbitrary scores but we may skip this drawback if we want conversely the b_j to be centroids of the a_i with weights $n_{ij}/n_{\cdot j}$

$$\underline{b} = D_2^{-1} N' \underline{a} \qquad (3)$$

and by substitution we will get two separate equations.

Unfortunately equations (2) and (3) do not hold simultaneously unless all coordinates are identical which is rather uninteresting: we cannot have at the same time the b_j's as means of the a_i's and vice-versa.

So we need weaker conditions:

$$\underline{a} = \alpha D_1^{-1} N \underline{b} \; ; \; \underline{b} = \beta D_2^{-1} N' \underline{a} \qquad (4)$$

where the constants α and β have to be as close as possible to one.

Then by substituting we get with $\lambda = (\alpha\beta)^{-1}$

$$\lambda \underline{a} = D_1^{-1} N D_2^{-1} N' \underline{a}$$

$$\lambda \underline{b} = D_2^{-1} N' D_1^{-1} N \underline{b} \qquad (5)$$

The best solution is obtained when choosing the largest eigenvalue less than 1 (ignoring the unit eigenvalue).

The coordinates of rows and columns are given by the corresponding eigenvectors $\underline{a}$ and $\underline{b}$ of the two matrices $D_1^{-1}\ N\ D_2^{-1}\ N'$ and $D_2^{-1}\ N'\ D_1^{-1}\ N$ which are the products of the two arrays of conditional frequencies.

It is straightforward to prove that $\alpha = \beta = \lambda^{-1/2}$ and that the elimination of the extraneous solution $\lambda=1$ provides solutions where rows and colums are of zero mean:

$$\sum_i \frac{n_{ij}}{n_{i.}} a_i = \sum_j \frac{n_{ij}}{n_{.j}} b_j = 0 \qquad (6)$$

So far we have obtained a one-dimensional display of the categories of the two variables x_1 and x_2.

To obtain two- or higher dimensional plots we have just to take the eigenvectors associated with the second and third largest non-unit eigenvalues of equations (5), as the following subsection will demonstrate.

2.2. Connection with Principal Components Analysis

Correspondence analysis is nothing else than two separate but dual principal components analyses performed successively on the rows and then on the columns of N. The difference from ordinary P.C.A. consists in using weights for both rows and columns of N.

Let us consider the array $D_1^{-1}\ N$, that is to say, the matrix of row frequencies associated with N. Region i is described by m_2 variables (the row-frequencies) and must of course be weighted by its marginal frequency $n_{i.}/n$. If we want to apply P.C.A. to get a low-dimensional configuration of the 22 regions we have moreover to specify some measure of distance between regions. In many respects the so-called chi-square distance (Guttman, 1941):

$$d^2_{x^2}(i,i') = \sum_{j=1}^{m_2} \frac{n}{n_{.j}} \left(\frac{n_{i'j}}{n_{i'.}} - \frac{n_{ij}}{n_{i.}}\right)^2 \qquad (7)$$

is preferable to the usual one without the $\frac{n}{n_{.j}}$ terms since it avoids the implicit discarding of categories with small frequencies in computing the distance.

But the importance of the chi-square distance relies on the following property: the two possible P.C.A., for rows and for columns, are in a strict duality when choosing chi-square distance for rows and for columns respectively.

More precisely, in the first analysis the matrix of weighted sums of squares and products is

$$(D_1^{-1}\ N)'\ D_1\ (D_1^{-1}\ N) = N'\ D_1^{-1}\ N$$

Since we use the chi-square distance associated with the quadratic form of matrix D_2^{-1}, we have to pre-multiply $N'D_1^{-1}\ N$ by D_2^{-1} to get the matrix which has as its eigenvectors $\underline{u}$ the linear combinations providing the principal components, $D_2^{-1}\ N'\ D_1^{-1}$. The values of the principal components or coordinates along the principal axis are thus given by $\underline{a} = D_1^{-1}\ N\ \underline{u}$. Since

$$D_2^{-1}\ N'\ D_1^{-1}\ N\ \underline{u} = \lambda \underline{u}$$

we have $D_1^{-1}\ N\ D_2^{-1}\ N'\ \underline{a} = \lambda \underline{a}$

The other P.C.A. comes down to exchange the principal coordinates with the principal components loadings.

Since the vectors of coordinates are principal components they are in a natural way normalized by:

$$\sum_i \frac{n_{i.}}{n} (a_i)^2 = \sum_j \frac{n_{.j}}{n} (b_j)^2 = \lambda \qquad (8)$$

and the various eigenvectors of $D_2^{-1}\ N'\ D_1^{-1}\ N$ are orthogonal: the scores of various order are uncorrelated variables.

The quantity $\frac{n_{i.}}{n} (a_i)^2/\lambda$ is called the "contribution of the i^{th} category to the eigenvalue".

The reader will have noticed that the two P.C.A. are performed without setting to zero the means of the variables: once the trivial solution $a_i = b_j = 1, \forall\ i, j$, is discarded, the other solutions are necessarily of zero-mean.

The simultaneous representation of rows and columns of N is therefore nothing else than the superposition of the two separate scatter-plots provided by the two P.C.A.

The closeness of two row-points or of two column-points is easy to understand: they have roughly the same conditional distributions to the degree that the reduction of dimensionality is not misleading. However, it seems more difficult to interpret the proximity between a row-point and a column-point. We will see further in Section 3 that this proximity can be interpreted in terms of proximity of means of categories.

2.3. Canonical Decomposition of Contingency Tables

The whole set of eigenvectors $\underline{a}^{(k)}$ and $\underline{b}^{(k)}$ can give an exact reconstruction of the table:

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{n} \left(1 + \sum_{k=2}^{\min m_1 m_2} \frac{a_i^{(k)}\, b_j^{(k)}}{\sqrt{\lambda_k}}\right) \qquad (9)$$

(see Kendall and Stuart, 1961, for instance).

If we use only the first k eigenvectors (including the trivial solution) we obtain the best approximation of the array N by a matrix of rank k in the following sense:

$$\text{if } \Phi_{ij} = \frac{n_{i.} n_{j.}}{n} \left(\sum_{1=1}^{k} \frac{a_i^{(1)} b_j^{(1)}}{\lambda_1}\right)$$

then the Φ_{ij} are such that they minimize:

$$\sum_i \sum_j \frac{(n_{ij} - \Phi_{ij})^2}{n_{i.}\, n_{.j}} \qquad (10)$$

Since $\mathrm{tr}\,(D_1^{-1} N D_2^{-1} N') = \sum_i \sum_j \frac{n_{ij}^2}{n_{i.}\, n_{.j}}$, the sum of non trivial eigenvalues is equal to the Pearson's $\emptyset^2$, i.e., to the usual chi-square statistic for testing the independence divided by

$$n\left(\sum_{k=2} \lambda_k\right) = x^2 \qquad (11)$$

The preceding properties show that correspondence analysis is a way of analyzing the structure of dependencies in a contingency table. The scores $\underline{a}^{(k)}$ and $\underline{b}^{(k)}$ are pairs of artificial variables each pair representing in some sense a part of the association between x_1 and x_2. It may be proved too that the eigenvalue λ_k is the squared correlation coefficient between the artificial variables $\underline{a}^{(k)}$ and $\underline{b}^{(k)}$ (Hirschfeld, 1935).

2.4. Example

A correspondence analysis applied to the data of Table 1 gives the following results:

	eigenvalue	percent	cumulated percent
1.	.407	.389	.389
2.	.247	.236	.626
3.	.148	.142	.768
4.	.113	.108	.875
5.	.067	.065	.940
6.	.033	.031	.971
7.	.023	.022	.993
8.	.007	.007	1.000

Table 2: Results of correspondence analysis

Thus the first four dimensions "extract" near 90 percent of the structure.

The decomposition of the first four eigenvalues according to rows and columns (formula (8)) is useful for interpretative purposes (Table 3).

Contributions of colums (with the sign of the coordinate)

CERE	.278+	.000-	.068-	.460-
AGRG	.222+	.090+	.063+	.330+
VINE	.001-	.116-	.318+	.014-
FRUI	.001-	.060-	.142+	.010-
MILK	.330-	.331+	.034+	.075-
MEAT	.056-	.123+	.295-	.061+
MIX	.048-	.005-	.060-	.023+
SHEEP	.039-	.270-	.020+	.002+
AGSH	.025+	.004+	.000-	.024+

Contributions of rows

ILDF	.097	.004	.009	.109
CHAM	.083	.019	.006	.026
PICA	.108	.054	.027	.203
HNOR	.008	.010	.000	.044
CENT	.200	.000	.043	.281
BNOR	.087	.102	.001	.009
BOUR	.002	.044	.133	.005
NORD	.030	.034	.021	.166
LORR	.001	.009	.004	.000
ALSA	.003	.000	.004	.001
FRCO	.041	.048	.004	.019
LOIR	.068	.001	.052	.013
BRET	.013	.183	.020	.033
POIT	.001	.029	.005	.002
AQUI	.002	.026	.077	.005
MIDI	.000	.030	.001	.002
LIMO	.043	.133	.131	.062
RHON	.028	.001	.040	.007
AUVE	.082	.000	.019	.000
LANG	.009	.152	.240	.006
PROV	.000	.093	.164	.004
CORS	.003	.031	.000	.002

Table 3: Decomposition of eigenvalues according to rows and columns

We see clearly on the diagram (Fig. 1) that the horizontal axis separates the regions into two main categories: those specializing in milk production at the left side (Bretagne, Basse-Normandie.) and those specializing in cereals and general agriculture on the right side (Ile de France, Centre, Picardie mainly). This represents the main feature of the data set.

The second dimension is characterized by milk production again, opposed this time by "sheep", "wines" and "fruit". We find at the bottom of the map regions of the Mediterranean coast.

The third axis is characteristic of an opposition between regions devoted mainly to "wine production" and regions devoted to "meat production". The fourth dimension provides no new information but allows the separation between "cereals" and "general agriculture" which was not distinguishable along the first axis.

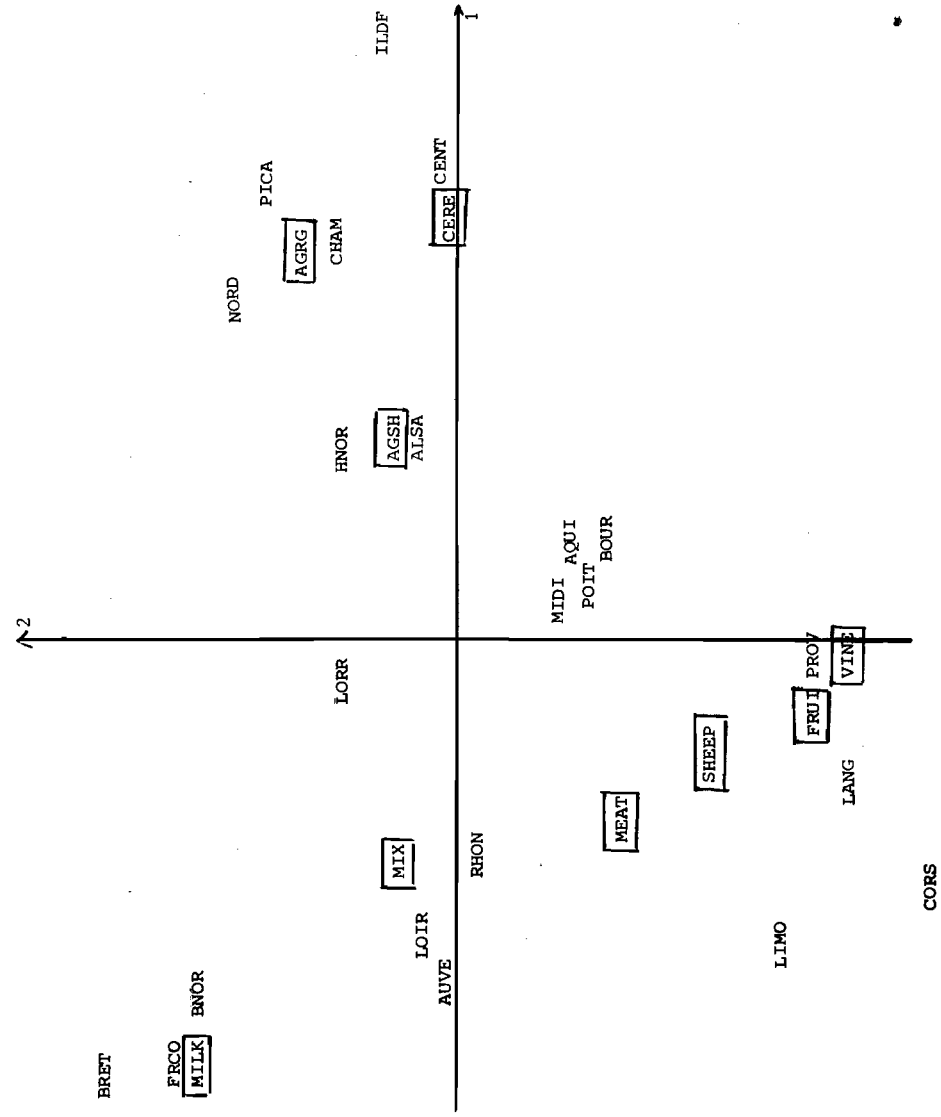
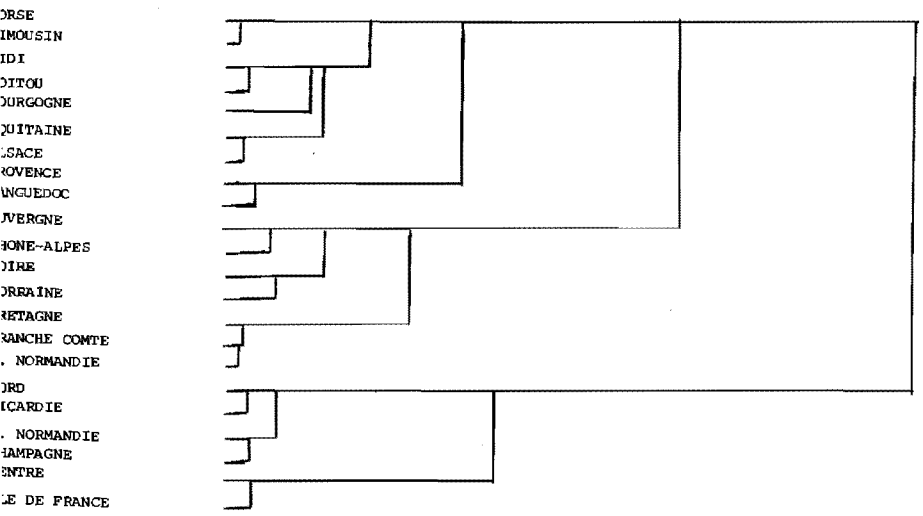


Figure 1: Joint representation of regions and agricultural land users.

nce the chi-square distance is an Euclidean distance, it is pos-
ble to use metric methods of cluster analysis in order to classify
d identify common spatial patterns. For clustering regions,
rd's hierarchical technique seems appropriate because for contin-
ncy tables it comes down to the following algorithm: at each step
merge the two lines of the contingency table that lead to the
nimal loss of the phi-square measure of association. The hierar-
ical level of a cluster is exactly equal to that loss and the
ms of all levels is thus, of course, equal to the $\emptyset^2$ statistic
r the whole array (Benzecri, 1973; Jambu, 1978; Bouroche and
oorta, 1980).

e reader should note that Ward's method consists of maximizing
each step the intercluster dispersion measured as the weighted
an of the squared distances between centroids, or equivalently
minimizing the within-cluster dispersion.

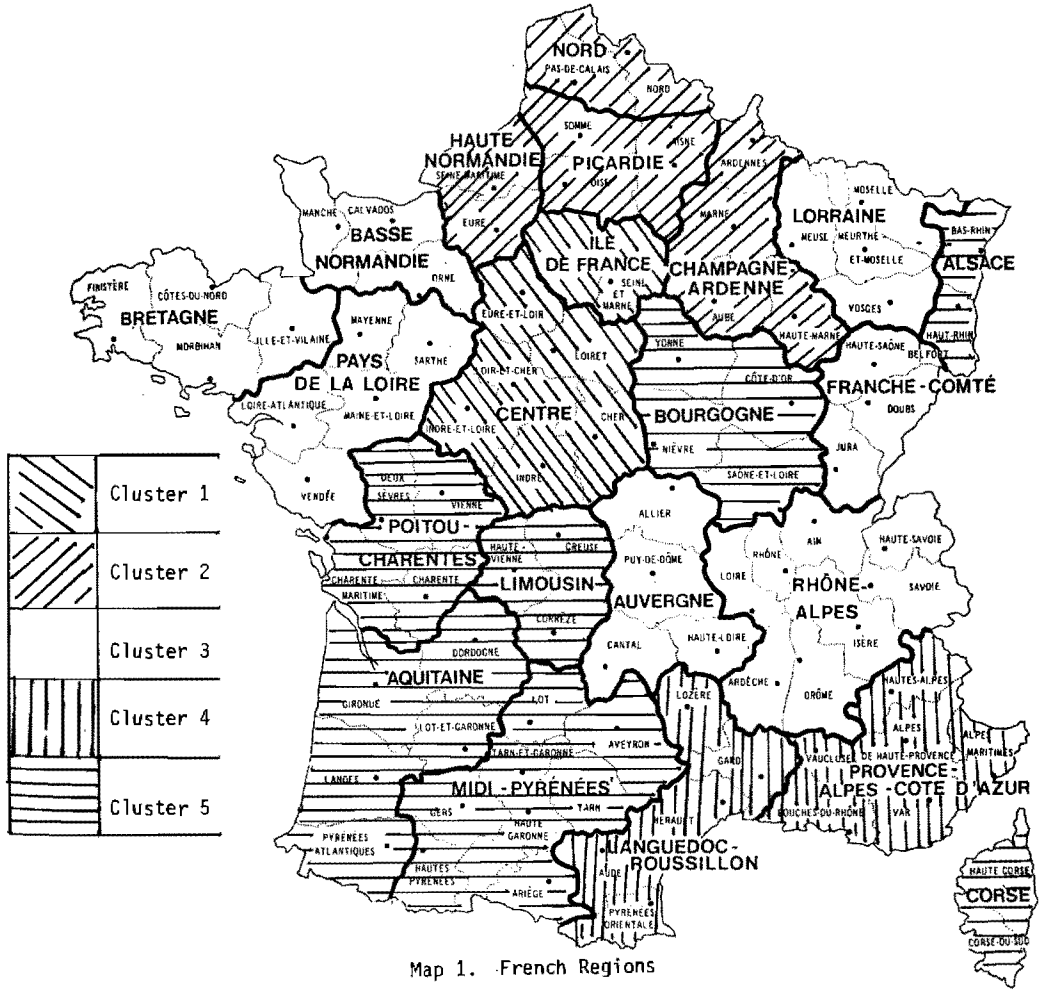
e result is the following tree diagram (Figure 2).



jure 2: A tree diagram of French regions

We may identify distinctly 5 clusters followed by their main characteristics:

1. Centre, Ile de France/ cereals
2. Nord, Picardie, Haute-Normandie, Champagne/ general agriculture
3. Basse-Normandie, France-Comté, Bretagne, Lorraine, Loire, Rhône-Alpes, Auvergne/ milk production
4. Provence, Languedoc/ fruits, wines
5. Alsace, Aquitaine, Bourgogne, Poitou, Midi, Limousin, Corse/breeding



Map 1. French Regions

3. GENERAL CASE: p CATEGORICAL VARIABLES

3.1. Notation

N individuals are described by p categorical variables $x_1, x_2, \ldots, x_p$ with $m_1, m_2, \ldots, m_p$ categories. The data are usually presented in the form of a nxp array with entries x_i^j = "symbol of the category of x^j for individual i".

For mathematical convenience only, it is interesting to use the following representation called dijunctive form: we associate with each categorical variable x_j the n x m_j matrix X_j where the columns are the indicator variables of x_j:

X_j (i,k) = 1 if individual i belongs to category k

X_j (i,k) = 0 if not

$X = (X_1, X_2, \ldots, X_p)$ is then the supermatrix with Σm_j columns, with the X_j as its blocks.

D is the diagonal matrix with, as its diagonal blocks, the matrices D_j of marginal frequencies of the x_j's.

The method that will be considered in the next subsection is used in different countries under a great variety of names: homogeneity analysis in the Netherlands (van Rijckevorsel and de Leeuw, 1979), quantification in Japan (Hayashi, 1950), principal components of scales in the U.S.A (Guttman, 1941), dual scaling in Canada (Nishisato, 1980), and so on. The pioneering work is indeed Guttman (1941), but at that time the method was considered as a technique for obtaining only a one-dimensional scale. (For a comprehensive monograph, see McKeon, 1966).

In this paper we use the name "multiple correspondence analysis" since it is the terminology used in France (Lebart Morineau and Tabard, 1977; Deville and Saporta, 1983) but it is only for convenience since there is not yet an international agreement about a unique name. Indeed, the presentation will be very similar to what is commonly regarded as homogeneity analysis.

3.2. Multiple Correspondence Analysis

If the variables were metric, say x_j, the method of principal components would provide a graphical display of the individuals by

means of the coordinates along the principal axes.

If $\underline{z}^{(1)}$ denotes the n-vector of coordinates along the first principal axis it is well known that $\underline{z}^{(1)}$ is the variable which maximizes:

$$\sum_{j=i}^{p} r^2 (\underline{z}^{(1)}; \underline{x}^j) \qquad (13)$$

If we want to make a similar analysis for categorical variables it seems natural to modify the preceding criterion to take into account the nature of the variables. This will briefly be described here. Taking as a measure of relationship the squared correlation ratio $\eta^2 (\underline{z} ; x_j)$ instead of the squared correlation coefficient r^2, leads to a generalization of correspondence analysis.

$\eta^2(\underline{z}; x_j)$ is defined as a variance ratio:

$$\frac{\text{variance of the } m_j \text{ means of } \underline{z} \text{ for the categories of } x_j}{\text{variance of } \underline{z}}$$

If $\underline{z}$ is zero-mean and of unit variance a straightforward matrix calculation shows that

$$\eta^2(\underline{z} ; x_j) = \underline{z}' X_j (X_j' X_j)^{-1} X_j' \underline{z}$$

Maximization of (13) then reduces to an eigenvalue problem and the optimal $\underline{z}$ is the solution of

$$\sum_{j=i}^{p} X_j (X_j' X_j)^{-1} X_j' \underline{z} = \mu \underline{z}$$

or

$$\sum_{j=i}^{p} X_j D_j^{-1} X_j' \underline{z} = \mu \underline{z} \qquad (14)$$

Since $\sum_{j=i}^{p} X_j D_j^{-1} X_j'$ may be rewritten as $X D^{-1} X'$ if we divide both sides by p

$$\frac{1}{p} X D^{-1} X' \underline{z} = \frac{\mu}{p} \underline{z} = \lambda \underline{z} \qquad (15)$$

equation (15) looks like a correspondence analysis equation (cf. equation (5)). Indeed, pI is the diagonal matrix of row totals of X, D is the diagonal matrix of column totals of X and $\frac{1}{p} X D^{-1} X'$ is the product of the two "conditional" arrays associated with X. Thus it is necessary simply to apply correspondence analysis to X, formally considered as a contingency table.

Using the duality between solutions associated with rows and columns it follows that:

$$\underline{z} = \lambda^{-1/2} \frac{1}{p} X \underline{a} = \lambda^{-1/2} \left(\frac{1}{p} \Sigma X_j \underline{a}_j \right) \qquad (17)$$

where vector $\underline{a}$ (made of subvectors $\underline{a}_j$) is a solution of:

$$\frac{1}{p} D^{-1} X'X \underline{a} = \lambda \underline{a} \qquad (18)$$

which is a more convenient equation than (15) since it is of size Σm_j and not of size n.

Applying again the duality relations gives:

$$\underline{a} = \lambda^{-1/2} D^{-1} X' \underline{z} \qquad (19)$$

Apart from the constant $\lambda^{-1/2}$ the coordinate of a column of X along an axis of correspondence analysis is just the mean value of variable $\underline{z}$ for those individuals belonging to the category corresponding to this column.

Equation (17) establishes conversely an analogous property for the rows: we find again here the principle of dual scaling used in section 2 which could have been taken as a definition of multiple correspondence analysis.

It must be pointed out that here the sum of non-trivial eigenvalues (there exist Σm_j-p such eigenvalues) is a constant equal to $(\frac{1}{p} \Sigma m_j)$-1 and has no statistical meaning. One should not attach much importance to the percentages of explained variance which are generally small since $\lambda<1$.

If p = 2 multiple correspondence analysis gives the same results as usual correspondence analysis: precisely the super vector $\underline{a}$ contains as subvectors the two-vectors $\underline{a}$ and $\underline{b}$ of coordinates of rows and columns of the contingency table $N = X_1'X_2$. The eigenvalues, however, are not the same: they are transformed by

$$\frac{1 \pm \sqrt{\lambda}}{2}$$

If in the data set some variables are numerical, it is possible to handle them after splitting their values into categories.

Surprisingly enough, the discretization of numerical variables is very efficient and does not lead to a loss of information: actually, it is a way to avoid the linearity of classical treatments (Masson, 1974; Gifi, 1981).

3.3. An Example

From a study concerning the standard of living in Europe ("L'Expansion" April, 1979) we have taken the following array giving the household expenditure according to eight variables, each with three categories (1 = low, 2 = medium, 3 = high). (see Table 4).

Since there are 8 variables, each with three categories, there are 24-8 = 16 non trivial eigenvalues the sum of which equals 2.

The first four eigenvalues are:

$$\lambda_1 = 0.380 \quad \lambda_2 = 0.349 \quad \lambda_3 = 0.219 \quad \lambda_4 = 0.184$$

The first principal plane (Figure 4) reveals the following features: the horizontal axis shows a strong difference between countries where household expenditures for food are high (Greece, Portugal, Spain, Italy, Ireland) and those where household expenditures for home are high (Belgium, West Germany, France, Finland). The first principal component is positively correlated with variable HOME and negatively with variable FOOD as it can be seen with the order of categories along the first axis.

The vertical axis isolates a group of three countries (Switzerland, Sweden, Denmark) characterized by a high level of housing and heating expenditure and a low level of clothing expenditures.

A cluster analysis using Ward's algorithm gives a confirmation of the three clusters outlined in Figure 4.

Here cluster analysis, applied to the rows of the disjunctive array with the chi-square metric, presents some interesting features. The similarity measure associated with the chi-square metric is a (weighted) scalar product between rows:

	Food (FO)	Clothing (WE)	Housing & Heating (LO)	Home Equipment (HO)	Health (HE)	Transport (TR)	Education (ED)	Others (OT)
W. Germany	1	2	2	3	2	3	2	2
Austria	1	3	2	3	2	2	3	1
Belgium	1	1	2	3	3	2	3	2
Denmark	2	1	1	3	1	1	3	3
Spain	3	3	1	1	2	2	2	3
Finland	2	2	2	3	2	3	2	2
France	1	2	1	3	3	2	2	2
Greece	3	3	3	1	2	1	1	1
Ireland	3	2	1	2	1	2	1	1
Italy	3	3	2	2	1	1	2	1
Luxembourg	1	2	2	2	2	2	1	3
Norway	1	3	1	2	1	3	3	2
The Netherlands	1	2	2	2	3	1	2	3
Portugal	3	2	1	1	1	1	1	2
United Kingdom	2	2	3	2	2	2	3	1
Sweden	1	1	3	1	1	3	3	3
Switzerland	2	1	3	1	3	1	3	3

Table 4: Qualitative cross-national data on household expenditures

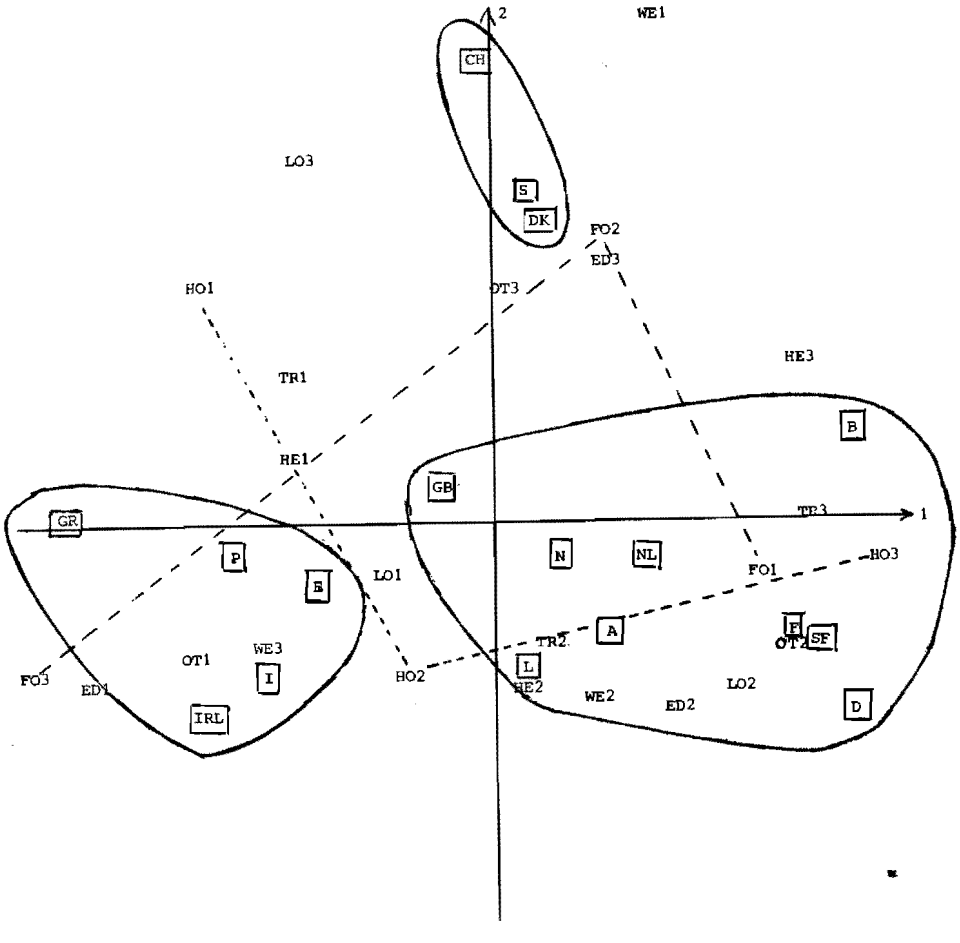


Figure 3. Results of a qualitative correspondence analysis

$$S(i,i') = \sum_j \frac{np}{n_{\cdot j}} x_i^j x_{i'}^j \qquad (20)$$

Since the elements x_i^j are 0 or 1, $S(i,i')$ is the sum over the p categorical variables of the following elementary similarity measure:

$$S_x(i,i') = 0 \text{ if } i \text{ and } i' \text{ do not belong to the same category of } x$$

$$= \frac{np}{n_{\cdot j}} \text{ if } i \text{ and } i' \text{ belong to the category } j \text{ of } x \text{ .}$$

In other words the similarity between i and i' for variable x is greater if i and i' belong both to an uncommon category than if i and i' belong to a common one. This seems to be a nice property for categorical variables.

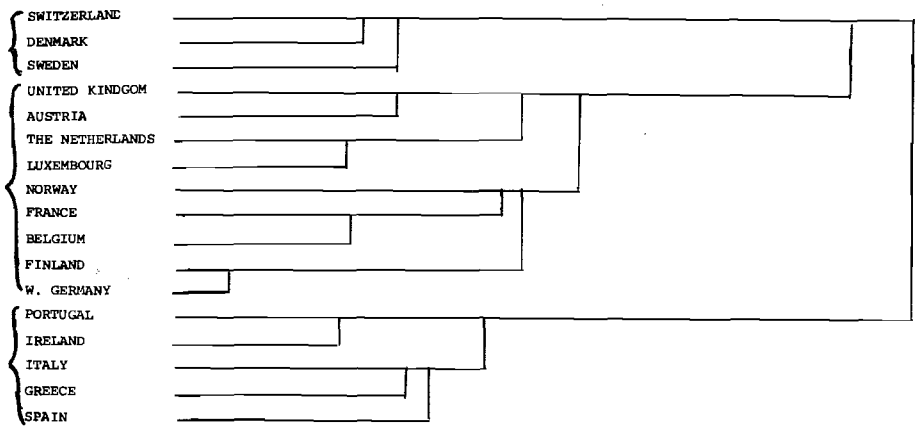


Figure 4. A tree diagram of nations

4. CONCLUSION

The techniques described above are, of course, only exploratory. Nevertheless the use of additional variables and additional cases may help us to test hypotheses in the following way. Suppose we wonder if a particular group of individuals have a special behaviour which differs from the average; if that group is defined by a category of a variable which was not part of the analysis it is possible to test if the mean value of that group for the various principal components differs significantly from zero. The use of additional variables is a major practice in the screening of opinion surveys where one splits the variables into two groups. The first one, generally the sociological and cultural variables, is processed by multiple correspondence analysis, the second one with the opinion variables is projected afterwards upon the principal axes allowing a form of categorical regression (Lebart et al., 1977).

Are correspondence analysis and multiple correspondence analysis really multivariate techniques? At first glance they are, and from two different points of view. The first one is that we use several dimensions, i.e., several principal components for describing the data, and the second is that data are multidimensional (for multiple correspondence analysis).

Of course one may question the advantage of taking more than one solution to the eigenequations (5) or (14), for the statistical meaning of suboptimal solutions is not obvious. Our answer is that all these analyses try, as a matter of fact, to reconstruct distances between objects (rows or columns) and that this is not possible with a single dimension.

However, even for multiple correspondence analysis the true multivariate nature of data is not used but only its bidimensional facets. As Gifi (1981, p. 50) points out these techniques are essentially bivariate for they "give the same results if we apply them to another multivariate distribution with the same bivariate marginals", that is to say, in our context of categorical variables with the same 2x2 contingency tables. This is clear in equation (18). This is certainly a limitation but the counterpart is that these techniques may process a very large number of variables.

Finally, though these techniques may be applied without difficulty to spatial data, there is a need for further extensions. When individuals are spatial units, the assumption of independence between them is certainly inadequate and the phenomenon of spatial correlation due to continguity must be taken into account.

REFERENCES

Benzécri, J.P., 1973, L'analyse des données vol. 1 Taxonomie (Dunod, Paris).

Benzécri, J.P., 1973, L'analyse des données vol II Correspondances (Dunod, Paris).

Bouroche, J.M. and G. Saporta, 1980, L'analyse des données, Collection Que sais-je no. 1854 (Presses Universitaires de France, Paris).

Deville, J.C. and G. Saporta, 1983, Correspondence analysis with an extension towards nominal time series, Journal of Econometrics 22, 152-164.

De Leeuw, J. and J. van Rijckevorsel, 1979, Homals and Princals, in E. Diday (ed.), Data analysis and Informatics (North-Holland, Amsterdam) 231-242.

Gifi, A., 1981, Non-linear Multivariate Analysis (Department of Data Theory, Leyden University, Leyden).

Gnanadesikan, R., 1977, Methods for Statistical Data Analysis of Multivariate Observations (John Wiley, New Yrok).

Guttman, L., 1941, The quantification of a class of attributes: a theory and method of scale construction, in: P. Horst (ed.), The prediction of personal adjustment (Social Science Research Council, New York) 319-348.

Hayashi, C., 1950, On the quantification of qualitative data from the mathematico-statistical point of view, Annals of the Institute of Mathematical Statistics 2, 35-47.

Hill, M.O., 1974, Correspondence analysis: a neglected multivariate method, Applied Statistics 23, 340-354.

Hirschfeld, H.O., 1935, A connection between correlation and contingency. Proceedings of the Cambridge Philosophical Society 31, 520-524.

Jambu, M., 1978, Classification automatique pour l'analyse des donées (Dunod, Paris).

Kendall, M.G., and A. Stuart, 1961, The advanced theory of statistics Vol. 2: Inference and relationship (Griffin, London).

Lebart, L., A. Morineau, and N. Tabard, 1977, Techniques de la description statistique (Dunod, Paris).

MacKeon, J.J., 1966, Canonical analysis: some relations between canonical correlation, factor analysis, discriminant functions analysis and scaling theory. Psychometric monograph no. 13 (The Psychometric Society, New York).

Masson, M., 1974, Processus lineaires et analyse de données non linéaire, Thèse (Université Pierre et Marie Curie, Paris).

Tukey, J., 1977, Exploratory data analysis (Addison Wesley, Reading).