Multidimensional data analysis and quantification of categorical variables

Gilbert Saporta

New Trends in Data Analysis and Applications
J. Janssen, J.-F. Marcotorchino, J.-M. Proth (eds.)

MULTIDIMENSIONAL DATA ANALYSIS AND QUANTIFICATION OF CATEGORICAL VARIABLES

Gilbert SAPORTA

IUT de Paris
143 Avenue de Versailles, 75016 Paris

Quantification is a powerful tool for dealing with qualitative variables in data analysis : it is both an efficient way for description and prediction with nominal data and a conceptual framework for presenting many techniques. This paper attempts to be a survey of this topic.

1. USE AND DEFINITION OF QUANTIFICATION

By "quantification" (the words "scaling" or "scoring" are also often used) we mean a transformation of one or several categorical variables into numerical ones. Before describing the way to achieve a quantification (and the problems it involves) it is necessary to argue about the rationale of this approach.

1) Quantification : a device or a fundamental method ?

a) Numerical methods for data which are not numerical.

The main consequence of quantifying qualitative variables is to allow the use of usual statistical techniques such as principal components analysis, multiple regression or discriminant analysis for instance, by coming down to the usual situation of numerical variables.

That convenience was during a long time the only justification of the quantification techniques, what may seem a somewhat poor argument from a theoretical point of view : isn't it by lack of imagination or laziness that one does not develop methods which would be fitted to the qualitative nature of data ?

Let us notice here that the quantification is a way of processing variables of different kinds (numerical and categorical)

by giving the same part to each variable. Let us suppose that we have a first set of m numerical variables $x_1 \ldots x_n$ and a second set of f qualitative variables $\chi_1, \chi_2, \ldots \chi_m$ and that we want to make a descriptive data analysis of all the (m+p) variables by means of a component-like analysis ? There are actually four possibilities, the first two giving non symmetrical parts to the two sets :

- Perform a principal component analysis with the x_j $j=1,2,..m$ and use the χ_k as additional variables by representing the categories of each χ_k by the averages of the individuals which belongs to them . We have here a representation of the χ_k in the space of the individuals.

- Perform a multiple correspondence analysis of the χ_k and use the x_j as additional variables by computing the correlation coefficient of the x_j with the principal components. The representation of the x_j is here in the space of variables.

- Split into categories the numerical variables and perform a multiple correspondence analysis with (m+p) qualitative variables.

- Quantify the χ_k into x_k^* and perform a principal components analysis with (m+p) numerical variables.

This last possibility is what we proposed above ; the third one seems different but we will see that it is also a quantification technique.

b) Multivariate analysis as quantification techniques.

Actually many classical methods which are dealing with categorical variables may be considered as quantification techniques.

For instance analysis of variance or covariance are performing quantification of the nominal "factors of variability" when estimating the effects upon the dependent variable (in the case of the no-interaction model). Canonical discriminant analysis is merely a regression of the quantified group-variable upon a set of predictors with an "optimal" quantification i.e. giving the largest determination coefficient among all possible quantifications (see for instance Mc Donald and Cailliez-Pages).

Furthermore any multidimensional scaling method for qualitative variables (such as correspondence analysis) which gives

coordinates for the categories of a set of qualitative variables is in fact a multidimensional quantification technique : the coordinates along an axis are numerical values (scores) to be assigned to a qualitative variable.

c) An approximation of non-linear analysis.

Multidimensional data analysis is mostly a set of methods using linear algebra and the vector-space structure of the sets of variables and individuals : these methods can only place in prominent position linear relationships between variables what is well fitted to the case of multinormal distribution which is exceptional in most applications.

In most cases, the study of linear relationships is not sufficient and a non-linear analysis is actually necessary.

The quantification of qualitative variables provides an approximation of non linear analysis : let us transform a numerical variable x into a qualitative one X by simply splitting into classes the set of its values and then let us quantify X into x^*.

The set of all possible quantifications of X is an approximation, by piecewise constant functions, of the set of all functions of x f (x).

As the study of linear relationships between any functions f(x) and g(y) of two variables x and y is nothing else than the study of non-linear relationships between x and y, we therefore get a tool for non linear multivariate analysis (Masson). In fact it is possible to define less simple transformation of variables to get non-linearity, such as transformation by spline-functions of higher order (splitting and quantifying is equivalent to splines of degree zero) (Van Rijckevorsel and De Leeuw, Lafaye de Michaux), but it is slightly more complicated (spline functions are a kind of regular functions, very popular in the field of interpolation).

As Dauxois and Pousse quoted it, these techniques are not purely non-linear but only semi-linear in the following sense : if we use transformed variables x_j^* instead of raw variables x_j in a multidimensional analysis we will deal with linear combinations of the x_j^* but not with any functions of the x_j.

2) How to quantify ?

If we accept quantification as a justifiable practice we have however to make clear some points such as : among all possible quantifications of a variable how much must we choose and according to which criteria ?

a) Quantification and type of variables

When a qualitative variable X is purely nominal (without any structure upon the set of its categories) a quantification is a transformation of X into a discrete numerical variable : we will assign the same numerical value, say a_i, to all individuals belonging to the ith category of X.

If the variable X is ordinal (there is a total natural order for its categories such as level of education for instance), people often suggest to use only quantifications respecting the order of the categories : the numerical values to be assigned to the m ordered categories of X must be such as $a_1 \leq a_2 \leq \ldots \leq a_m$. Some authors (Nishisato mainly) consider the more general situation of a partial order of the categories.

Quantification under order restrictions leads to a more sophisticated theory than quantification without constraints, using convex cones instead of vector subspaces (see Barlow and al. or Tenenhaus) and to more difficult computations. But we get essentially a one-dimensional quantification as we will see below.

Apart from these difficulties we may wonder if it is actually necessary to respect order constraints : for instance in a prediction problem where an explanatory variable is ordinal and the variable to be predicted is numerical, quantification with order restrictions comes down to postulate the existence of a monotonic relationship. Do we have to introduce such an a priori before having studied the relationship ? It may be more interesting to make an analysis without order restrictions and see afterwards if we get a quantification which respects the ordinal nature of the categories. If not, it will be a proof of a non monotonic relationship provided there is no sampling errors. Of course if we have strong reasons for a monotonic relationship we need to use this information.

Conversely if it is the dependent variable which is ordinal we must respect its nature, as in the situation where we have to describe the relationships between several ordinal variables.

In most cases the quantification of a qualitative variable involves the assignment of a single numerical value to each category. F.W. Young and his team have attracted notice to the difference between the underlying process and the measurement level : for instance a phenomenon may be continuous (wave-length for colour perception) and the observation discrete (colour). So why quantify a category by a single value ? A more general quantification implies that a category may be represented by an interval of values, the different intervals being a partition of an interval of $\mathbb{R}$.

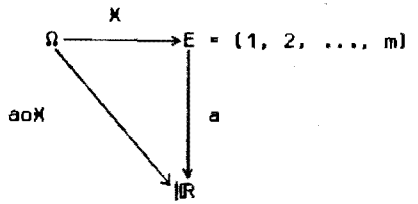
For ordinal measurements associated to continuous process there is in addition a constraint to order for the intervals.

We may notice that we are looking in fact for a quantification of observations more than of categories. Apart of the techniques proposed by Young, De Leeuw and Takane, quantification with intervals is not widely used and we will not refer to it in the remaining of this paper.

b) Mathematical formulation of quantification.

Let X be a qualitative variable with m categories and E the set of its categories.

If Ω is the usual universe, X is a mapping of Ω onto E.
A quantification of X is defined by a mapping a of E onto $\mathbb{R}$, the quantified variable being a o X :



The variable aoX takes only at most m values $a_1, a_2, \ldots a_m$ corresponding to the m categories of X.

If we introduce the m following indicator variables of the categories Π_j $j = 1,2,\ldots,m$ such that :

$$\Pi_j(\omega) = \begin{cases} 1 \text{ if } X(\omega) = j \\ 0 \text{ if not} \end{cases} \quad \text{for } \omega \in \Omega$$

we have the elementary but capital result : the quantified

variable aoX is nothing else than the linear combination of the indicator variables defined by the scores a_j :

$$aoX = \sum_{j=1}^{m} a_j \Pi_j$$

Therefore the quantification of variables belongs to the field of linear multidimensional analysis.

If there is no restriction upon the scores a_j (purely nominal variables) the set of numerical variables which are a quantification of X is a closed subset of dimension m of $L^2(\Omega)$, the vector subspace spanned by the Π_j.

If X is an ordinal variable with the natural order upon its categories a quantification of X must verify $a_1 \leq a_2 \leq .. \leq a_m$.

This set of constraints may be written as follows with the non negative numbers $b_0, b_1, .., b_m$

$$\begin{cases} a_1 = b_1 - b_0 \\ a_2 = b_1 + b_2 - b_0 \\ a_m = b_1 + b_2 + .. + b_m - b_0 \end{cases}$$

The quantified variable a . X is then equal to :

$$\sum_{j=1}^{m} a_j \Pi_j = \sum_{j=1}^{m} (b_1 + b_2 + .. + b_j - b_0) \Pi_j$$

$$= \sum_{j=0}^{m} b_j \tilde{\Pi}_j \quad \text{with } b_j > 0$$

where $\tilde{\Pi}_j = \sum_{i \geq j} \Pi_i$ and $\tilde{\Pi}_0 = -1$.

The Π_j are the "indicator variables" of the order (Bouroche, Dupont-Gatelmand, Tenenhaus) in the sense where :

$$\tilde{\Pi}_j(\omega) = \begin{cases} 0 \text{ if } X(\omega) < j \\ 1 \text{ if } X(\omega) \geq j \end{cases} \qquad j = 1, 2, .., m.$$

The set of all possible quantifications of X with the order restriction is thus the polyedric convex cone C spanned by the variables $\tilde{\Pi}_j$.

$$C = \{ x^* \mid x^* = \Sigma b_j \tilde{\Pi}_j \quad b_j \geq 0 \}$$

If the variable X has been observed upon n individuals we have the following matrix representation of the data and of the quantification :

for a purely nominal variable X may be represented as the matrix with n rows and m columns of the indicator variables.

$X = \begin{pmatrix} 0100 \\ 1000 \\ .... \\ .... \end{pmatrix}$ A numerical variable $\underline{x}^*$ obtained by a quantification of X

is expressed by $\underline{x}^* = X\,\underline{a}$ where $\underline{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$ is the vector of the scores of the categories

The set of all quantified variables is W, the subspace of $\mathbb{R}^n$ of dimension m defined by :

$$W = \{ \underline{x}^* \mid \underline{x}^* = X\underline{a} \quad \underline{a} \in \mathbb{R}^m \}$$

For an ordinal variable X with 3 categories we have for instance for the following five individuals and $a_1 \leq a_2 \leq a_3$

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} - & 1 & 1 & 0 & 0 \\ - & 1 & 1 & 1 & 0 \\ - & 1 & 1 & 1 & 1 \\ - & 1 & 1 & 0 & 0 \\ - & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} b_1 - b_0 \\ b_1 + b_2 - b_0 \\ b_1 + b_2 + b_3 - b_0 \\ b_1 - b_0 \\ b_1 + b_2 - b_0 \end{pmatrix}$$

The set of all possible $\underline{x}^*$ is the cone C defined by

$$C = \{ \underline{x}^* \mid \underline{x}^* = \tilde{X}\,\underline{b} \quad b_j \geq 0 \}$$

Frequently one has to deal only with zero-mean variables : if $\underline{1}$ is the constant variable which components are all equal to 1, the set of all possible $\underline{x}^*$ reduce to $W \cap \underline{1}^\perp$ or $\mathbb{I} \cap \underline{1}^\perp$ where $\underline{1}^\perp$ is the vector subspace orthogonal for $\underline{1}$ (i.e. made of zero-mean variables).

For nominal variables the equivalence between a quantification and a linear combination of indicator variables shows that the study of relationships between a set of quantified variables comes down to canonical analysis problems for it is nothing else than the study of linear relations between sets of numerical variables (taking only values 0 or 1).

c) Uni or multidimensional quantification ?

As the set of all possible quantifications of a qualitative variable is not of dimension one, it may be necessary to use several quantifications in order to describe a single qualitative variable. Thus it is not obvious that a nominal variable may be reduced to a single dimension ; correspondence analysis of contingency tables is an example of such a situation : we get two sequences of quantifications for taking into account the dependency between the two cross-classified variables.

In many cases, however, the complexity of a multidimensional quantification may be considered as a drawback and the practitioner would like to satisfy himself with a single quantification but this leads to some difficulties, as we will see below, except for two situations. The first one is that of prediction problems where there is only one best possible quantification and the second one is that of ordinal variables in descriptive studies : multidimensional quantifications are generally obtained by search of orthogonal basis (uncorrelated variables). But, if we have a total order upon the categories of X, the admissible quantifications of zero mean belong to the cone $\mathcal{C} \cap \underline{1}^\perp$: when a variable x belongs to this cone, the variables which are orthogonal to x do not belong to the cone and do not fill the order restrictions.

d) "Optimal" quantification.

In order to quantify qualitative variables we thus have to respect the nature of the variables but there remains an infinite number of possible quantifications : quantification is meaningful only if we have a precise aim which consists generally in the maximisation of some criterium of fit. For

instance if we deal only with two nominal variables, it seems natural that the two quantified variables should be maximally correlated what allows the best prediction of one by the other in the least-square sense.

In the same way if we have to predict one variable (qualitative or not) by several variables which may also be qualitative or not, there is a natural criterium of quantification: the maximisation of the square correlation between the dependent (possibly quantified) and a linear combination of the (possibly quantified) explanatory ones.

But if we have to quantify simultaneously more than two nominal variables without an external dependent variable, there is no unique criterium and there will be many "optimal" quantifications as we will see in the next part.

II. SIMULTANEOUS QUANTIFICATION OF SEVERAL QUALITATIVE VARIABLES

1) The case of two variables

Analyzing contingency tables by means of a quantification of the margins is very classical and leads to an unique solution, whatever is the criterium, that one gives by correspondence analysis or by canonical analysis.

Without discussing about rather uninteresting questions of anteriority, it seems instructive to remind briefly of the different points of view appeared along the years.

For more details the reader may refer to publications of Benzecri or De Leeuw.

In 1935 Hirschfeld, remarking that the analysis of the correlation between two variables gives the best results if the two lines of regression are linear, settles and solves the following problem :

"Given a discontinuous distribution, is it always possible to introduce new values for the variates such as both regressions are linear ?"

He then proves that for an array of size (m,p), there exists $\min[(p-1) ; (m-1)]$ orthogonal solutions and gives the equality between the Pearson's $\emptyset^2$ and the sum of squares of correlation between the pairs of quantified variables.

This result remained apparently ignored during more than twenty years until the work of Lancaster (1958) who generalized it to continuous variables and made the synthesis of the papers of Fisher (1940) Maung (1941) and Williams (1952).

Fisher introduced the quantification in order to apply his linear discriminant function when the predictor is itself nominal and settled the equivalence with Hotelling's canonical analysis. His algorithm to get the simultaneous quantification is interesting by its closeness to Hirschfeld's problem and to the alternate least squares method of Young, De Leeuw and Takane : in a first step we give an arbitrary quantification to one variable, then the second variable is quantified according to the means of the first in each of its categories. The first variable is now quantified again according to this principle of "reciprocal averaging" (Hill), apart from a factor of normalization to avoid degenerate solution, and so on until convergence (due to the fact that at the optimum the two quantified variables have both linear regressions).

Formally the solution is given by the canonical analysis of the two sets of indicator variables X_1 and X_2 : the quantified variables are the canonical variables and the optimal scores are given by the eigenvectors of the products of the two arrays of conditional frequencies.

2) The case of p nominal variables

The simultaneous quantification of more than two nominal variables have as many solutions as criteria, on the contrary of the case where p=2 where all criteria were equivalent. This is due to the fact that there does not exist a single measure of correlation between more than two variables.

However there are two main kinds of quantifying p variables, leading both to relative simple computations. The first comes down to an eigenvector problem and provides multidimensional quantifications of the variables, the second looking for a single quantification of each variable such as we get an optimal representation of the set of individuals upon a subspace of fixed dimension.

We will need the following matrices :

$X = (X_1 \mid X_2 \mid \ldots \mid X_p)$ disjunctive array with n rows and Σm_i columns of the indicator variables of the categories of all the X_i.

$X^* = \begin{pmatrix} x_1^* & x_2^* & \ldots & x_p^* \\ | & | & & | \end{pmatrix}$ array of size (n,p) of the quantified variables.

The so-called Burt's matrix :

$B = X'X = \begin{pmatrix} C_{11} & C_{12} & \ldots & C_{1p} \\ C_{21} & C_{22} & \ldots & C_{2p} \\ \vdots & & & \\ C_{p1} & \ldots\ldots & & C_{pp} \end{pmatrix}$ where C_{ij} is the contingency table between X_i and X_j ; i.e. $C_{ij} = X_i' X_j$.

B is a square-symmetrical matrix of size Σm_i.

D is the diagonal matrix with the same diagonal terms as B.

a) From Guttman to Benzecri : or the long story of multiple correspondence analysis.

Following L.L. Guttman, various authors (Horst, Lord, Bock, Nishisato, de Leeuw ...) have introduced the problem according to the following formulation : "Find the quantification of $X_1, X_2, \ldots X_p$ such as the measures of the p quantified variables $x_1^*, x_2^*, \ldots x_p^*$ be as homogeneous as possible for any individual and as scattered as possible between individuals". This is a very clear criterium from the classical point of view of factor analysis which comes down to an optimal analysis of variance : we have to maximize the dispersion of the means of each row of X^* in relation to the total variance of X^*.

If the x_j^* are of zero mean, we have :

$$\sum_{i=1}^{n} \sum_{j=1}^{p} (x_j^*(i))^2 = p \sum_{i=1}^{n} (\bar{x}^*(i))^2 + \sum_{i=1}^{n} \sum_{j=1}^{p} (x_j^*(i) - \bar{x}^*(i))^2$$

where $\bar{x}^*(i) = \frac{1}{p} \sum_{j=1}^{p} x_j^*(i)$.

As $\underline{x}_j^* = X_j \underline{a}_j$, if $\underline{a}$ is the supervector of all scores $\underline{a} = \begin{pmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \underline{a}_p \end{pmatrix}$ we may write again this analysis of variance equation in matrix notations, the lefthand side is equal to $\underline{a}'D\underline{a}$

and $\Sigma((\bar{x}^*(i))^2$ as $\frac{1}{p^2}\ a'\ Ba$. Thus the problem is the maximisation of $\eta^2 = \frac{a'Ba}{p\ a'Da}$

The solution is given by the eigenvector a related to the first non extraneous eigenvalue λ_1 of $\frac{1}{p} D^{-1}B$.

This is the same eigen-equation as that of correspondence analysis of X and that of Carroll's generalized canonical analysis of the p sets of indicator variables (see Bouroche-Saporta-Tenenhaus). Furthermore, as Guttman quoted it, the variable $\bar{x}^*$ is even the first principal component of the x_j^*, hence the name of "principal component of scale". In other words we have a quantification x^* of each X such as the first eigenvalue of the correlation matrix of the x^* is maximal. (For an historical survey and other criteria leading to the same result see Tenenhaus 1981).

To the other eigenvalues of $\frac{1}{p} D^{-1}B$ correspond different principal components each one corresponding to a <u>different</u> quantification of the X s : so we get actually a multidimensional quantification of the nominal variables, but the quantifications of various orders of a variable are non uncorrelated ; we have only the following mean and weak orthogonality properties between quantification of order k and l :

$$\text{cov}\left(\sum_{j=1}^{p} x_j^{*(k)}\ ,\ \sum_{j=1}^{p} x_j^{*(l)}\right)=0$$

$$\sum_{j=1}^{p} \text{cov}\left(x_j^{*(k)}\ ,\ x_j^{*(l)}\right) = 0$$

(see Saporta 1979).

There are many computer programs performing that kind of quantification, the best known being MULTM by Lebart-Morineau-Tabard, HOMALS by De Leeuw and Van Rijckevorsel and OPSCAL by Nishisato.

b) Obtaining a single quantification.

The problem here may be considered as a modification of the "principal component of scale" property of the above method:

we look for a quantification of the variables such as we get the best principal component analysis of the quantified variables i.e. that the sum of the variances of the first k principal components be maximised.

If k = 1 we obtain again the solution of multiple correspondence analysis i.e. the principal component of scale associated to the first eigenvalue of $\frac{1}{p} D^{-1}B$. If k > 1. The problem may be formalised simply in geometrical terms : we try to force the set of individuals to be in a k-dimensional subspace.

As we have no longer an eigenvector problem if k > 1, the solution is obtained by iterative algorithms. The three best known computer programs use alternating least squares methods and are chronologically PRINCIPALS (Young, De Leeuw, Takane), PRINQUAL (Tenenhaus), PRINCALS (De Leeuw, Van Rijckevorsel).

PRINQUAL processes only nominal variables while the two others (PRINCALS supersedes PRINCIPALS) accept ordinal variables and missing data (for PRINCALS).

Compared to the first method of quantification (Guttman-Hayashi-Benzecri..) this approach has the following advantages :

- the configuration of the individuals upon the k-dimensional subspace is optimal in the meaning of explained variance while the principal components of scale do not fulfill this condition. The first component of scale is the first principal component of the quantified variables, but the following components are not principal components for the same quantification.

- It is possible to compute correlation coefficient between quantified variables and principal components for instance, for we have here representation of the variables itself and not only of the categories.

- The method is well fit to ordinal variables for, as we have already seen, a single quantification is necessary. Conversely there are some drawbacks :

- The most important being in our sense that the quantification changes with the number of desired principal components and that the configuration of individuals is not the same while in the first method the coordinates of the individuals along the first k principal axis remain unaltered when we need further axis.

. If k increases, the solution becomes undetermined for k = p

. If there are two variables the second method does not come down to canonical analysis as it would have been desirable.

3) Further considerations

a) Extension to an infinite number of variables

The first method of quantification has been generalized (Saporta-Deville 1979) to nominal stochastic processes i.e. in the situation where one knows the evolution of a categorical variable (for instance hierarchical status or place of living) through the time between two instants 0 and T. The method comes down basically to a quantification of the nominal process X_t into a numerical one X_t such as the Karhunen-Loeve decomposition (P.C.A. of processes) of the standardized process X_t/δ_t gives a maximal first eigenvalue for the covariance operator.

b) Mixture of nominal and numerical variables.

It is very easy to generalize P.C.A. to that situation by quantifying the nominal variables.

Suppose that we have a set of numerical variables $y_1 \ldots y_q$ and p nominal variables X_1, X_2, ... X_p. To obtain the best P.C.A. of all variables in the sense of the first method of quantification (λ_1 maximal) we just have to perform a P.C.A. upon the array : $(Y \mid X_1 \mid X_2 \mid \ldots \mid X_p)$

with the metric $M = \begin{pmatrix} \begin{matrix} s_1^2 & & \\ & \ddots & \\ & & s_q^2 \end{matrix} & 0 & 0 \\ 0 & X_1' X_1 & 0 \\ 0 & 0 & X_p' X_p \end{pmatrix}^{-1}$

Again in the first method of quantification the following components will be associated to other quantifications of the X_s which may be uninteresting. PRINCIPALS, PRINQUAL and PRINCIPALS give single quantification and accept numerical variables.

III. QUANTIFICATION WITH DEPENDENT VARIABLES (Optimal scaling for prediction)

The problem is now to quantify qualitative variables in order to predict an external variable or criterium which may itself be nominal or ordinal.

Let X_1, X_2, ... X_p be the explanatory variables and Y the criterium, the general problem is thus to obtain a quantification x_1^*, x_2^*, ... x_p^* and y^* such that the square multiple correlation between y^* and the x_j^* be maximised.

Differences between algorithms are due essentially to the number and the nature of variables.

We will briefly give a few indications about the main situations.

1) Numerical criterium : regression with qualitative predictors

If all the explanatory variables are nominal we have here a particular case of the linear model which can be described as the estimation of effects in an analysis of variance with p factors of variability without interaction.

Mathematically this comes down to the multiple regression of the dependent variable y over the set of indicator variables of all the categories of the X s.

As the sum of indicator variables of the categories of each X is equal to 1, the problem is not of full rank and there is an infinite number of equivalent quantifications leading to the same prediction.

To get a solution we need linear constraints : the most usual being that the quantified variables should be of zero-mean. It is worth noting here that solution with zero-mean quantification may be obtained by regressing y onto all the $(\sum m_j - p)$ principal components of the multiple correspondence analysis of the X's. An approximate solution is then given by the regression of y onto the first k principal components (or onto the k components best correlated with y).

If one or several explanatory variables are ordinal we get a problem of monotonic regression to obtain quantifications respecting the order.

For a single ordinal X, the Kruskal's algorithm is well known. For several variables computer programs ADDALS and MORALS (Young, De Leeuw, Takane) provide good solutions using an alternating least squares algorithm (quantification, regression, new quantification and so on).

2) Nominal criterium, discrimination with qualitative predictors

This situation is very common in many applications such as credit-scoring or risk-evaluation where we have to predict the belonging of an individual to some group (good or bad behaviour for instance) with qualitative variables.

Like in regression, we cannot use without caution the classical procedures of discriminant analysis with indicator variables for they are linearly dependent and one usually need constraints of zero-mean quantification.

Let us remark that as early as in 1952, Hayashi programmed a method for nominal discrimination with that constraint.

It is possible to use MORALS and ADDALS for both accept a nominal criterium.

Another computer program DISQUAL (Saporta 1977) performs a quasi-optimal quantification of purely nominal variables. Its principle consists in selecting a subset of principal components of the correspondence analysis of the X's (in other words a multidimensional quantification of the X's) and then in a linear discriminant analysis upon that numerical components. Each linear discriminant function is a linear combination of the multidimensional quantifications of X's and gives thus an unique quantification (discriminant scores) of the X's.

The reader will find in appendix an example of application of DISQUAL.

3) Ordinal criterium

In that situation the quantification of the criterium must satisfy the order constraints.

The case of nominal predictors was first solved by Kruskal with his algorithm called MONANOVA.

Computer programs ADDALS and MORALS are also well fitted to this kind of problem and accept predictors of ordinal kind.

REFERENCES

BARLOW, R.E., BARTHOLOMEW, D.J., BREMMER, J.M., BRUNK, H.D., Statistical inference under oder restrictions,Willey,New York,1972.

BENZECRI J.P., L'analyse des données, Dunod Paris, 1979.

BENZECRI J.P., Histoire et préhistoire de l'analyse des données, CAD vol. I, 1976.

BOUROCHE J.M., DUPONT-GATELMAND C, TENENHAUS M., L'analyse canonique des préférences, in Data Analysis and Informatics, E. Diday (ed.), 631-649, North-Holland, 1980.

BOUROCHE J.M., SAPORTA G., TENENHAUS M., Generalized canonical analysis of qualitative data - U.S.-Japan, Seminar on Theory, Methods and Applications of Multidimensional Scaling, 1975.

CAILLIEZ F., PAGES J.P., Introduction à l'analyse des données, SMASH, 1976.

CARROLL J.D., Categorical conjoint measurement - Bell Telephone Laboratories

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

DAUXOIS J., POUSSE A., K-analyses canoniques Publ. n° 3, Laboratoire de Statistique, Université Paul Sabatier, Toulouse, 1975.

DEHEDIN J., Discrimination sur variables qualitatives, Thèse 3ème cycle, Université Paris VI, 1975.

DE LEEUW, J., Canonical analysis of categorical data, 1973, University of Leiden.

DE LEEUW, J., VAN RIJCKEVORSEL, J., Homals and Princals in Data Analysis and Informatics, E. Diday, ed., 231-242, North Holland,1980.

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

DROUET d'AUBIGNY G., Description statistique des données ordinales : analyse multidimensionnelle, Thèse de 3ème cycle, Grenoble 1975.

GUTTMAN L., 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 1941.

GUTTMAN L., The principal components of scale analysis, in Stouffer Ed. Measurement and Prediction, Princeton University, 1950.

HAYASHI C., On the quantification of qualitative data from the mathematico-statistical point of view, Annals Inst. Stat. Math. 2, N° 1, 1950.

HAYASHI C., On the prediction of phenomena from qualitative data and quantification of qualitative data from the mathematico-statistical point of view - Annals Inst. Stat. Math. 3, N° 2, 1952.

HILL M.O., Reciprocal averaging, J. Ecol. 61, 237-251, 1973

HILL M.O., Correspondence analysis : a neglected multivariate method, Appl. Stat. 23 n° 3, 304-354, 1974.

HIRSCHFELD H.O., A connection between correlation and contingency, Proc. Camb. Phil. Soc. 31, 520-524, 1935.

HORST P., Obtaining a composite measure from a number of different measures of the same attribute, Psychometrika 1, 53-60, 1936.

HORST P., Relations among m sets of measures, Psychometrika 26, 129-149, 1961.

KETTENRING R.J., Canonical analysis of several sets of variables, Biometrika 58, 433-451, 1971.

KRUSKAL J.B., Nonmetric multidimensional scaling, Psychometrika 29, 115-129, 1964.

KRUSKAL J.B., Analysis of factorial experiments by estimating monotone transformation of the data, JRSS B 27, 251-263, 1965.

LANCASTER H.O., The structure of bivariate distribution, A.M.S. 2, 719-736, 1958.

LEBART L., MORINEAU A., TABARD N., Traitement des données statistiques, Dunod 1979.

LORD F.M., Some relations between Guttman's principal components of scale analysis and other psychometric theory, Psychometrika 23, 291-296, 1958.

MAC DONALD R.P., A unified treatment of the weighting problem, Psychometrika 33, 351-381, 1968.

MASSON M., Processus linéaire et analyse des données non linéaire, Thèse d'Etat, Université de Paris VI, 1974.

MAUNG K., Measurement of association in a contingency table with spacial reference to the pigmentation of hair and eyes, Annals of Eugenics , 11, 189-205, 1941.

NISHISATO S., Dual scaling and its application, University of Toronto Press, 1980.

SAPORTA G., Liaison entre plusieurs ensembles de variables et codages de données qualitatives, Thèse de 3ème cycle, Université de Paris VI, 1975

SAPORTA G., Discrimination when all the variables are nominal : a stepwise method, Spring Meeting of the Psychometric Society, Manay Hill, NJ, 1976.

SAPORTA G., Disqual une méthode et un programme d'analyse discriminante sur variables qualitatives, Journées internationales Analyse des Données et Informatique, INRIA 1977.

SAPORTA G., About some remarkable properties of Carroll's generalized canonical analysis, European Meeting of the Psychometric Society, Groningen NL 1979.

SAPORTA G., Méthodes exploratoires d'analyses de données temporelles, Thèse Doctorat es Sciences, Université Paris VI, 1981.

TENENHAUS M., Analyse en composantes principales de variables qualitatives, Publ. N° 01-81 du Laboratoire de Statistique, Université Paul Sabatier, Toulouse 1981.

TORGERSON W.S., Theory and methods of scaling, Wiley, 1958.

WILLIAMS E.J., Use of scores for the analysis of association in contingency tables, Biometrika 39, 274-289, 1952.

WOLD S., Q Q regression : A. NIPALS procedure for regression with qualitative and quantitative variables. University of Umea, Sweden.

YOUNG F.W., DE LEEUW J., TAKANE Y., Multiple and canonical regression with a mix of qualitative and quantitative variables : an alternating least squares method with optimal scaling features. Psychometric Laboratory, University of North Carolina, 1975.

YOUNG F.W., DE LEEUW J., TAKANE Y., Additive structure in qualitative data, Psychometric Laboratory, University of North Carolina, 1975.

YOUNG F.W., DE LEEUW J., TAKANE Y., How to use PRINCIPALS, Psychometric Laboratory, University of North Carolina, 1975.

YOUNG F.W., DE LEEUW J., TAKANE Y., Quantifying qualitative data in M. Feger (ed.), Similarity of Choice,Academic Press, 1979.

YOUNG F.W.,TENENHAUS M., Multiple correspondence analysis and the principal components of qualitative data, Psychometrika (to be published).

APPENDIX

An example of discrimination with nominal predictors by DISQUAL : risk estimation of accident for young drivers

5814 customers of less than 25 years or having their driving licence from less than 2 years have been selected from the files of an important insurance company.

3849 were good drivers, i.e. without accident during the past three years.

2325 were bad drivers, i.e. responsible of three or more accidents during the same three years or of more than two the first year.

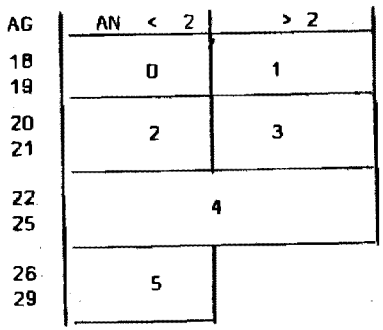
The true proportion of bad drivers was 4%, but in order to get good estimates this category was overrepresented in the sample.

After primary statistical investigations 11 explanatory variables were selected

FRAC	kind of payment (yearly, each 6 months, each 3 months)
AG	age at the date of licence (7 categories)
AN	number of years since the licence (less than 2 years or more than 2 years)
SEXE	(male, female)
MATR	matrimonial status (bachelor or other)
ZONT	area of tariff (3 categories)
USAG	type of use of the car (9 categories)
CLGR	tariff group (5 categories)
RAFR	special payment for a policy without deduction (yes, no)
CRED	car bought with a credit or not
CAPD	life insured or not.

Among these variables some may present interactions concerning the category of drivers (good or bad) and it is necessary to create crossed variables before a linear (additive formula) discrimination.

A systematic study of interactions by means of a log-linear model showed that there was a significant interaction between AG and AN. It was thus necessary to create a new variable AGAN,replacing the two preceding ones according to the following diagram :



There is now 10 predictors with a total number of categories 36. A multiple correspondence analysis will provide 36-10 = 26 non-trivial components.

Here is the output of DISQUAL

Results of diagonalisation

Non trivial eigenvalues :

1.897 1.529 1.312 1.126 1.119 1.103 1.092 1.046 1.026 1.022
1.013 1.002 0.484 0.981 0.960 0.915 0.933 0.932 0.912 0.902
0.860 0.756 0.703 0.678 0.615 0.541

Number of factors to be selected (amount of accounted variance superior to 1%) = 26.

Selection of factors

Rank	Number	Relative discriminant power	% of variance accounted
1	1	0.624	0.073
2	8	0.101	0.029
3	2	0.077	0.059
4	10	0.035	0.043
5	11	0.032	0.043
6	6	0.019	0.026
7	12	0.018	0.042
8	4	0.014	0.050
9	16	0.013	0.040
10	5	0.011	0.024
11	15	0.010	0.035
12	26	0.010	0.039
13	22	0.010	0.039
14	24	0.005	0.038
15	7	0.005	0.027

Number of selected factors	15
Cumulated % of discriminant power	98,22
Cumulated % of variance	60,75

Scoring rule

FRAC	0	3.68 (yearly)
	1	0.50 (each 6 months)
	2	-4,75 (each 3 months)
AGAN	0	-1.73
	1	4.40
	2	-1.25
	3	3.51
	4	0.40
	5	-4.83
SEXE	0	-1.56 (male)
	1	5.50 (female)
MATR	0	-1.01 (bachelor)
	1	2.53 (other)
ZONT	0	3.24
	1	-1.06
	2	-3.85

USAG	0	-7.05
	1	0.57
	2	1.72
	3	0.91
	4	-1.62
	5	0.60
	6	-8.64
	7	0.99
	8	-7.68
CLGR	0	3.88
	1	-1.48
	2	-4.58
	3	-7.20
	4	-6.10
RAFR	0	-6.80 (yes)
	1	10.20 (no)
CRED	0	0.09 (yes)
	1	-2.75 (no)
CAPD	0	0.17 (yes)
	1	-4.35 (no)

Threshold value -4.27.

We use the scoring rules as follows :

Let us take a male driver who wants to pay each 3 months if he is bachelor and of category 2 of variable AGAN (20 or 21 years old and licence from less than 2 years) living in ZONT number 0 for the USAG category 5, the CLGR category 3, taking a no-deduction policy (RAFR = 0) with a car bought with a credit and without a life insurance. His score will be

-4.75 -1.25 -1.56 -1.01 +3.24 +0.60
-7.20 -6.80 +0.09 -4.35 =-12.26

As this score is less than the threshold value -4.27, he will be classified into the "bad drivers" population.

The efficiency of this scoring rule is described by the following table obtained by classifying the individuals of the sample :

true group	classified in good	bad
good	2830	659
bad	518	1807

There is an amount of about 80% of correct classifications and 78% of bad drivers are detected.