Simultaneous Analysis of Qualitative and Quantitative Data

Analisi simultanea di dati qualitativi e quantitativi

Gilbert Saporta
Conservatoire National des Arts et Métiers
Paris

Estratto da:

SOCIETÀ ITALIANA DI STATISTICA
ATTI DELLA XXXV RIUNIONE SCIENTIFICA
Padova, 18 - 21 aprile 1990 - Volume I

PADOVA
CEDAM - CASA EDITRICE DOTT. ANTONIO MILANI
1990

Simultaneous Analysis of Qualitative and Quantitative Data

Analisi simultanea di dati qualitativi e quantitativi

Gilbert Saporta
Conservatoire National des Arts et Métiers
Paris

1. Description of mixed data

The basic technique for describing relationships between numerical variables is principal component analysis (PCA), which provides a description of the observations and of the variables in a low dimensional euclidian space.

We will present five generalizations of PCA to handle nominal and numerical data.

1.1. Eigenanalysis of a matrix of correlation coefficients

Since ordinary PCA consists in finding the eigenvalues and the eigenvectors of the correlation matrix between p numerical variables, one solution when we have both categorical variables and numerical variables may consist in defining a matrix of correlation coefficients between variables of different kinds.

To do this we need measures of relationships between two categorical variables and between a categorical and a numerical variable which would have the same interpretation as the usual product-moment correlation, and lead to semi-definite positive matrix of coefficients.

Since negative correlation is meaningless when a categorical variable is involved, the coefficients which we may use are generally homogeneous to squared correlation and not to correlation.

For a couple of categorical variables, the coefficient will be a function of the chi-square, and for a couple between a categorical and a numerical variable, the coefficient will be a function of the correlation ratio η.

The RV coefficients proposed by Escoufier for measuring the relationships between vector-valued variables give an elegant solution to our problem if we identify a categorical variable with k categories to the set of the k indicator variables of its categories.

When comparing two data matrices X_1 and X_2 with k_1 and k_2 variables, RV is defined as

$$RV = \frac{\mathrm{Trace}(X_1 M_1 X_1' \quad X_2 M_2 X_2')}{\sqrt{\mathrm{Trace}(X_1 M_1 X_1')^2 \quad \mathrm{Trace}(X_2 M_2 X_2')^2}}$$

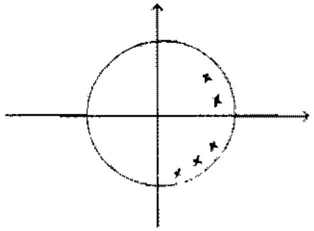
Where M_1, M_2 are matrices associated to the two data sets.

If we choose $M_1 = (X_1' X_1)^{-1}$, the RV coefficient is a function of the canonical correlations between X_1 and X_2.

It is straightforward to prove that with zero-mean variables, the RV between two numerical variables is equal to the squared correlation coefficient r^2, the RV between a k-categorical variable and a numerical variable is equal to $\frac{\eta^2}{\sqrt{k-1}}$ and the RV coefficient between two categorical variables is equal to $\frac{\phi^2}{\sqrt{(k_1-1)(k_2-1)}} = T^2$ (Tschuprow's coefficient) see Saporta (1976).

So, the analysis of the relationships between p variables of various kinds may be performed by diagonalizing the matrix with elements r^2, $\frac{\eta^2}{\sqrt{k-1}}$, T^2.

This matrix is sdp. Since all its coefficients are positive, the first eigenvector has all its elements of the same sign and the correlation circle is actually a half-circle:



Of course this kind of analysis provides only a mapping of the proximities between variables, but the eigenvectors cannot be used to get a satisfactory representation of the individuals since it is not possible to define a linear combination of categorical and numerical variables.

However in the STATIS method, we obtain a linear combination of scalar product matrices between individuals

$$\tilde{W} = \sum_{i=1}^{p} \alpha_i X_i X_i' \;, \quad \text{the "compromise"}$$

which in turn may be analysed and provide a representation of units.

Here the α_i are the components of the first eigenvector of the matrix of RV coefficients. But $\tilde{W}$ does not use the whole information (i.e. all the correlation structure).

1.2. Multidimensional scaling of similarity coefficients

One other important drawback of the last method is that it is not correct to compare T^2 or η^2 coefficients when the number of categories of the nominal variables are not identical.

It is well known that the chi-square measure of independence is a monotonic function (in a probabilistic sense) of the degrees of freedom: higher is the number of categories, higher is the chi-square.

The division by the square root of the degree of freedom does not make a full correction to this fact (its effect consists mainly in normalizing between 0 and 1 the chi-square).

Moreover, even if r^2, $\frac{\eta^2}{\sqrt{k-1}}$, T^2 are cosines of elements of some vector spaces, they have not the same distribution under the null hypothesis of independence which prevents a complete comparison.

For these reasons, we propose to use, as a measure of similarity between variables of different kinds, the probability of getting a value less than the correlation coefficient (r^2, η^2, T^2) under the hypothesis of independence. Since these similarities are probabilities they may be compared and the problem of the degrees of freedom vanishes.

There is no reason why the p×p matrix S of these similarity coefficients should be positive.

So we propose not to do an eigenanalysis, but to perform a multidimensional scaling to get a mapping of the variables.

However, like in any method of multidimensional scaling we need to fix the dimension of the representation since the solutions are not nested.

1.3. An extension of principal component and of multiple correspondence analysis

It is well known:

a) that in PCA of standardized variables the principal components maximize $\sum_j r^2(\underline{c},\underline{x}^j)$ where the $\underline{x}^j$ are the numerical data variables.

b) that in MCA the components maximize $\sum \eta^2(\underline{c};\chi^j)$ where the χ^j are categorical variables.

A natural extension of both PCA and MCA to a mixture of qualitative and quantitative variables consists in maximizing

$$\sum r^2(\underline{c};\underline{x}^j) + \sum \eta^2(\underline{c};\chi^j)$$

to derive generalized principal components (see Saporta (1988) or Tenenhaus (1977)).

This method provides a simple representation of individuals (the solutions are nested) and comes down to a PCA of the following matrix:

$$\left(\begin{array}{c||c||c|c} & 0\ 1\ 0 & & 1\ 0\ 0 \\ X & X_1 & & X_q \end{array}\right)$$

numerical variables — indicator matrices of categorical variables

with the metric:

$$\left(\begin{array}{cc|cc} 1/s_1^2 & 0 & & 0 \\ 0 & 1/s_p^2 & & \\ \hline & & 1/n_1 & 0 \\ 0 & & 0 & 1/n_m \end{array}\right)$$

which is the concatenation of the D_{1/s^2} metric and the chi-square metric.

Unfortunately this method does not give a satisfactory repreeentation for the variables:

The mapping of the variables with the η^2 and the r^2 does not lead to clear graphics since all the variables are in the first quadrant.

1.4. PCA with optimal scaling of the categorical variables

Following the works by Young (1981), Young, De Leeuw, Takane (1978), Tenenhaus (1977) this technique consists in transforming each categorical variable into a numerical variable by allotting numerical scores to the categories.

These scores are optimally calculated in order to get an optimal PCA according to some criterium; the most popular criterium being the amount of variance accounted for, by the first k eigenvalues of the correlation matrix.

The algorithm is usually of the alternating least square family (ALS).

Starting from an initial quantification of the categorical variables, a PCA is performed which gives k components c_j. Knowing these components, a set of projections onto the indicator variables of the categorical variables (first canonical variable between the components and the indicator variables matrix X_j) leads to a different quantification and so on.

The criterium $\sum_{i=1}^{k} \lambda_k = \sum_{i=1}^{k}\sum_{j=1}^{p} r^2(c_i;x^j) + \sum_{i=1}^{k}\sum_{j=1}^{q} r^2(c_i;X_j\underline{a}_j)$ is thus optimized over the c_i and the $\underline{a}_j$, because it increases at each step.

The Proc PRINQUAL of the release 6.03 of SAS-System is an implementation of this technique. In addition to the usual criterium of the sum of the first k eigenvalues, there are two other criteria (one is based on the minimization of det R) and various options to transform the numerical variables: (functional, splines, M-splines).

Since after the optimal transformations, this method is a standard PCA, the usual outputs may be produced: in particular one has correlation coefficients between transformed variables and numerical variables (principal components and variables of the data set). However a local and not global optimum may occurr, depending on the startup point.

An other drawback is that the solution depends on k, the number of components retained for the representation: solutions are not nested.

The robustness of this method may also be questionable and has not yet given raise to publications.

1.5. A variant of INDSCAL: INDOMIX

Recently Kiers (1989) proposed a method based on an application of the INDSCAL model of Carrol and Chang (1970) to a set of similarity matrices between the n observations.

Each similarity matrix corresponds to one of the variables. So it is necessary to define similarities between units according to the nature of the variables, categorical or numerical. Although one may think of the Gower's coefficients (1971), Indomix uses similarity matrices based on orthogonal projectors.

For some numerical variable x we get

$$s_{ii'} = \frac{1}{n} \frac{(x_i - \overline{x})\ (x_{i'} - \overline{x})}{s^2}$$

$S = \frac{1}{n} \underline{z}\ \underline{z}'$ where $\underline{z}$ is the vector of standardized values.

For a categorical variable we get

$$s_{ii'} = \begin{cases} 0 & \text{if } i,i' \text{ do not belong to the same category} \\ n/n_j & \text{if } i \text{ and } i' \text{ belong to category of frequency } n_j. \end{cases}$$

$S \ = X(X'X)^{-1}$ where X is the indicator matrix of the categories.

A normalization factor may be used here, $\|S^j\|=1$ for any variables j, in order to compare variables with different number of categories.

An INDSCAL analysis is then performed which gives a mapping of the individuals in a common space and a mapping of the variables according to the weights given to the underlying dimensions.

Of course some other variants are possible: such as blocking for instance the numerical variables in a single array X, or analyzing with a classical scaling technique the average matrix of the S^j or applying STATIS techniques, or any kind of three-way methods since the data may be considered as a set of matrices each one associated to a single variable, see for instance see Coppi, Bolasco (1989) and Lavit (1988).

One problem with methods such as Indomix is that they cannot handle a large amount of data, since like every multidimensional technique, scaling the critical dimension is the number of observations, not the number of variables.

1.6. Cluster analysis

Two approaches are feasible when want to cluster observations with both qualitative and quantitative descriptors.

1.6.1

The direct approach consists in defining a global similarity measure incorporating all the variables such as the sum of the Gower's similarity coefficients or the sum of similarity matrices used in Indomix. Once this global similarity matrix W is defined, any method of hierarchical clustering may be applied; furthermore with the former choices (Gower or Indomix) the matrix W is positive definite and may be considered as a matrix of scalar-product. It implies that method for euclidean data such as Dynamic Clustering or Ward's hierarchical method are applicable.

As in section 1.5 the difficulty here consists in defining measures of similarity for qualitative and quantitative data which may be compared and a correct way of aggregating these similarities.

1.6.2

We may also use an non-direct approach based on one of the generalizations of principal components analysis presented in the previous sections. We just have to perform a cluster analysis of the individuals described by their coordinates along the principal axes.

The use of cluster analysis with principal coordinates is a well established methodology (refer to SPAD-N software) but it is highly recommended to retain all the coordinates to have a complete recovery of the interindividuals distances: retaining only the 5 or 6 first principal axes, for instance, may lead to wrong conclusions; some particular groups of individuals may be revealed only in a high-dimension representation.

In this respect there is no problem to use the extension of PCA in section 1.3 because it is possible to have a full reconstitution of the data matrix with all the components. However the use of an optimal scaling method such as Prinqual may be subject to some questions: the solution of this kind of PCA relies heavily on the number of components choosed by the user; this number should be small to prevent indeterminacy or instability of the solution and there is no garantee to recover the data.

2. Explanatory methods with qualitative and quantitative predictors

This situation is better known and since there is an objective criterium, linked to the predictability of the dependent variable, the problems are rather different than in the case of component analysis.

2.1. Linear effects, interactions, reversal

If we restrict our topic to linear models (regression or discrimination according to the nature of the dependent variables) the main question concerns the type of influence of the categorical variables upon the structure of dependency.

a) The simplest case is of course the additive effect on the mean: the decision function is a linear combination of the numerical variables and of the indicator variables of the categories of the nominal variables.

b) When there are interaction between the categorical variables, one has only to insert in the previous model the indicator variables of the significant crossings of the categorical variables.

c) But the most problematic case occurs when the correlational structure of the numerical predictors is a function of the categories of the nominal variables. An extreme situation is when the signs of some correlation is changed according to the fact that an observation belongs to some category or to another: it is the reversal case. In this case, differents models have to be fitted.

In multiple regression there is no particular problem to handle case a) and b) which corresponds to models of covariance analysis.

Less attention has been given to the case c); a formal solution is given by separate regression for each group of individuals defined by combinations of categorical variables but the number of groups makes it generally unfeasible. A methodology derived from regression tree Breiman et al (1984), seems one of the way to solve this problem.

2.2. Discriminant analysis

When there is no risk of reversal, a linear discriminant analysis with optimal scaling of the categorical predictors may be performed: since it is equivalent to a discriminant analysis with numerical variables and indicator variables it does not present any difficulty. Like in the general linear model some constraints on the coefficient of the indicator variables are necessary, since they add-up to unity; the most usual constraint being to put a zero coefficient to the first (or the last) indicator of each nominal variable.

Logistic regression is an alternative method which is in favour by the econometricians; its superiority over discriminant analysis seems to be doubtful except when there is strong nonnormality of the numerical variables or strong difference between covariance matrices. (Efron 1975).

When the discriminant behaviour of the numerical variables differs according to the subgroups defined by the categorical variables ("reversal"), the location model developped by Olkin-Tate (1961) and Krzanowski (1975, 1980) may be very useful: this model assumes that the conditional distribution of the numerical variables X for each group G_i and for a fixed value of the categorical variable χ is normal with a mean $E(X/G_i,\chi) = m_{i,\chi}$ and a constant matrix of covariance Σ. $m_{i,\chi}$ is fitted with MANOVA model. The model may be completed by a log linear model for $P(\chi/G_i)$. The parameters are estimated by maximum likelihood. An implementation of this method is the program ADM by Daudin-Soukal (1989).

Due to the complexity of the method, this model is limited to a small number of variables.

Other proposals have been made such as using as predictors the products of the numerical variables by the indicator variables of the nominal predictors but it leads very quickly to a too large number of parameters.

3. Discussion

From this short overview of the problem, we may draw some conclusions.

For component analysis there are many solutions and the practitioner has to choose between them. If his purpose is only the study of the relationships between variables, multidimensional scaling of P-values seems to be the best choice. If the purpose is a mapping of units we would recommend a PCA with indicator variables. If we want both simultaneously, a compromise is necessary: such as an optimal quantification technique like Proc Prinqual or a scaling technique like Indomix. But we have to be cautious with the number of components retained; further comparison studies and sensitivity analysis are necessary.

In cluster analysis the main problem relies upon the definition of an adequate measure of similarity but this a common feature to all clustering techniques.

For explanatory problems, there is no difficulty when there are only linear effects and classical software is available. For more complex interaction effects, modelling is more difficult, but cross-validation techniques may provide goodness of fit criteria since there is usually a simple criterium to optimize (R^2 in regression, or error rate in discrimination for instance).

REFERENCES

BREIMAN L., FRIEDMAN J., OHLSEN, STONE (1984), "Classification and regression trees", Wadsworth.

CARROL J.D., CHANG J.J. (1970), "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition", Psychometrika, 35, 283-319.

COPPI R., BOLASCO S. Ed. (1989), "Multiway data analysis", North-Holland.

DAUDIN J.J., SOUKAL M. (1989), "Analyse discriminante sur variables continues et qualitatives, Notice du logiciel ADM", INAPG 16 rue Claude Bernard, 75005 Paris.

EFRON B. (1975), "The efficiency of logistic regression compared to normal discriminant analysis", JASA, 70, 892-898.

GOWER J.C. (1971), "A general coefficient of similarity", Biometrics, 27, 857-871.

KIERS H. (1989), "Three-way methods for the analysis of qualitative and quantitqtive data", DSWO Press.

KRZANOWSKI (1975), "Discrimination and classification using both binary and continuous variables", JASA, 70, 782-790.

KRZANOWSKI (1980), "Mixture of continuous and categorical variables in discriminant analysis", Biometrics, 36, 486-499.

LAVIT Ch. (1988), "Analyse conjointe de tableaux quantitatifs", Masson.

OLKIN, TATE (1961), "Multivariate correlation models with mixed discrete and continuous variables", AMS, 32, 448-465.

SAPORTA G. (1976), "Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives", Statistique et Analyse des Données, 1, 38-46.

SAPORTA G. (1983), "Multidimensional data analysis and quantification of categorical data", New Trends in Data Analysis and Applications, J. Janssen Ed., North-Holland, 73-97.

SAPORTA G. (1988), "About maximal association criteria in linear analysis and in cluster analysis", Classification and related techniques of data analysis, H. Bock Ed. North-Holland.

SAS Institute (1988), Additional SAS-Stat procedures release 6.03, Technical Report P 179.

TENENHAUS M. (1977), "Analyse en composantes principales d'un ensemble de variables nominales ou numériques", Revue de Statistique Appliquée, 25, 39-56.

YOUNG F.W. (1981), "Quantitative analysis of qualitative data", Psychometrika, 46, 357-388.

YOUNG F.W., TAKANE Y., DE LEEUW J. (1978), "The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features", Psychometrika, 43, 279-281.