Journal of Econometrics 22 (1983) 169–189. North-Holland Publishing Company

CORRESPONDENCE ANALYSIS, WITH AN EXTENSION TOWARDS NOMINAL TIME SERIES

J.-C. DEVILLE

Institut National de la Statistique et des Etudes Economique, 75675 Paris, France

G. SAPORTA

Université René Descartes, 75016 Paris, France

Correspondence analysis is a technique for studying the relationship between two nominal variables which uses mainly simultaneous graphical displays. It has been generalized to more than two variables under the name of 'multiple correspondence analysis'. 'Qualitative harmonic analysis' is an other extension towards individual time-series where one observes the evolution of a nominal variable through a finite period of time. The present paper is based essentially on the concept of multidimensional scaling by means of barycentric representation.

1. Introduction

Correspondence analysis is perhaps the most popular method of multidimensional data analysis in France: for J.P. Benzécri and his school, correspondence analysis has even become almost synonymous with data analysis.

The reasons of such a success are multiple and are due mainly to the suggestive power of the graphical displays. A whole set of interpretation rules using specific measures such as the contributions and the practice of additional points has been developed. Moreover, correspondence analysis proved to be a robust method as its results are remarkably stable if the data are perturbated [Lebart et al. (1977)].

Correspondence analysis handles only categorical data (numerical variables have to be discretized). Its initial field of application is that of contingency tables where data are cross-classified according to two categorical variables. This is the topic of section 2. Although the principles are not new [Hirschfeld (1935), Fisher (1940)] as, mathematically, correspondence analysis is identical to canonical analysis of contingency tables [Kendall and Stuart (1961, pp. 568–574)], its rediscovery as a method of exploratory data analysis is recent (J.P. Benzécri in the sixties). The fundamental point is the use of the various canonical variables — not only of the first one — to obtain

0165–7410/83/$03.00 © Elsevier Science Publishers

graphical displays for the categories of the variables or, in other words, of the rows and columns of a contingency table in a common space of low dimension.

Instead of the classical derivations of the correspondence analysis as a canonical correlation analysis or as two simultaneous principal components analyses of rows and columns proportions, we propose here a presentation based upon the graphical displays of categories over a 'good' system of axes. This presentation has the advantage of facilitating, in section 3, the presentation of multiple correspondence analysis, that is, the case where more than two categorical variables have been observed for a set of n individuals. Once more we arrive at known results: the equations are that of the principal components of scale [Guttman (1941)] but the use is different and is similar to that of HOMALS [see Gifi (1981)]. Finally, in section 4, we present a continuous extension of multiple correspondence analysis to an infinite number of categorical variables associated to the different states of a set of individuals through the time [Deville and Saporta (1979)].

2. Correspondence analysis of a contingency table

Let N be a contingency table with m_1 rows and m_2 columns whose elements are n_{kl}; χ_1 and χ_2 will be the two related nominal variables with m_1 and m_2 categories, and n is the total number of individuals. We suppose $m_1 \leqq m_2$. Let X_1 and X_2 denote the matrices of size (n, m_1) and (n, m_2), associated with χ_1 and χ_2 respectively, of the indicator variables of their categories,

$$\begin{aligned} X_1(i,k) &= 1 \quad \text{if individual } i \text{ belongs to category } k \text{ of } \chi_1, \\ &= 0 \quad \text{if not.} \end{aligned}$$

So, $N = X_1'X_2$. Denote by D_1 and D_2 the diagonal matrices of marginal frequencies of χ_1 and χ_2,

$$D_1 = X_1'X_1, \qquad D_2 = X_2'X_2. \tag{1}$$

$D_1^{-1}N$ and ND_2^{-1} are therefore the two arrays of conditional frequencies whose elements are $n_{kl}/n_{k.}$ and $n_{kl}/n_{.l}$, respectively, where a dot denotes summation over an index.

2.1. Simultaneous representation of the categories of χ_1 and χ_2

The categories of χ_1 and χ_2 define subgroups of size $n_{k.}$ or $n_{.l}$ of the population. If, by hypothesis, we know a numerical variable z measured for each of the n individuals, we are able to compute the average value of z for each subgroup; the different means for the m_1 categories of χ_1 can be arranged as an m_1-vector $\boldsymbol{a}_1$, such that

$$\boldsymbol{a}_1 = (X_1'X_1)^{-1}X_1'z = D_1^{-1}X_1'z, \tag{2}$$

and for χ_2,

$$\boldsymbol{a}_2 = D_2^{-1}X_2'z.$$

It is then easy to display simultaneously the (m_1+m_2) categories of χ_1 and χ_2 on an axis according to the category means. If z has zero mean, the dispersion of the categories of χ_1 along this axis (i.e., the variance due to χ_1) is

$$z'X_1D_1^{-1}X_1'z = \boldsymbol{a}_1'D_1\boldsymbol{a}_1 = z'A_1z, \tag{3}$$

where $A_1 = X_1(X_1'X_1)^{-1}X_1'$ is the orthogonal projector (regression operator) onto the subspace spanned by the columns of X_1. The greater this quantity is, the better the categories χ_1 will be separated on the axis associated with z. Of course the correlation ratio $\eta^2(z\,|\,\chi_1) = \boldsymbol{a}_1'D_1\boldsymbol{a}_1/z'z$ is maximum and equal to 1 if $z = X_1\boldsymbol{a}_1$, that is to say if all the individuals of each category of χ_1 have the same value of z. Such a variable z would be optimal for χ_1.

Actually we do not know such a variable z and we just have χ_1 and χ_2. We will now try to find an artificial variable z that is optimal for both χ_1 and χ_2. Since it is generally impossible to have simultaneously $\eta^2(z\,|\,\chi_1)=1$ and $\eta^2(z\,|\,\chi_2)=1$ we will look for a variable z with fixed variance such that *on average* the variance of z due to χ_1 and χ_2 be maximum. So the problem is to maximize

$$\tfrac{1}{2}(z'A_1z + z'A_2z) \quad \text{with} \quad z'z = c,$$

the solution of which is to take for z the eigenvector of $\frac{1}{2}(A_1+A_2)$ associated with its greatest non trivial eigenvalue $\mu=\mu_1$. This artificial variable provides the 'best' representation of the categories of χ_1 and χ_2 along a unique axis and the coordinates of the categories χ_1 and χ_2 are given by the components of $\boldsymbol{a}_1 = D_1^{-1}X_1'z$ and $\boldsymbol{a}_2 = D_2^{-1}X_2'z$.

Notice that, from $\frac{1}{2}(A_1+A_2)z = \mu_1 z$,

$$\mu z = \tfrac{1}{2}(X_1D_1^{-1}X_1' + X_2D_2^{-1}X_2')z \tag{4}$$

implies that $\mu z = \frac{1}{2}(X_1\boldsymbol{a}_1 + X_2\boldsymbol{a}_2)$. The variable z takes the same value for all the n_{kl} individuals of the cell (k,l) of N. Since $\mu_0=1$ is always the greatest eigenvalue of $\frac{1}{2}(A_1+A_2)$ associated with a constant z, this trivial solution must however be eliminated.

Suppose now that we want to display the categories of χ_1 and χ_2 in a plane and not just along a single axis (for there is no reason why the phenomenon might be represented by a single dimension). We then need a second variable z uncorrelated with the first one, like in principal components analysis, providing an other extremum of $\frac{1}{2}z'(\dot{A}+A_2)z$. The solution is then provided by the second non-trivial eigenvector of $\frac{1}{2}(A_1+A_2)$ associated with the second largest eigenvalue μ_2, and so on.

Since z is an n-vector the solution of the preceding eigenequation is cumbersome and it is simpler to obtain the coordinates of the categories directly along the axes. This is simply done by substituting $\boldsymbol{a}_1=D_1^{-1}X_1'z$ and $\boldsymbol{a}_2=D_2^{-1}X_2'z$ into $\mu z=\frac{1}{2}(X_1\boldsymbol{a}_1+X_2\boldsymbol{a}_2)$. As $X_1'X_2=N$, we have

$$(2\mu-1)\boldsymbol{a}_1=D_1^{-1}N\boldsymbol{a}_2 \quad \text{and} \quad (2\mu-1)\boldsymbol{a}_2=D_2^{-1}N'\boldsymbol{a}_1. \tag{5}$$

Hence, by substituting again,

$$(2\mu-1)^2\boldsymbol{a}_1=D_1^{-1}ND_2^{-1}N'\boldsymbol{a}_1. \tag{6}$$

As $D_1^{-1}N$ and $D_2^{-1}N'$ are the two arrays of conditional frequencies we now have established the following property:

Property 1. The vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ of coordinates of the categories of variables are eigenvectors of the product of the two matrices obtained by multiplication of the arrays of conditional frequencies.

If we define $\lambda=(2\mu-1)^2$, we have $0\leqq\lambda\leqq1$ since $0\leqq\mu\leqq1$. Omitting the trivial eigenvalue $\lambda_0=1$, there exist at most m_1-1 non-trivial eigenvalues (as $m_1\leqq m_2$). Knowing $\boldsymbol{a}_1$ we simply get $\boldsymbol{a}_2$ by

$$\boldsymbol{a}_2=\lambda^{-\frac{1}{2}}D_2^{-1}N'\boldsymbol{a}_1. \tag{7}$$

Usually $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are normalized as follows (see subsection 2.2):

$$(1/n)\boldsymbol{a}_1'D_1\boldsymbol{a}_1=(1/n)\boldsymbol{a}_2'D_2\boldsymbol{a}_2=\lambda. \tag{8}$$

Since $\lambda_0=1$, and

$$1+\sum_{i=1}^{m-1}\lambda_i=\operatorname{tr} D_1^{-1}ND_2^{-1}N'=\sum_k\sum_l(n_{kl}^2/n_{k.}n_{.l}), \tag{9}$$

we have:

Property 2. The sum of the non-trivial eigenvalues is equal to Pearson's ϕ^2 of dependence, that is to say the usual chi-square divided by n.

We also have [see Kendall and Stuart (1967, p. 574)]:

Property 3. Reconstruction of N:

$$n_{kl}=(n_{k.}n_{.l}/n)\left[1+\sum_{i=1}^{m_1-1}\lambda^{-\frac{1}{2}}a_{1k}^{(i)}a_{2l}^{(i)}\right]. \tag{10}$$

These two properties illustrate the fact that correspondence analysis is an analysis of the departure from independence in a contingency table.

2.2. How this presentation can be linked to other ones

Correspondence analysis may be considered as a special case of canonical correlation analysis between the two sets of indicator variables associated with the categories χ_1 and χ_2 [Cailliez and Pagès (1976)]. It is also a method of assigning scores to the categories (the scores are the coordinates derived in the preceding subsection), such that the following holds:

— the bivariate distribution of $(\xi_1=X_1\boldsymbol{a}_1, \xi_2=X_2\boldsymbol{a}_2)$ has both regressions linear [Hirschfeld (1935)], that is to say: the conditional means of ξ_1 given ξ_2 (or χ_2) are linear functions of ξ_2 and vice versa;

— the linear correlation coefficient between ξ_1 and ξ_2 is maximal (and equal to $\lambda^{\frac{1}{2}}$) [Lancaster (1957), Williams (1952), Kendall and Stuart (1967)];

— the discrimination of the categories of χ_1 by means of ξ_2 is maximal [Fisher (1940)].

The most popular presentation in France, used by J.P. Benzécri and his team, is related with principal components analysis. Consider the array $D_1^{-1}N$ as a data matrix of m_1 'individuals' (the rows) described by m_2 variables, the row frequencies $n_{kl}/n_{k.}$ $(l=1,2,\dots,m_2)$ being the variable values. The 'individuals' here are weighted according to the matrix $(1/n)D_1$ of the row-marginal frequencies. The distance between rows is the so-called chi-square distance [Guttman (1941)],

$$d^2(k,h)=\sum_{l=1}^{m}(n/n_{.l})(n_{kl}/n_{k.}-n_{hl}/n_{h.})^2, \tag{11}$$

i.e., the metric in R^{m_2} is nD_2^{-1}. We therefore obtain directly the coordinates of the rows, $\boldsymbol{a}_1$, on the principal axis by solving the following equation [principal coordinate analysis, see Mardia et al. (1979)]:

$$(D_1^{-1}N)(nD_2^{-1})(D_1^{-1}N)'\left(\frac{1}{n}D_1\right)\boldsymbol{a}_{1.}=\boldsymbol{a}_1, \tag{12}$$

where the four matrices at the left-hand side are the data matrix, the metric, the transposed data matrix and the weights, respectively. This reduces to $D_1^{-1}ND_2^{-1}N'\boldsymbol{a}_1=\lambda\boldsymbol{a}_1$.

Conversely, the coordinates of the columns are obtained by submitting to a principal coordinate analysis the data matrix ND_2^{-1} with the weights $(1/n)D_2$ and the metric $D_1^{-1}n$; the eigenequation then is $D_2^{-1}N'D_1^{-1}N\boldsymbol{a}_2=\lambda\boldsymbol{a}_2$. There is a duality between these two principal components analyses, but here the simultaneous representation is merely a device and has no strong theoretical background because the categories of χ_1 and χ_2 belong to two different vectorspaces.

As the eigenvalues λ are the variances of the principal components, the natural normalization of the coordinates is

$$(1/n)\boldsymbol{a}_1'D_1\boldsymbol{a}_1=(1/n)\boldsymbol{a}_2'D_2\boldsymbol{a}_2=\lambda,$$

but this holds only for the analysis of a contingency table and not for multiple correspondence analysis (see section 3).

2.3. *Use and interpretation of correspondence analysis*

Let us take the following demographic example as an illustration. N is the table (omitted here) giving the distribution of the population of twelve European countries (West Germany, France, Italy, The Netherlands, Belgium, Luxemburg, Great Britain, Ireland, Denmark, Greece, Portugal, Spain) over sixteen age groups (0–4, 5–9,..., 75 and over) in 1979. Of the eleven non-trivial eigenvalues the first three explain 90% of the ϕ^2: 61.4%, 17.4% and 10.7%, respectively. The simultaneous representation of the first two dimensions is given in fig. 1.

Table 1 helps us.to know which categories have contributed mainly to the determination of the axes. As

$$\lambda=\sum_k(n_{k.}/n)(a_{1k})^2=\sum_l(n_{.l}/n)(a_{2l})^2, \tag{13}$$

the contribution of category k to λ is defined by

$$(n_{k.}/n)(a_{1k})^2/\lambda, \tag{14}$$

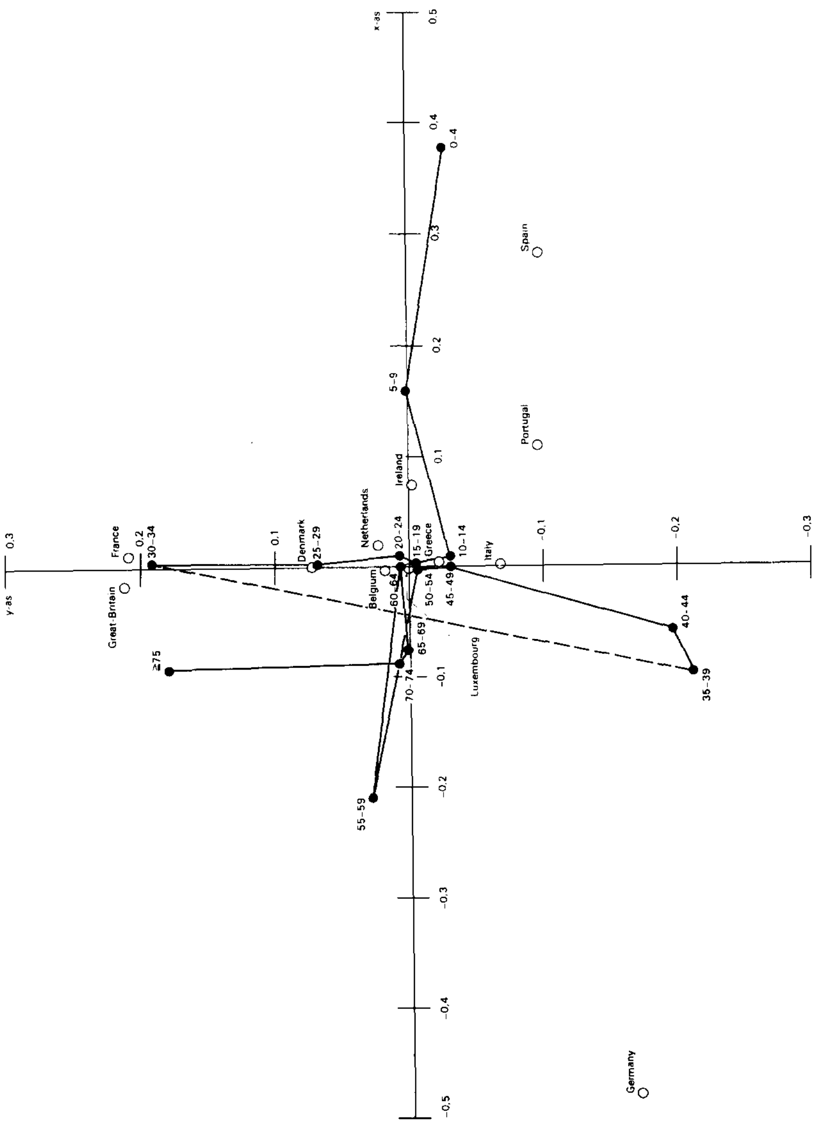
x-as
y-as
0.5
0.4
0.3
0.2
0.1
-0.1
-0.2
-0.3
-0.4
-0.5
0.3
0.2
0.1
-0.1
-0.2
-0.3
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
≥75
Spain
Portugal
Ireland
Netherlands
Denmark
France
Great-Britain
Belgium
Greece
Italy
Luxembourg
Germany

Fig. 1

Table 1

Contributions to the eigenvalues.

	Eigenvalues			Marginal frequencies
	λ_1 (61.4%)	λ_2 (17.4%)	λ_3 (10.7%)	
Countries				
West Germany	(−)0.478	(−)0.170	(−)0.011	0.1939
France	0.012	0.211	(−)0.456	0.1688
Italy	0.001	(−)0.066	0.000	0.1797
The Netherlands	0.019	0.023	(−)0.005	0.0442
Belgium	(−)0.003	0.019	(−)0.016	0.0311
Luxemburg	(−)0.001	0.000	(−)0.001	0.0011
Great Britain	(−)0.016	0.213	0.460	0.1769
Ireland	0.072	(−)0.001	0.007	0.0106
Denmark	0.000	0.073	0.028	0.0159
Greece	0.004	(−)0.021	0.001	0.0298
Portugal	0.110	(−)0.096	0.013	0.0310
Spain	0.284	(−)0.098	0.002	0.1169
	1.	1.	1.	1.
Age				
0–4	0.377	(−)0.026	(−)0.000	0.0664
5–9	0.160	0.001	0.001	0.0758
10–14	0.008	(−)0.030	0.022	0.0832
15–19	0.003	(−)0.004	0.000	0.0811
20–24	0.010	0.008	(−)0.043	0.0747
25–29	0.001	0.069	(−)0.148	0.0719
30–34	0.004	0.192	0.000	0.0690
35–39	(−)0.096	(−)0.212	(−)0.000	0.0626
40–44	(−)0.058	(−)0.195	(−)0.017	0.0630
45–49	0.000	(−)0.030	(−)0.038	0.0609
50–54	(−)0.003	(−)0.005	(−)0.007	0.0594
55–59	(−)0.021	0.029	0.050	0.0555
60–64	0.000	0.007	0.616	0.0410
65–69	(−)0.076	0.001	0.024	0.0470
70–74	(−)0.088	0.008	0.002	0.0385
75 and over	(−)0.093	0.182	(−)0.032	0.0499
	1.	1.	1.	1.

which may be compared for instance to the importance of category k in the population, $n_{k.}/n$. It is also useful to write the contribution with the sign of the coordinate as some contributions may be of opposite meaning.

We can see that the first dimension reveals a strong difference between West Germany (on the left of the figure) and Ireland, Spain and Portugal (on the right), which is mainly due to the difference between the age groups over 65 on the one side and the age groups under 10 on the other. This dimension may be identified as a fertility one which is closely related with birth-rate.

On the vertical axis, France, Great Britain and Denmark stand apart from the other countries and this phenomenon is due to the fact that the 35–39 and 40–44 age groups are less numerous (in percentages) in these three countries than in the others. One may observe too that the 30–34 age group has a great positive contribution to this dimension; the second dimension may be explained by the diminution of the birth-rate during the years 1935–1939 combined with the baby-boom after the Second World War.

The third dimension reveals a difference between France and Great Britain due to the 60–64 age group, which may be explained by a gap in the birth-rate in France during the First World War, which did not occur in Great Britain.

Of course, much more could be said; for instance about the large contribution of the '75 and over' age group to the second eigenvalue, but we only intended to show that the essential information about the structure of a contingency table can be given very quickly by correspondence analysis.

3. Multiple correspondence analysis

3.1. *Extension to p variables*

Suppose now that the n individuals are described by p nominal variables with $m_1, m_2, \ldots, m_p$ categories. Let X denote the supermatrix $(X_1, X_2, \ldots, X_p)$ of all indicator variables and $D = \operatorname{diag}(D_1, \ldots, D_p)$ the diagonal supermatrix of all marginal frequencies. Let z be again a zero-mean numerical variable providing an unidimensional scale for the n individuals.

The criterium presented in section 1 may be generalized as follows. The best representation of categories of all variables will be obtained if

$$(1/p) \sum_{j=1}^{p} \boldsymbol{a}_j' D_j \boldsymbol{a}_j \quad \text{is maximum,} \tag{15}$$

in other words, if the p possible analyses of variance are maximized on average. We have obviously the following results: z is the eigenvector of $(1/p)\sum A_j$ associated with the greatest non-trivial eigenvalue μ and

$$z = (1/p\mu) \sum_{j=1}^{p} X_j \boldsymbol{a}_j. \tag{16}$$

A multidimensional configuration is then obtained with the other eigenvectors. The total number of non-trivial eigenvalues is $\sum m_j - p$, as we have $p-1$ linear relationships between the indicator variables.

Unlike $p=2$, it is not possible to find an eigenequation for each $\boldsymbol{a}_j$ but only for the whole set of the $\boldsymbol{a}_j$, that is to say, for the supervector $\boldsymbol{a} = (\boldsymbol{a}_1', \ldots, \boldsymbol{a}_p')'$ with $\sum m_j$ components: substituting (16) in the equation $(1/n)\sum A_j z = \mu z$ gives

$$\sum_j X_j D_j^{-1} X_j' \left(\sum_k X_k \boldsymbol{a}_k \right) = \mu p \sum_k X_k \boldsymbol{a}_k. \tag{17}$$

The preceding equation is the expansion in a partitioned form of $(1/p)XD^{-1}X'X\boldsymbol{a}=\mu X\boldsymbol{a}$, as $X=(X_1,X_2,\dots,X_p)$; hence,

$$(1/p)D^{-1}X'X\boldsymbol{a}=\mu\boldsymbol{a}, \tag{18}$$

where $X'X$ is the super-array of all 2×2-contingency tables of the χ_j's. The $\boldsymbol{a}_j$ are normalized by

$$(1/np)\boldsymbol{a}'D\boldsymbol{a}=(1/np)\sum_j \boldsymbol{a}_j'D_j\boldsymbol{a}_j=\mu. \tag{19}$$

We notice that except for $p=2$ the $\boldsymbol{a}_j'D_j\boldsymbol{a}_j$ do not necessarily have the same value. As

$$(1/p)\sum A_j=(1/p)\sum X_jD_j^{-1}X_j'=(1/p)XD^{-1}X', \tag{20}$$

the sum of each row of X equals p, and terms of D are the sums of the columns of X. Property 1, given in subsection 2.1, applies here and we have the fundamental result:

Property 4. Multiple correspondence analysis of $\chi_1,\chi_2,\dots,\chi_p$ is identical to the formal correspondence analysis of the disjunctive array X considered as a contingency table.

It must be pointed out here that the sum of non-trivial eigenvalues equals $(1/p)\sum(m_j-1)$ and has no statistical importance: not much meaning can be attached to the percentage of explained variance, since, as a matter of fact, $\mu\leqq 1$: if for instance, the average number of categories is 5 for the p variables, $\mu_1\leqq 25\%$. The interpretation of the axes will be based essentially upon the contributions of the categories and the variables.

Finally, notice that the same coordinates of all categories (apart from the multiplicative constant $\mu^{\frac{1}{2}}$) may be obtained by performing a correspondence analysis of $X'X$. In this case the eigenvalues are μ^2 instead of μ, and we have

$$\sum\mu^2=(1/p^2)\sum\sum\phi_{kl}^2, \tag{21}$$

the average value of all K. Pearson's ϕ^2 between χ_k and χ_l for $k=1,2,\dots,p$ and $l=1,2,\dots,p$. This clearly shows that multiple correspondence analysis is a method for studying the relationships between p variables χ_j using only the two-by-two-dependencies.

3.2. Some other enlightening presentations

The preceding solution may be obtained in various ways. The first one is related with Guttman's principal components of scale, or homogeneity analysis. One of the main ideas of factor analysis is that different variables may measure the 'same thing' and can thus be represented by a unique scale. When the variables are nominal, Guttman (1941) proposed the following quantification technique: assign a numerical value to each category of the χ_j such that the scores of the individuals be as homogeneous as possible for the p variables and as different as possible between individuals.

Let ξ_j (vector of length n) be a zero-mean quantification of χ_j: $\xi_j = X_j \boldsymbol{a}_j$, where $\boldsymbol{a}_j$ is the m_j-vector *of scores for* χ_j. Let $\xi_{.}$ be the vector of average scores, which will be taken as the unique desired scale,

$$\xi_{.} = (1/p) \sum_{j=1}^{p} \xi_j = (1/p) \sum_{j=1}^{p} X_j \boldsymbol{a}_j = (1/p) X \boldsymbol{a}.$$

For each individual i, the heterogeneity of his scores is measured by

$$(1/p) \sum_j (\xi_{ij} - \xi_{i.})^2, \tag{22}$$

which, averaged over all individuals, is equal to

$$(1/np) \sum_i \sum_j (\xi_{ij} - \xi_{i.})^2. \tag{23}$$

The problem is to minimize the latter quantity. Since we have the classical analysis of variance equation,

$$(1/n) \sum_i \xi_{i.}^2 + (1/np) \sum_i \sum_j (\xi_{ij} - \xi_{i.})^2 = (1/np) \sum_i \sum_j \xi_{ij}^2,$$

an equivalent problem is to maximize the correlation ratio

$$(1/n) \sum_i \xi_{i.}^2 \Big/ (1/np) \sum_i \sum_j \xi_{ij}^2 = (1/np^2) \boldsymbol{a}' X' X \boldsymbol{a} \Big/ (1/np) \sum_j \boldsymbol{a}_j' D_j \boldsymbol{a}_j$$

$$= (1/p)(\boldsymbol{a}' X' X \boldsymbol{a} / \boldsymbol{a}' D \boldsymbol{a}). \tag{24}$$

The maximum is attained if $\boldsymbol{a}$ is the eigenvector of $(1/p) D^{-1} X' X$ corresponding to its maximal eigenvalue and we find again the first dimension of multiple correspondence analysis.

A second and closely connected way of obtaining the multiple correspondence analysis is based upon principal components analysis of nominal variables. It is well known that the first principal component z of a set of p numerical variables provides the best unidimensional configuration of n individuals in the sense of $\operatorname{var}(z)$ being maximal. Hence, the idea of quantifying the x_j's into $\xi_j = X_j \boldsymbol{a}_j$ such that the first principal component of the ξ_j's be of maximal variance. In other words, we look for $\boldsymbol{a}_j$'s such that the first eigenvalue μ_1 of the correlation matrix of the ξ_j can be maximized.

The result is that the $\boldsymbol{a}_j$ are the coordinates of the categories of χ_j on the first axis of multiple correspondence analysis. For any set of known ξ_j their first principal component z maximizes $\sum r^2(z; \xi_j)$. When looking for the best set of ξ_j, we have to maximize that quantity over ξ_j and z. Since $\xi_j = X_j \boldsymbol{a}_j$, the problem may be formulated as

$$\max_{\boldsymbol{a}_j} \max_{z} \sum_{j} r^2(z; X_j \boldsymbol{a}_j).$$

But the maximum over the $\boldsymbol{a}_j$ of $r^2(z; X_j \boldsymbol{a}_j)$ for fixed z is equal to the squared multiple correlation coefficient between z and the columns of X_j, $R^2(z; X_j)$, which is equal to the correlation ratio $\eta^2(z \mid \chi_j) = z' A_j z$, since X_j is the indicator matrix of χ_j. Thus the problem reduces to

$$\max_{z} \sum_{j} z' A_j z \quad \text{subject to } z'z = 1, \tag{25}$$

and the solution is given by the first eigenvector of $\sum_j A_j$. Since the two criteria of optimality,

$$\max \sum_{j} r^2(z; \xi_j) \quad \text{for numerical variables,}$$

$$\max \sum_{j} \eta^2(z \mid \chi_j) \quad \text{for nominal variables,} \tag{26}$$

are similar, we may consider multiple correspondence analysis equivalent to principal components analysis for nominal data, and since $\eta^2(z \mid \chi_j) = R^2(z; X_j)$, as a special case of Carroll's (1968) generalized canonical analysis.

The method of reciprocal averaging is another approach very classical in psychometric literature and is one of the easiest way of presenting multidimensional scaling of nominal variables [Kuder and Richardson (1933), Nishisato (1980)]. It consists of a simultaneous representation of individuals and of categories of the nominal variables such that:

(1) the coordinate of a category be equal to the average coordinate of individuals who belong to that category; and
(2) the coordinate of an individual be equal to the average values of the coordinates of the categories to which he belongs.

Starting from an arbitrary set of values for the individuals, say z, we may obtain by an iterative process the coordinates of the categories and then a new variable z, and so on. Provided we impose a normalization constraint, the algorithm converges very quickly and is the basis of some alternating least squares quantification techniques such as HOMALS [see Gifi (1981)]. Actually the two conditions mentioned above cannot both be satisfied as $\boldsymbol{a}_j = D^{-1}X'_j z$ and $z = (1/p)\sum X_j \boldsymbol{a}_j$ do not hold simultaneously; we need a constant α as small as possible (since it may be shown that $\alpha \geqq 1$), such that

$$\boldsymbol{a} = \alpha D^{-1}X'z \quad \text{and} \quad z = (1/p)\alpha X\boldsymbol{a}. \tag{27}$$

By substituting we find

$$(1/\alpha^2)z = (1/p)XD^{-1}X'z \quad \text{and} \quad (1/\alpha^2)\boldsymbol{a} = (1/p)D^{-1}X'X\boldsymbol{a}, \tag{28}$$

and we have again the first solution of multiple correspondence analysis, since $\mu_1 = 1/\alpha^2$ must be maximized:

3.3. Use and interpretation of multiple correspondence analysis

This technique is widely used for the screening of surveys. The graphical representation of all response categories allows for a very fast detection of the more interesting relationships and directs the researcher towards the more interesting cross-tabulations.

A major practice is to use additional variables. Usually the set of variables is split up into two groups: the working variables, with which the axes are computed, and the passive ones, which may be easily represented on the system of axes as usual. The categories of the passive variables are represented through the z variables by the mean-value of individuals which belong to them. If the additional variables are numerical, we can compute of course the product–moment correlation coefficient with the z variables.

The use of additional variables may serve two different goals. First it provides a quick approximate regression in the sense that we actually project the passive variables upon the subspace spanned by the z variables. We may interpret this as an 'explanation' of some dependent variables by nominal predictors. For instance [Bouroche and Saporta (1980)], a sample of 6,083 individuals are described by a set of twelve sociological and cultural variables with a multiple correspondence analysis. Then the subgroup of individuals who have seen a certain movie may be projected upon the system of the first axes, which allows us to determine quickly the average pattern of these individuals.

The second feature is concerned with the validation of the results. Interpreting the outcome of an analysis using the working variables may be subject to criticism: maybe the results are nothing but an artefact due to the mathematical technique that has been used. If, however, the meaning of an axis is obtained by a correlation with a variable that has not contributed to its determination the interpretation will be more convincing; moreover some approximate statistical significance tests may be performed such as a one-way analysis of variance to test equality of means of an additional categorical variable over an axis.

4. Individual time-series: Qualitative harmonic analysis

Correspondence analysis can also be extended to qualitative data varying over a finite time interval T. For convenience, we will assume that T is $[0,1]$ and that all the time-data are expressed within this interval by means of a linear transformation. The category space is finite with elements k numbered from 1 to m. For every individual i $(i=1,\dots,n)$, the data consist of the records of the successive states in which he has been, with the exact dates when he has moved from one state to another. We are able, therefore, to compute for each time t the matrix X_t, with n rows and m columns, indicating the state k to which the individual i belongs at time t. In the same way as correspondence analysis dealt with two nominal variables, and multiple correspondence analysis dealt with p variables, we now have to deal with a continuous infinity of nominal variables indexed by time.

4.1. The equations

We try, once more, to define an artificial variable z, independent of time which describes the trajectories of each individual over time 'as well as possible'. Formally, z is an n-vector having one coordinate per individual. At time t, we can compute the mean of the coordinates of z corresponding to individuals being in state k. Those means can be arranged in the m-vector $\boldsymbol{a}_t$ given by

$$\boldsymbol{a}_t=(X_t'X_t)^{-1}X_t'z. \tag{29}$$

It is well-known that this vector minimizes the sum of the squres of the coordinates of $z-X_t\boldsymbol{a}_t$ and that we have the identity (with the notation $\|x\|^2=x'x$ for every n-vector x)

$$\|z\|^2=\|X_t\boldsymbol{a}_t\|^2+\|z-X_t\boldsymbol{a}_t\|^2. \tag{30}$$

Integrating over T, and considering the second term at the right-hand side as a residual, we get, using (29),

$$z'z=\int_0^1 z'X_t(X_t'X_t)^{-1}X_t'z\,\mathrm{d}t+\text{mean of residuals}. \tag{31}$$

The variable z we want to compute maximizes the quantity

$$z'Qz/z'z, \tag{32}$$

with

$$Q=\int_0^1 X_t(X_t'X_t)^{-1}X_t'\,\mathrm{d}t, \tag{33}$$

the mean over time of the projection operator onto the subspace generated by the columns of X_t.

The matrix Q is symmetric positive definite and its elements are easy to compute. If individuals i and j are not in the same state at time t, the corresponding element in $X_t(X_t'X_t)^{-1}X_t'$ is 0; if they are in the same state at time t, say k, its value is $1/n_k^t$, with n_k^t denoting the number of individuals who are in state k at time t. The matrix Q appears to be a measure of similarity between the individuals, integrating elementary similarities indexed by the time.

The problem to solve remains formally the same as in the previous sections and its solution is given by the eigenvector of Q associated with the largest non-trivial eigenvalue. Notice that the vector $\boldsymbol{l}$ with all its coordinates equal to 1, is an eigenvector of Q associated with eigenvalue 1. This solution has no statistical interest; it shows, however, that all its other solutions satisfy $\boldsymbol{l}'z=0$ and that we could, without any loss of generality, impose on z the restriction of zero mean.

The relations between z and the vector valued function $\boldsymbol{a}_t$ are of interest. Starting from the eigenvalue equation $\lambda z=Qz$, we get by straightforward calculation

$$\lambda \boldsymbol{a}_s=\int_0^1 (X_s'X_s)^{-1}X_s'X_t\boldsymbol{a}_t\,\mathrm{d}t. \tag{34}$$

The matrix $N_{st}=X_s'X_t$ is the $m\times m$ array of numbers $n_{kl'}^{st}$ of those individuals who were in state k at time s and in state l at time t. Of course, N_{tt} is the diagonal matrix of the number n_k^t of individuals who were in state k at time t. The matrix $(X_s'X_s)^{-1}X_s'X_t$ is the array of conditional frequencies with entries p_{st}^{kl}, the probability of being in state l at time t, knowing that one was in state k at time s. Therefore, the $\boldsymbol{a}_t$ function is the solution of the equations

$$\lambda \boldsymbol{a}_s^k=\sum_{l=1}^{m}\int_0^1 p_{st}^{kl}\boldsymbol{a}_t^l\,\mathrm{d}t. \tag{35}$$

It is formally equivalent to solve this set of equations or to solve $\lambda z=Qz$, but, curiously, the former turns out to be easier from a computational point of view.

4.2. Practical solution: Approximation procedure

The matrix Q is generally too large to be handled numerically. In practice we will try to compute an approximation of the $\boldsymbol{a}_t$-function in a prescribed form. Here, we will outline an approach that satisfies the usual practical needs, although it is not the most general one. We look for a solution in the form

$$\boldsymbol{a}_t = \sum_{t=1}^{L} \boldsymbol{\alpha}_l f_l(t), \tag{36}$$

where $\boldsymbol{\alpha}_l$ is a family of unknown m-vectors, and $f_l(t)$ a family of known numerical functions, chosen to be easy to handle. As a matter of fact, we replace the set of all m-vector valued functions on T by a finite-dimensional subspace generated by (36). The functions $f_l(t)$ must of course be linearly independent. It is also convenient that the subspace generated by the f_l contains the constant function. The reason is that the vector function having all its coordinates equal to 1 is a trivial solution of (34), since $z=1$ is the solution of the initial eigenvalue problem. It is useful not to lose this solution in the approximation procedure in order to be sure, by orthogonality, that the non-trivial solutions will be centered and uncorrelated. The $\boldsymbol{\alpha}$ vectors may satisfy some other constraints such as nullity of some specific coordinates.

We look for the 'best substitute' that satisfies (34) for the true solution of (34). It turns out to be also the solution of a 'projected statistical problem' in a finite-dimensional space. Under some mild assumptions it can be shown that the approximate solution converges to the true ones, when the dimension l of the approximation increases.

The problem is now to find a z and $\boldsymbol{a}_t$ that satisfy (36) and minimize

$$\|z\|^{-2} \int_0^1 \|z - X_t \boldsymbol{a}_t\|^2 \, \mathrm{d}t. \tag{37}$$

The simplest and most usual way to choose the $f_l(t)$ functions consists in choosing $(L+1)$ points t_l in T, such that $0=t_0<\cdots<t_l<t_{l+1}<\cdots<t_L=1$, and to define

$$f_l(t) = 1 \quad \text{if} \quad t_{l-1} \leqq t < t_l,$$

$$\quad\;\; = 0 \quad \text{elsewhere.}$$

In this case $\boldsymbol{a}_t = \boldsymbol{\alpha}_l$ if t belongs to $[t_{l-1}, t_l]$ and we have simply to minimize

$$\sum_{l=1}^{L} \int_{t_{l-1}}^{t_l} [-2z'X_t\boldsymbol{\alpha}_l + \boldsymbol{\alpha}_l' X_t' X_t \boldsymbol{\alpha}_l]\,\mathrm{d}t. \tag{38}$$

Define

$$V_l = \int_{t_{l-1}}^{t_l} X_t\,\mathrm{d}t \quad \text{and} \quad D_l = \int_{t_{l-1}}^{t} X_t' X_t\,\mathrm{d}t. \tag{39}$$

The $m \times m$ matrix V_l has as its entries the time spent by the ith individual in the kth state between t_{l-1} and t_l. The $m \times m$ matrix D_l is diagonal and has as its entries the total time spent by all individuals in the kth state during the lth interval. It is clear that $\boldsymbol{\alpha}_l = D_l^{-1} V_l z$ and that the calculations are going on exactly as in subsection 2.1. Matrix D_l plays the role of the D_j and V_l plays the role of the X_j. In fact the computation comes down to performing correspondence analysis on the table $(V_1, V_2, \ldots, V_L)$.

This table is no longer a disjunctive table, but a table of the $L \times m$ new variables, ‘time spent during the lth interval of time in the kth state’. For the computation we have only to create these variables and then to use standard correspondence analysis software. All the interpretations of multiple correspondence analysis can be utilized; see, for instance, Benzécri (1973) and Lebart et al. (1977). The time dimension, however, is especially useful to determine the meaning of the factors.

The most general way to examine the question of approximation is to consider the $\boldsymbol{a}_t$ as a numerical function of two variables: time and category. The solution is searched for, in a prescribed finite-dimensional subspace of this set of functions. The computation generally does not come down to correspondence analysis, but rather to principal components analysis with a special metric derived from the data [see Deville (1982)].

As in multiple correspondence analysis, there are many different ways to present qualitative harmonic analysis. One of them seems to be of special interest because it generalizes to qualitative data the Karhunen–Loève expansion of a stochastic process. The point is to define an operator-valued covariance of two nominal variables by the product of their conditional expectation operator [Deville and Saporta (1980), Saporta (1981), Deville (1982)].

4.3. An application: Women who have been married at least three times

The data come from two French retrospective surveys in fertility. Qualitative harmonic analysis has been performed using data about 423 women who had been married three times or more for whom sufficient information was available, viz. the precise date of each marriage and dissolution date of those marriages which were dissolved, and the cause, death or divorce. The time series under study are the marital status of each woman from age 15 to 45. The state space has four categories: single, married, widowed and divorced. Some other variables are also available for every individual, notably: the number of children and their dates of birth, social status of the successive husbands, place of residence, date of birth. They are used as supplementary variables in the analysis and proved powerful tools for the interpretation of the factors.

Three different approaches have been used in the calculations and the three results are very similar [for a complete report, see Deville (1982)]. The main results are summarized in figs. 2 and 3. Only those given by the first approach (six intervals of five years) have been plotted. Fig. 2 presents the working variables of the analysis, i.e., the times spent in each state for every time interval. The coordinates are the value of the α_t-vectors for the first two factors. Straight lines connect points representing the same state at successive intervals. Fig. 3 shows passive variables in the same plane but with a larger scale.

The horizontal axis draws apart women with several divorces from women who where widowed more than once. It appears also to be related to time. Women born before 1900 are associated with widowhood, women born after 1925 with divorce. This is consistent with the decrease of mortality and the

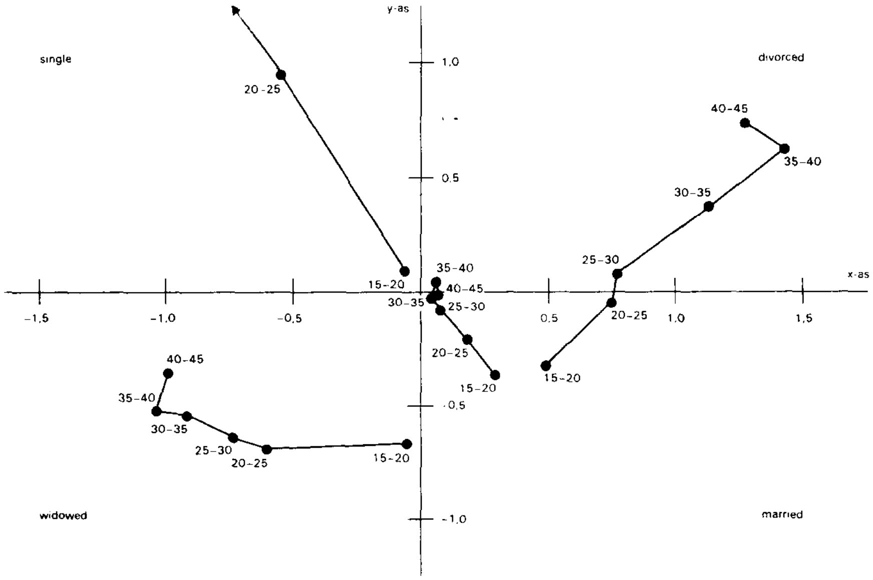

Fig. 2

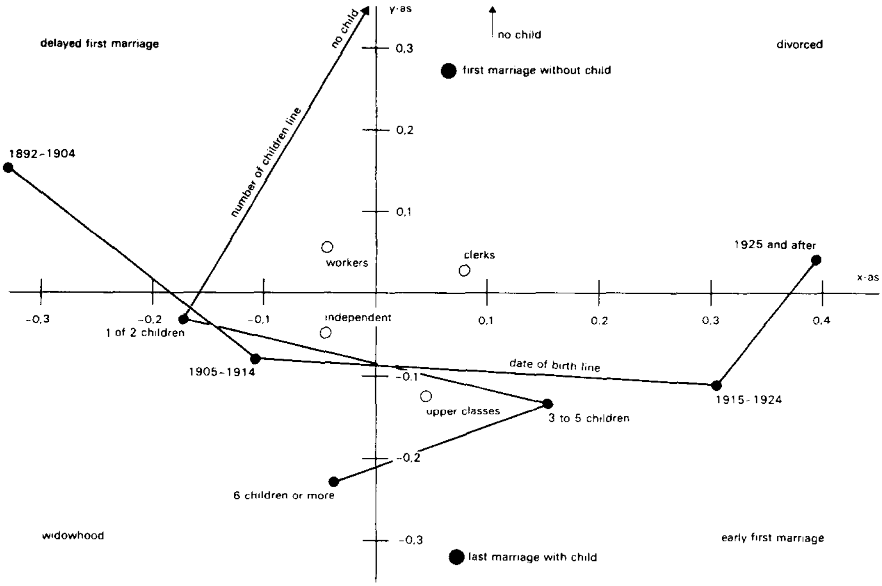

Fig. 3

rise of divorce during the 20th century. Women who married (at least once) a white-collar worker divorce more often than women who married only workers or men having an independent occupation (farmers of merchants). The marriages of women with four marriages end more frequently by divorce than by death of husband.

The vertical axis is related with the age at the time of the first marriage. The 'single' state contributes for 75% in the determination of this factor. However, it opposes much more women without children (about 15% of the sample) — who generally got married late — to women having one or several children. Workers' wives are here associated with the absence of children, although they are on the average more fertile than other women. The meaning of this association could be the following: in lower social classes, a divorced or widowed woman with a child has less chance to get remarried than a woman in the same situation belonging to the upper classes.

The third factor also allows a clear interpretation. It is determined essentially by events occurring between the ages of 35 and 45. It splits the sample in two major groups. In the first one, there is a stable third marriage, very often with children. Women belonging to the second one continue to have a very eventful life, even after 35 years.

5. Conclusion

Unlike other methods, correspondence analysis is not concerned with model building and is not oriented towards prediction. Its aim is essentially exploratory and consists in clarifying the essential structures of a large data set and, if these exist, the features that differentiate several subpopulations. Correspondence analysis may be applied at a very large scale to real data thanks to efficient software [such as SPAD of Lebart and Morineau (1982)] which is able to process arrays of very large dimension (several hundreds of variables and thousands of individuals).

References

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

Boumaza, R., 1980, Contribution à l'étude descriptive d'une fonction aléatoire qualitative, Thèse 3è cycle (Université Paul Sabatier, Toulouse).

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

Burt, C., 1950, The factorial analysis of qualitative data, British Journal of Psychology 3, 166–185.

Cailliez, F. and J.P. Pagès, 1976, Introduction à l'analyse des données (SMASH, Paris).

Carroll, J.D., 1968, A generalization of canonical correlation analysis to three or more sets of variables, Proceedings of the 76th convention of the American Psychological Association, 227–228.

De Leeuw, J., 1973, Canonical analysis of categorical data, Ph.D. thesis (Leyden University, Leyden).

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

Deville, J.-C., 1974, Méthodes statistiques et numériques de l'analyse harmonique, Annales de l'INSEE 15, 3–101.

Deville, J.-C., 1982, Analyse des données chronologiques qualitatives, Annales de l'INSEE 45, 45–104.

Deville, J.-C. and G. Saporta, 1979, Analyse harmonique qualitative, in: E. Diday, ed., Data analysis and informatics (North-Holland, Amsterdam) 375–389.

Fisher, R.A., 1940, The precision of discriminant functions, Annals of Eugenics 10, 422–429.

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

Greenacre, M.J., 1981, Practical correspondence analysis, in: V. Barnett, ed., Interpreting multivariate data (Wiley, New York).

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.

Guttman, L., 1959, Metricizing rank-ordered or unordered data for a linear factor analysis, Sankhyā 21, 257–268.

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., 1973, Reciprocal averaging: An eigenvector method of ordination, Journal of Ecology 61, 237–251.

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.

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

Kuder, G.F. and M. Richardson, 1933, Making a rating scale that measures, Personnel Journal 12, 36–40.

Lancaster, H.O., 1957, Some properties of the bivariate normal distribution considered in the form of a contingency table, Biometrika 44, 289–292.

Lancaster, H.O., 1969, The chi-squared distribution (Wiley, New York).

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

Lebart, L. and A. Morineau, 1982, SPAD système portable pour l'analyse des données (CESIA, 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).

Mardia, K.V., J.T. Kent and J.M. Bibby, 1979, Multivariate analysis (Academic Press, London).

Maung, K., 1941, Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children, Annals of Eugenics 11, 189–223.

Nishisato, S., 1980, Analysis of categorical data: Dual scaling and its applications (University of Toronto Press, Toronto).

Saporta, G., 1975, Dépendance et codages de deux variables aléatoires, Revue de Statistique Appliquée 23, 43–63.

Saporta, G., 1981, Méthodes exploratoires d'analyse de données temporelles, Thèse de doctorat es sciences (Université de Paris 6, Paris).

Tenenhaus, M. and F.W. Young, 1982, Multiple correspondence analysis and the principal components of qualitative data, Psychometrika, forthcoming.

Williams, E.J., 1962, The use of scores for the analysis of association in contingency tables, Biometrika 39, 274–289.