Correspondence analysis for categorical stochastic processes

Gilbert Saporta

INTERNATIONAL CONFERENCE ON
ADVANCES IN MULTIVARIATE STATISTICAL ANALYSIS
INDIAN STATISTICAL INSTITUTE
CALCUTTA - INDIA

16 - 20 DECEMBER 1985

CORRESPONDENCE ANALYSIS FOR CATEGORICAL STOCHASTIC PROCESSES

G. SAPORTA

Conservatoire National des Arts et Métiers
292 rue Saint-Martin
75141 PARIS CEDEX 03
FRANCE

Under the name of qualitative harmonic analysis we present an extension of multiple correspondence analysis which is fitted to handle event-history data in an exploratory way.

Principal features of correspondence analysis are reminded in parts 1 and 2.

1 - CORRESPONDENCE ANALYSIS (C.A.)

Strictly speaking, C.A. is a way of displaying simultaneously rows and columns of a contingency table as a set of points in a low-dimensional space. A recent presentation in English is Greenacre (1984).

1.1 Notations and equations

Let N be a contingency table of m_1 rows and m_2 columns with entries n_{ij} of sum n. Let D_1 and D_2 be the diagonal matrices of rows and column marginal frequencies

$$D_1 = \text{diag}\ (n_{1.}\ ,\ n_{2.}\ ,\ \ldots\ ,\ n_{m_1.})$$

Then the coordinates of the m_1 rows along the first axis are given by the eigenvector $\underline{a}$ associated to its largest eigenvalue λ_1 of

$$D_1^{-1}\ N\ D_2^{-1}\ N' \qquad (1)$$

(the trivial solution 1 being rejected)

Conversely the coordinates of the columns are given by the eigenvectors $\underline{b}$ of $D_2^{-1}\ N'\ D_1^{-1}\ N$.

With $\frac{1}{n}\ \underline{a}'\ D_1\ \underline{a} = \frac{1}{n}\ \underline{b}'\ D_2\ \underline{b} = \lambda$ we have

$$\begin{cases} \underline{b} = \frac{1}{\sqrt{\lambda}}\ D_2^{-1}\ N'\ \underline{a} \\ \underline{a} = \frac{1}{\sqrt{\lambda}}\ D_1^{-1}\ N\ \underline{b} \end{cases} \qquad (2)$$

$D_1^{-1}\ N$ and $D_2^{-1}\ N'$ are the tables of conditional frequencies

1.2 Various presentations

The preceding equations have been rediscovered a great number of times in various contexts. Among them, quantification of categorical variables plays a great part ; $\underline{a}$ and $\underline{b}$ are vectors of category scores giving a maximal correlation between the quantified variables (Fisher 1940) or optimal linear regression (Hirschfeld 1935).

Dual scaling (Nishisato 1980) or reciprocal averaging (Hill 1973) are basically equivalent and consist in displaying rows and columns categories along a single line in order that the coordinate of category i , a_i , be (apart from a scaling factor) equal to the weighted mean of the coordinates of the columns category $\sum_j \frac{n_{ij}}{n_{i.}}\ b_j$ and conversely.

However all these presentations are essentially unidimensional and the need for solutions associated to the eigenvectors of order 2, 3 etc is not obvious. It is undoubtedly the merit of J.P. Benzecri (1973) of having transformed what

was only a mathematical property into an efficient tool of graphical data analysis : basically his approach comes down to a principal component analysis of both arrays of conditional frequencies;the row percentages are considered as a set of m_1 weighted points in $\mathbb{R}^{m_2}$, the vector-space having a scalar product defined by the matrix $n\ D_2^{-1}$ (the chi-square metric).

In addition to the graphical display, various index-numbers such as contributions to inertia provide helpful informations for the interpretation of the outputs.

2 - MULTIPLE CORRESPONDENCE ANALYSIS (M.C.A.)

Like C.A. , M.C.A. has been presented and rediscovered several times and is known under various names."Principal components of scale"Guttman (1941), "Quantification of Type II"Hayashi (1950),"Homogeneity Analysis"Gifi (1981). See Tenenhaus and Young (1985) for a review and Lebart and al (1984).

The name of multiple correspondence analysis used here is due to the following property : if X_1 and X_2 are the two indicator matrices of size $n\times m_1$ and $n\times m_2$ associated to the contingency table N , then a formal correspondence analysis of $X = (X_1 \mid X_2)$ gives results equivalent to the C.A. of N. The coordinates of the columns of X are proportional to $\underline{a}$ and $\underline{b}$.

2.1 Notations and equations

The data consist in a set of n observations of p categorical variables $\mathcal{X}_1 , \mathcal{X}_2 , \ldots \mathcal{X}_p$ with $m_1 , m_2 , \ldots , m_p$ categories.

X is the supermatrix of indicator variable $X = (X_1 \mid X_2 \mid \ldots \mid X_p)$

$$D = \begin{bmatrix} D_1 & & \bigcirc \\ & D_2 & \\ \bigcirc & & \ddots D_p \end{bmatrix}$$ the superdiagonal array of category counts.

$B = X'X$, the so-called Burt's matrix is the superarray of all bivariate cross-tables

$$B = \begin{bmatrix} D_1 & N_{12} & \cdots & N_{1p} \\ N_{21} & D_2 & \cdots & N_{2p} \\ \vdots & & \ddots & \\ \cdots & & & D_p \end{bmatrix}$$

M.C.A. gives a graphical display of the n observations and of the Σm_i categories of the p variables.

If $\underline{z}^{(k)}$ and $\underline{a}^{(k)}$ are the respective vectors of coordinates along the k^{th} axis then

$$\frac{1}{p} D^{-1} B \underline{a}^{(k)} = \lambda_k \underline{a}^{(k)}$$

$$\frac{1}{p} X D^{-1} X' \underline{z}^{(k)} = \lambda_k \underline{z}^{(k)} = \frac{1}{p} \left(\sum_{i=1}^{p} X_i (X_i' X_i)^{-1} X_i' \right) \underline{z}^{(k)}$$

Apart from a scaling factor, the coordinate of each observation is the mean of the coordinates of the categories it belongs to :

$$\underline{z} = \frac{1}{\sqrt{\lambda}} \frac{1}{p} X \underline{a} \qquad (4)$$

and the coordinate of each category is the mean of the coordinates of the observations which belong to them :

$$\underline{a} = \frac{1}{\sqrt{\lambda}} D^{-1} X' \underline{z} \qquad (5)$$

Equations (3) to (5) are those of correspondence analysis where X stands instead of N in equations (1) and (2).

2.2 M.C.A. as an extension of principal components analysis.

The variables $\underline{z}$ give extremal values to the sum of their square correlation ratio with the $\mathcal{X}_j$ (Saporta 1980)

$$\max \sum_{j=1}^{b} \eta^2 (\underline{z} ; \mathcal{X}_j) \qquad (6)$$

This property is similar to that of principal components $\underline{c}$ of p standardized variables $\underline{x}^1 \underline{x}^2 \ldots \underline{x}^p$

where $\sum_{j=1}^{p} r^2 (\underline{c} ; \underline{x}^j)$ is maximized.

M.C.A. is thus a generalization of usual principal components to nominal variables.

Furthermore, M.C.A. is a particular case of nonlinear principal components analysis where one looks for the maximum of $\sum_j r^2 (\underline{c} ; \varphi_j(\underline{x}^j))$ over

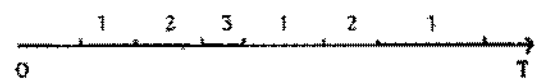
$\underline{c}$ and φ_j ; the φ_j being here stepfuctions (see Dauxois, Pousse (1976), De Leeuw (1982)).

3 - QUALITATIVE HARMONIC ANALYSIS (Q.H.A.)

Presented by Deville and Saporta (1980) as an extension of the Karhünen-Loeve decomposition of real valued stochastic processes to categorical ones, Q.H.A. is a generalization of multiple correspondence analysis for handling event-history data in continuous time : Saporta (1981), Deville (1982), Deville and Saporta (1983), De Leeuw (1984). We outline here some of its properties.

3.1 Field of application

Q.H.A. handles data which are a set of trajectories of a categorical stochastic process: with a finite number of states, say m, each observation is a sequence of states and dates of transition during a time interval $[0 ; T]$ (where transitions may appear at any time)



Evolution of marital status of a sample of french women are studied in Deville (1982) and Deville-Saporta (1983). De Leeuw, Van der Heijden, Kreft (1984) present time-activity data (with a discrete time).

At any t between 0 and T a categorical variable $\mathcal{X}_t$ is thus known. The problem is to deal with an infinite set of $\mathcal{X}_t$: of course if there is a finite number of observations and a finite number of transitions, there is only a finite number of distinct $\mathcal{X}_t$ but the theory of Q.H.A. works for an (even not countable) infinity of $\mathcal{X}_t$.

3.2 Notation

1_t^x will be the indicator variable of state x at time t : $1_t^x(i) = 1$ if observation i belongs to the x-th category of $\mathcal{X}$ at time t $1_t^x(i) = 0$ if not.

X_t is the $m \times m$ matrix of the indicator variables at time t.

Thus $X_t' X_s = N_{ts}$ is the $m \times m$ matix with elements n_{ts}^{xy} = number of observations being in category x at time t and in category y at time s.

$X_t' X_t = N_{tt}$ is the diagonal matrix of marginal frequencies of the m states at time t.

$(X_t' X_t)^{-1} X_t' X_s = N_{tt}^{-1} N_{ts}$ is the transition matrix from time t to time s, its elements are the conditional frequencies of being in some category y at time s knowing the category at time t (notice that t is not necessarily less than s).

3.3 The method

Whe look for a display of observations with time-invariant coordinates.

Using the criterium of M.C.A. the vector $\underline{z}$ of coordinates of the n observations along an axis should maximize :

$$\int_0^T \eta^2 (\underline{z} ; \mathcal{X}_t)\, dt \tag{7}$$

Since $\eta^2 (\underline{z} ; \mathcal{X}_t) = \dfrac{\underline{z}' X_t (X_t' X_t)^{-1} X_t' \underline{z}}{\underline{z}' \underline{z}}$

$\underline{z}$ is solution of the following eigenequation.

$$\lambda \underline{z} = \left[\int_0^T X_t (N_{tt})^{-1} X_t' \, dt \right] \underline{z} \tag{8}$$

Equation (8) is a matrix-equation of the form $Q\underline{z} = \lambda \underline{z}$ where Q is a matrix of size n which has an interesting interpretation in terms of inter-individual similarities :

$X_t (N_{tt})^{-1} X_t'$ is a $n \times n$ matrix where element (i,j) is zero if observations i and j are not in the same category at time t and $1/n_t^x$ if observations i and j are in the same category x at time t : i and j are more similar if they belong both to a seldom category with a small value of n_t^x than to a common one.

Q is then an integrated matrix of similarities and coordinates of individuals are given by the successive eigenvectors of their similarity matrix like in classical scaling or principal coordinate analysis.

Knowing the vector $\underline{z}$ of coordinates of the n observations along an axis, it is natural to display the category x of $\mathcal{X}$ at time t as the mean of the n_t^x points corresponding to observations which are actually in category x at time t.

The vector $\underline{a}_t = (a_t^1, a_t^2, \ldots, a_t^m)$ of these coordinates is such that :

$$\underline{a}_t = (N_{tt})^{-1} (X_t)' \underline{z} \tag{9}$$

An easy substitution in (8) gives the integral equation :

$$\lambda \underline{a}_t = \int_0^T (N_{tt})^{-1} (N_{ts})\, \underline{a}_s \, ds \tag{10}$$

It may be proved that $\underline{z}$ ant $\underline{a}_t$ are canonical variables of the following problem in the context of canonical analysis of two σ-algebras (Dauxois-Pousse (1976)). Data are probability measures over the product space $\Omega \times (T \times M)$ (Ω space of observations, M space of states) and we look for functions of Ω and functions of $T \times M$ maximally correlated Saporta (1981).

If changes of categories can only occur at fixed equally-spaced instants $0, t_1, t_2, \ldots t_p = T$, integrals are finite sums and equations (8) and (10) become :

$$\lambda \underline{a}_j = \sum_{j=1}^{p} N_{jj}^{-1} N_{ji} \, \underline{a}_i \tag{10'}$$

$$\lambda \underline{z} = \sum_{i=1}^{p} X_i (X_i' X_i)^{-1} X_i' \underline{z} \tag{9'}$$

Or in terms of the supermatrices $X = (X_1 : X_2 : \dots X_p)$

$$B = \begin{pmatrix} N_{11} & N_{12} & \dots & N_{1p} \\ \vdots & & \ddots & \vdots \\ N_{p1} & \dots & & N_{pp} \end{pmatrix} \qquad D = \begin{pmatrix} N_{11} & & \\ & N_{22} & \\ & & N_{pp} \end{pmatrix}$$

$$\begin{cases} \lambda \underline{a} = D^{-1} B \underline{a} \\ \lambda \underline{z} = X D^{-1} X' \underline{z} \end{cases}$$

which are the equations of multiple correspondence analysis of X apart from the constant p.

3.4 Approximate solution

Except for the previous particular case, equations (8) and (10) are not tractable : though for a finite sample equation (10) is not a true integral equation but a matrix equation (since there is a finite number of instants of transition, integral reduces to a sum. But the size of equation (8) and (10) are either the number of observations, or the number of dates of transition for all observations.

So, we have to discretize time into p periods $[t_0, t_1], [t_1, t_2], \dots [t_{p-1}, t_p]$ of length $T_1, T_2, \dots, T_p$ with $\sum_{j=1}^{p} T_j = T$ and look for solution $\underline{a}_t$ which are piecewise constant.
It is equivalent to approximate the process of the indicator variables 1_t^x by a piecewise constant vector-process.

Since, in the mean-square sense, the best approximation of a function by a constant is its mean it comes down to replace each trajectory i by a sequence of positive numbers $t_j^x(i)$ where $t_j^x(i)$ is the time spent by observation i in the state x during the period $[t_{j-1}, t_j]$.

Equation (8) and (10) are then equations of correspondence analysis of the following $nx\ mp$ table

$$\begin{array}{c|ccc|cccc|c|ccc|}
 & 1 & 2 & m & 1 & 2 & x & m & & 1 & 2 & m \\
1 & & & & & & & & & & & \\
2 & & & & & & & & & & & \\
i & t_1^1(i) & \dots & t_1^m(i) & & \dots & t_2^x(i) & & & & \dots & t_p^m(i) \\
n & & & & & & & & & & & \\
\end{array}$$

first period — second period — p^{th} period

where $\sum_{x=1}^{m} t_j^x(i) = T_j$ for every i

If the subdivision is such that each observation passes at most one time in each state there is no loss of information.

3.5 Decomposition of the eigenvalues

Since each eigenvalue λ is proportional to

$$\int_0^T \sum_{x=1}^{m} \frac{n_t^x}{n} (a_t^x)^2 \, dt$$

We are able to define the contributions of various entities to the eigenvalues :

contribution of instant t : $\sum_{x=1}^{m} \frac{n_t^x}{n} (a_t^x)^2$

contribution of category x at time t : $\frac{n_t^x}{n} (a_t^x)^2$

cumulated contribution of category x : $\int_0^T \frac{n_t^x}{n} (a_t^x)^2 \, dt$

which are useful for interpreting the principal axes.

4 - CONCLUSION

With few modification correspondence analysis may handle individual categorical time-series for exploratory purposes. The absence of a specified probabilistic model must be counterbalanced by a large amount of data. However the link between descriptive techniques and theoritical properties of processes should be of a great interest ; knowing results of Qualitative Harmonic Analysis for standard types of processes would be a precious help for modelling real-life situations.

REFERENCES

BENZECRI J.P. and al (1973) : L'analyse des données, Dunod, Paris.

DAUXOIS J. and POUSSE A. (1976) : Les analyses factorielles en calcul des probabilités et en statistique. Thèse Université Paul Sabatier, Toulouse.

DE LEEUW J. (1982) : Nonlinear principal components analysis, Compstat 82, Physica-Verlag 77-86.

DE LEEUW J. (1984) : Homogeneity analysis of curves and processus, "Table ronde analyse des données", Toulouse.

DE LEEUW J., VAN DER HEIJDEN P., KREFT I. (1985) : Homogeneity analysis of event history data. Methods of Operations Research 50, 299-316.

DEVILLE J.C. (1982) : Analyse des données chronologiques qualitatives, Annales de l'INSEE n° 45, 45-104.

DEVILLE J.C., SAPORTA G. (1980) : Analyse harmonique qualitative in Data Analysis and Informatics (E. Diday editor) 375-389, North-Holland, Amsterdam.

DEVILLE J.C., SAPORTA G. (1983) : Correspondence analysis, with an extension towards nominal time series. Journal of Econometrics 22, 169-189.

FISHER R.A. (1940) : The precision of discriminant functions. Annals of Eugenics 10, 442-449.

GIFI A. (1981) : Nonlinear multivariate analysis. University of Leiden. New edition Leiden DSWO Press 1984.

GREENACRE (1984) : Theory and application of correspondence analysis, Academic Press, New-York.

GUTTMAN L.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 Inst of Math. Stat. 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, Proc. Cambridge Phil. Soc. 31, 520-524.

LEBART L., MORINEAU A., WARWICK K. (1984) : Multivariate descriptive statistical analysis, Wiley, New-York.

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

SAPORTA G. (1980) : About some remarkable properties of Carroll's generalized canonical analysis, Paper presented at the 2nd European Meeting of the Psychometric Society, Groningen, N.L.

SAPORTA G. (1981) : Méthodes exploratoires d'analyse de données temporelles, Cahiers du Buro n° 37-38, Paris (Thèse Université Paris 6).

TENENHAUS M., YOUNG F.W. (1985) : An analysis and synthesis of multiple correspondence analysis, optimal scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika, [illegible], 91-119.