Recent Developments in Statistics
J.R.Barra et al., editors

SOME METHODS OF QUALITATIVE DATA ANALYSIS.

J.M. BOUROCHE
C O R E F

G. SAPORTA
I.U.T. PARIS V

M. TENENHAUS
C E S A

Most methods of data analysis have been conceived for numerical data (factor analysis, canonical analysis...), sometimes for both numerical and nominal data (analysis of variance and covariance, discriminant analysis). Other methods (multidimensional scaling, preference analysis, isotonic regression) enable the treatment of ordinal data but in a very restrictive and particular context.

Various statisticians have already proposed some methods for a mix of nominal and ordinal data e.g: Benzecri [1], Bock [3], Bouroche, Saporta and Tenenhaus [4], Carroll [5], Kruskal [8] [9], Lebart [10], de Leeuw [11], Masson [12], Nishisato [13], Pagès [14], Saporta [15], Tenenhaus [18], Young [19].

In the C.O.R.E.F-D.G.R.S.T. n° 75 07 0230 project the authors of this paper have attempted to achieve the following purposes :

- to complete existing syntheses
- to propose new methods
- to apply methods of analysis of qualitative data on real data to make sure of the value of results.

We present here a first synthesis of this project.

1.- DATA AND PROBLEMS

Before analysing a set of data, two questions arise :

- What is the nature of the data ?
- What are the problems to solve ?

The choice of a method derivesfrom the answers to these questions.

1.1. Nature of data.

A variable is nominal if the set of its categories is finite and has no ordinal structure.

A variable is ordinal if the set of its categories is finite with an ordinal structure.

A variable is numerical if it takes its values in $\mathbb{R}$.

Moreover F.W. Young [19] defines the concept of "underlying (or generating) process" which can be either discrete or continuous.

So we have six types of measurements : for instance a variable will be continuous-ordinal if the ordered set of its categories represents a continuous underlying process.

In a first stage, we have to determine exactly the nature of each variable. Even if a variable generally belongs to only one class, it may be interesting to modify arbitrarily its nature.

For instance we may consider a discrete ordinal variable as either discrete nominal (we neglect the ordinal structure) or continuous ordinal or discrete-ordinal.

1.2. Problems.

There are two kinds of problems. Either we have few information about the set of data and we are looking for a description, or we need a prediction of one (or more) specified variable through the others.

In the first case we shall use factor analysis or clustering techniques, in the second case methods derived from least squares and canonical analysis.

2.- METHODS

According to the nature of variables and to the problem, different methods are available. Before quoting a few of them in § 2.2., we present briefly the principle of optimal scaling.

2.1. Quantification of qualitative data.

Roughly speaking, the question is to allot a numerical value to each category of a discrete nominal or ordinal variable. If the variable is ordinal we require that the order of the numerical values represent the order of the categories.

Let E be the set of individuals, X a nominal variable. X is a mapping from E to $\mathfrak{X}$, set of categories. Let δ be a mapping from $\mathfrak{X}$ to $\mathbb{R}$. The variable $\delta \circ X$ is then numerical : it is a quantification (scaling) of X. As there is an infinite number of scalings, the choice of an optimal scaling is actually function of a criterion related to the method (cf. § 3).

If the variable X is continuous nominal or continuous ordinal, an interval of $\mathbb{R}$ corresponds to a category of X and the scaled observations are required to fall in the interval but have not necessarily to be equal. If X is ordinal, we require that the order of intervals corresponds to the order of categories, (cf. F.W. Young [19]).

2.2. Choice of methods.

We shortly quote some methods and their authors. For further details see the reference list or the final C.O.R.E.F report.

2.2.1. Descriptive methods.

We deal only with methods of principal component analysis type.

a). All variables are nominal.
See Bock [3], Bouroche, Saporta, Tenenhaus [4], Lebart [10], Nishisato [13].

b). Nominal and numerical variables
see Tenenhaus [18] and § 3.

c). Ordinal variables
see Kruskal and Shepard [9].

d). Mix of ordinal, nominal, numerical variables
see Young, de Leeuw, Takane [20], de Leeuw [11].

2.2.2. Predictive methods.

The method proposed by Young, de Leeuw and Takane [21] is the only one available for the general case. Let us point out two particular cases :

a). One nominal dependent variable, all predictors nominal : Carroll [5], Masson [12], Saporta [15].

b). One ordinal dependent variable, all predictors nominal : Kruskal [8].

Let us emphasize on the great flexibility of the Young, de Leeuw and Takane [21] approach : their MORALS/CORALS algorithm allows the treatment of any mix of variables in terms of regression analysis or canonical analysis.

More modestly, we shall now present two methods based upon optimal scaling of nominal data : a predictive method (DISQUAL) and a descriptive one (PRINQUAL).

3.- TWO NEW METHODS.

3.1. DISQUAL : a method and a program for stepwise discriminant analysis of nominal variables [17].

p nominal variables with $m_1, m_2, \ldots m_p$ categories are measured on n individuals and we attempt to predict the categories of an outside nominal variable with only k of these p predictors ($k < p$).

The method consists of two relatively separate parts : on the one hand the selection of the predictors, on the other hand the discrimination performed by means of the selected predictors.

3.1.1. Stepwise selection with Escoufier's operators.

According to Escoufier [6] and Pagès [14] we associate to each of the p+1 nominal variables the orthogonal projector P_i on the subspace of $\mathbb{R}^n$ of zero-mean variables spanned by the indicator variables of its categories This is the subspace of zero-mean discrete numerical variables obtained by scaling the nominal variable.

Projectors belong to the subset of the vector space of symetrical operators in which we define an inner product and a norm by

$$\langle P_i ; P_j \rangle = \text{Trace}(P_i P_j)$$
$$||P_i||^2 = \text{Trace } P_i^2.$$

In the case of nominal variables we know that $\text{Trace } P_i P_j = \Phi^2$ where Φ^2 is the K-Pearson measure of dependence between variables i and j (canonical or correspondence analysis of two nominal variables).

Furthermore $\text{Trace } P_i^2 = \text{Trace } P_i = m_{i-1}$ and the cosine of the angle between two operators associated to two nominal variables is nothing else than

the Tschuprow coefficient

$$T_{ij} = \frac{\phi^2}{\sqrt{(m_{i-1})\,(m_{j-1})}}$$

which is thus equivalent to a correlation coefficient between nominal variables if we identify a variable to its associated projector.

According to the usual geometry of correlation, we may then define the partial Tschuprow coefficient between nominal variables. For three variables i,j,k we have

$$T_{ij/k} = \frac{T_{ij} - T_{ik}\,T_{jk}}{\sqrt{(1-T_{ik}^2)\,(1-T_{jk}^2)}} \quad \text{and so on.}$$

Then stepwise selection method is :

- The first predictor is the one which maximizes Tschuprow's coefficient with the dependent variable
- The second predictor is the one which maximizes the partial Tschuprow coefficient with the dependent variable, given the first predictor.

3.1.2. Discriminant analysis.

k predictors being selected we have now an array of data of the following kind :

$$(A \,|\, X_1 \,|\, X_2 \,|\, \ldots \,|\, X_k)$$

where X_i (and A) are $n \times m_i$ matrices where the columns are the indicators of the categories. The rank of $X = (X_1 | X_2 \ldots | X_k)$ is inferior or equal to

$$\left(\sum_{i=1}^{k} m_i - k \right),$$

since the sum of the columns of each X_i is the vector $\underline{1}$.

An ordinary discriminant analysis being impossible here, because $X'X$ is not regular, we substitute to X a new matrix of nearly equivalent numerical variables : these new variables are the Guttman principal components of scale of the k nominal variables (which are also the components of Carroll's generalized canonical analysis [4] or of correspondence analysis of X). Among the $\sum_{i=1}^{k} m_i - k$ principal components, we retain only those which have a sufficient discriminating power.

A discriminant factor analysis performed on the selected components gives the discriminating scores of the nominal variables which we directly use for the classification technique if the dependent variable is binary (Fisher's function) ; otherwise we use a **classical procedure** based on the distances to the centroids.

3.2. PRINQUAL : a method and a program of principal component analysis of a set of nominal and numerical variables [18].

We are looking for a scaling of the nominal variables in order to get the best principal component analysis with factors in the sense of a maximum explained variance.

3.2.1. Method

Let E be the set of individuals, $D_1, D_2, \ldots D_k$ nominal variables, $X_1, X_2, \ldots X_\ell$ numerical variables. Let $\delta_i \circ D_i$ be the scaled variables. We know that the first m principal components of $\delta_i \circ D_i$ (supposed known) and X_i realize

$$\underset{Z_1, Z_2, \ldots Z_m}{\text{Max}} \; \sum_{i=1}^{k} \sum_{j=1}^{m} \text{cor}^2(\delta_i \circ D_i,\, Z_j) + \sum_{i=1}^{\ell} \sum_{j=1}^{m} \text{cor}^2(X_i\, Z_j)$$

where the Z_j are uncorrelated.

Thus the principle of the method consists in obtaining the optimal scaling δ_i by maximizing the previous expression both on δ_i and Z_j.

3.2.2. Algorithm.

The algorithm is iterative and maximizes the criterion alternatively on the δ_i and on the Z_j.

The initial solution $Z_j^{(o)}$ is optimally chosen by using generalized canonical analysis [18].

At step t we get the $\delta_i^{(t)}$ by

$$\underset{\delta_1, \delta_2, \ldots \delta_k}{\text{Max}} \; \Sigma\Sigma \text{cor}^2\,(\delta_i \circ D_i\,,\, Z_j^{(t-1)}) + \Sigma\Sigma \text{cor}^2\,(X_i\,,\, Z_j^{(t-1)})$$

Let $\lambda(t)$ be the value of this maximum.

We have now the $Z_j^{(t)}$ by

$$\underset{Z_1, Z_2, \ldots Z_m}{\text{Max}} \; \Sigma\Sigma \text{cor}^2\,(\delta_i^{(t)} \circ D_i\,,\, Z_j) + \Sigma\Sigma \text{cor}^2\,(X_i, Z_j)$$

Let $\mu(t)$ be the value of this maximum.

It is possible to show that :

$$\lambda(t) \leq \mu(t) \leq \lambda(t+1) \leq k+\ell$$

Thus the algorithm converges and

$$L = \lim_{t \to \infty} \lambda(t) : \lim_{t \to \infty} \mu(t)$$

Let

$$\delta_i^{*} = \lim \delta_i^{(t)}$$

$$Z_j^* = \lim Z_j^{(t)}$$

with $\delta_i \circ D_i$ and Z_j of zero-mean and unit-variance.

3.3. Principal component analysis of the $\delta_i^* \circ D_i$ and X_i

The Z_j^* are the first m principal components.

The part of explained variance is $\frac{L}{k+\ell}$.

We may represent observations and variables as usual in principal conponent analysis.

The categories are represented in $\mathbb{R}^m$ as follows

To the category k of variable D_i we associate the vector :

$$M_{i,k} = \frac{1}{|D_i^{-1}(k)|} \sum_{e \in D_i^{-1}(k)} \begin{pmatrix} Z_1^*(e) \\ Z_2^*(e) \\ \vdots \\ Z_m^*(e) \end{pmatrix}$$

REFERENCES

[1] Benzecri, J.P. (1974). L'analyse des données, Tomes I and II.

[2] Bertier P. & Bouroche J.M. (1975). L'analyse des données multidimensionnelles. (P.U.F.).

[3] Bock, D. (1960). Methods and applications of optimal scaling-Psychometric Laboratory, Report n° 26, University of North Carolina.

[4] Bouroche, J.M., Saporta, G., & Tenenhaus (1975). Generalized canonical analysis of qualitative data. U.S. Japan Seminar on multidimensional scaling and related methods, San Diego.

[5] Carroll, J.D. (1969). Categorical conjoint measurement, Ann Arbor, Michigan, Meeting of Mathematical Psychology.

[6] Escoufier, Y. Le traitement des variables vectorielles, Biometrics 29, p. 751.

[7] Hayashi, C. (1950). On the quantification of qualitative data from the mathematical-statistical point of view. Annals of the Institute of Statistical Mathematics, 2, 35-47.

[8] Kruskal, J.B. (1965). Analysis of Factorial Experiments by Estimating Monotone Transformation of the data, JRSS, Serie B, 27.

[9] Kruskal, J.B. and Shepard, R. (1974). A non metric variety of linear factor analysis, Psychometrika, 39, 123-157.

[10] Lebart, L. (1973). Recherche sur la description automatique des données socio-économiques, Rapport de recherche, CREDOC.

[11] De Leeuw, J. (1976). HOMALS, Spring meeting of the Psychometric Society, Murray Hill.

[12] Masson, M. (1974). Thèse d'Etat, Université de Paris VI.

[13] Nishisato, A.P.S. (1975). Non linear programming approach to optimal scaling of partially ordered categories, Psychometrika, 40, 522-548.

[14] Pagès, J.P. (1974). A propos des opérateurs d'Escoufier, Séminaires de l'I R I A.

[15] Saporta, G. (1975). Liaison entre plusieurs ensembles de variables et codage des données qualitatives, Thèse de 3ème Cycle, Paris VI.