Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 3255 – 3268

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data

Mónica Bécue-Bertaut[a,*], Jérôme Pagès[b]

[a]EIO, Universitat Politècnica de Catalunya, Edifici C5, C/Jordi Girona 1-3, 08034 Barcelona, Spain
[b]ENSA/INFSA, Rennes, France

Available online 29 September 2007

Abstract

Analysing and clustering units described by a mixture of sets of quantitative, categorical and frequency variables is a relevant challenge. Multiple factor analysis is extended to include these three types of variables in order to balance the influence of the different sets when a global distance between units is computed. Suitable coding is adopted to keep as close as possible to the approach offered by principal axes methods, that is, principal component analysis for quantitative sets, multiple correspondence analysis for categorical sets and correspondence analysis for frequency sets. In addition, the presence of frequency sets poses the problem of selecting the unit weighting, since this is fixed by the user (usually uniform) in principal component analysis and multiple correspondence analysis, but imposed by the table margin in correspondence analysis. The method's main steps are presented and illustrated by an example extracted from a survey that aimed to cluster respondents to a questionnaire that included both closed and open-ended questions.

Keywords: Mixed data; Textual data; Distance; Multiple factor analysis; Multiple factor analysis for contingency tables; Clustering; Survey

1. Introduction

In different studies, statistical units are simultaneously described by heterogeneous variables belonging to at least two of the following types: quantitative, categorical and frequency variables. A particular case, used as an application of the method that we propose, arises in surveys when a complex topic is tackled using both closed and open-ended questions; often, there are numerous closed questions gathered in several sets that can be quantitative and/or categorical, each one corresponding to a different topic.

Each set of quantitative variables is represented by a classical units × variables table, as in principal components analysis (PCA). Each set of categorical variables leads to a units × indicator variables table by associating an indicator variable with each category, as in multiple correspondence analysis (MCA). Each set of frequency variables, represented by a units × frequency variables table, corresponds to a list of related events whose occurrences are counted up in the sample; for each set, the total number of occurrences summed per unit (row) or event (column), as well as the grand total, are meaningful. In our example, frequency tables are derived from the open-ended questions: each of these

[*] Corresponding author. Tel.: +34 93 401 70 31; fax: +34 93 401 58 55.
E-mail address: monica.becue@upc.edu (M. Bécue-Bertaut).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.09.023

questions builds up a respondent-units × word-frequency variables table by counting the occurrences of the different words in every respondent answer. This kind of table is classically analysed using correspondence analysis (CA) (Lebart et al., 1998, Chapter 3). Thus, we have to deal with a global multiple table that juxtaposes sets of quantitative, indicator and frequency variables, i.e. heterogeneous (or mixed) data. However, within each set, the variable type is homogeneous.

To analyse and cluster units described by these kinds of data, the starting point is to define a global distance by combining the distances (called *separate* distances) issued from PCA (in the case of quantitative sets), MCA (in the case of categorical sets) and CA (in the case of frequency sets).

The problem of mixed data has already been studied. Gower (1971) first proposed a solution for balancing quantitative and categorical variables, whatever their type. Specific distances are used for categorical and quantitative variables; the range of variation of every distance is standardized to 1 before aggregation in a global distance. Moreover, by using a set of a priori weights, users can favour those variables that they consider to be important.

Podani (1999) extends Gower's general coefficient to ordinal variables. Grabmeier and Rudolph (2002) follow Gower's proposal to standardize the range of variation of any distance, but considering a larger range of distances; sets of variables are considered but only to group the variables depending on the associated distance. Balancing is performed at the variable level, in order to give the variables equal importance (or the importance decided by the user). Thus, neither the case of frequency variables nor balance the influence of different sets of variables is considered. Another possible approach is multiple factor analysis (MFA, Escofier and Pagès, 1988–1998, Chapters 7–9; 1994) which adopts a geometric approach and standardizes the highest axial inertia of every set of variables to 1 for balancing their importance.

We have already proposed an extension of MFA, in order to deal with multiple frequency or contingency tables (multiple factor analysis for contingency tables, MFACT; Bécue-Bertaut and Pagès, 1999, 2004). The fundamental aim of this paper is to extend MFA to deal with multiple tables in which quantitative, categorical and frequency variables are juxtaposed, and to show the interest of its use as a preprocessing step for clustering, following the proposals by Lebart (1994) and Chae and Warde (2006) to combine principal axes and clustering methods.

After presenting the notation (Section 2) and recalling the main principles of MFA and its extension to frequency tables (Section 3), in Section 4 we address the problem of unit weighting that arises when quantitative, categorical and frequency sets are simultaneously introduced. Section 5 sets out the properties of this MFA extension and Section 6 presents the clustering step, focusing on how to relate the variables to the final partition. An application of this extension to survey data (Section 7) illustrates the method and underlines the usefulness of open-ended questions for a better approach to complex topics in surveys.

2. Notation

I statistical units are described by J sets of variables: J_{q} sets of quantitative variables, J_{c} sets of categorical variables and J_{f} sets of frequency variables (Fig. 1). $J = J_{\mathrm{q}} + J_{\mathrm{c}} + J_{\mathrm{f}}$.

Below, the symbols I, J, J_{q}, J_{c}, J_{f}, K or K_j refer to both the set and its cardinal number.

Whatever the set type, the letter j refers to a set, the letter k refers to a column and K_j is the number of columns in set j. $K = \sum_{j \in J} K_j$ is the number of columns across all the sets.

A table $I \times K_j$ is associated with every set j. The J tables together make up a multiple or global table $I \times K$.

For a quantitative or frequency set j, K_j is both the number of columns and the number of variables.

For a categorical set j, with Q_j variables, K_j is both the number of columns and the number of categories summed across the Q_j variables. This kind of set is represented by a units × indicator variables table in which column k is associated with category k.

At the crossing of row i and column k (belonging to table j) we have:

- if j is a quantitative set, the value x_{ikj} of the variable k for the unit i;
- if j is a categorical set, $z_{ikj} = 1$ if i belongs to the category k and 0 if it does not;
- if j is a frequency set, the proportion f_{ikj}, computed as the ratio between the number of occurrences of event k (belonging to set j) for unit i and the grand total over the table that gathers all the J_{f} frequency tables; thus: $\sum_{j \in J_{\mathrm{f}}} \sum_{k \in K_j} \sum_{i \in I} f_{ikj} = 1$.

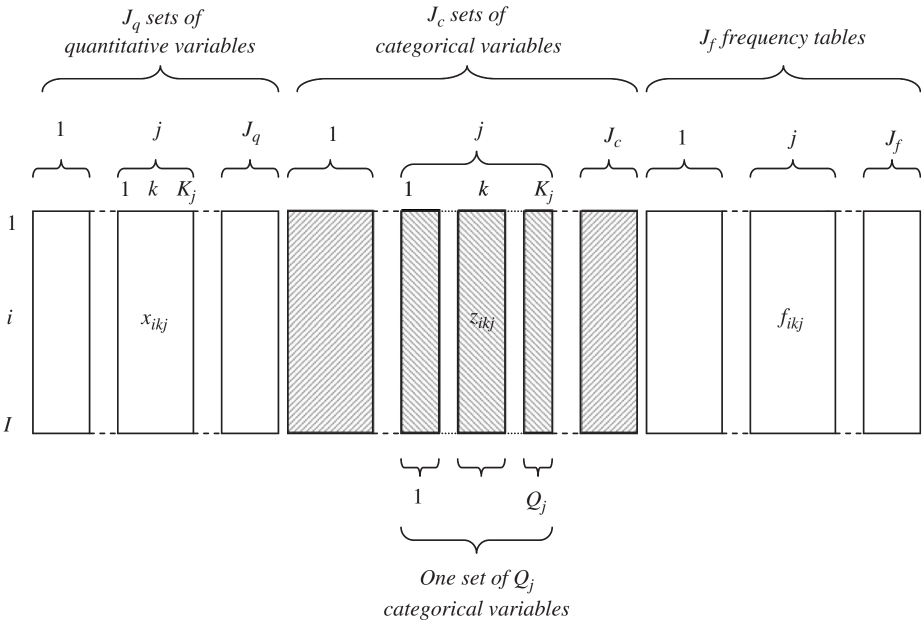

Fig. 1. Global table in which sets of quantitative, categorical and frequency variables are juxtaposed.

We use:

- $f_{i.j} = \sum_{k \in K_j} f_{ikj}$ and $f_{.jk} = \sum_{i \in I} f_{ikj}$ to denote the row and column margins of the frequency table j as a subtable of the global table;
- $f_{i..} = \sum_{j \in J_{\mathrm{f}}} \sum_{k \in K_j} f_{ikj}$ to denote the row margin of the table gathering all the J_{f} frequency tables.

3. MFA: a geometric approach to balancing the influence of the different sets

Mixed data require that the importance of the different variables in the global distance be balanced. In Gower's coefficient (1971), this balancing is achieved by standardizing the distances induced by every variable in such a way that they range between 0 and 1.

Here, the problem is different because we want to balance sets of variables. We therefore adopt the MFA approach. We first recall MFA principles in the case of quantitative and categorical variables and then in the case of frequency data.

Escofier and Pagès (1994, 1988–1998, pp. 149–169) proposed a principal axes method called MFA for dealing with a multiple table in which different sets of quantitative variables are juxtaposed. In order to balance the influence of the different sets, MFA adopts a geometric approach by considering the unit cloud associated with each set of variables and standardizes the inertia of every cloud on the first principal axis to 1. Technically, this property is obtained by dividing the weight of the columns belonging to set j by λ_1^j, the first eigenvalue of the separate analysis of set j. Thus, the contribution of any set to the global distance depends on the actual dimension of the unit cloud: a cloud with several important orthogonal inertia directions will have a greater influence than a one-dimensional cloud. The basis of MFA can be viewed as a weighted PCA applied to the multiple table. The separate tables are considered as in PCA (standardized or non-standardized). We note p_i the weight—usually uniform—assigned to unit i by the user.

To integrate categorical sets into MFA (Escofier and Pagès, 1988–1998, pp. 173–177; Pagès, 2002), the starting point is to make use of the equivalence between MCA and a non-standardized weighted PCA. The results of MCA can be obtained by performing PCA:

- applied to the table with the general term $(z_{ikj} - w_{kj})/w_{kj}$, where $z_{ikj} = 1$ if i belongs to the category k and 0 if it does not, and $w_{kj} = \sum_{i \in I} p_i \cdot z_{ikj}$ (*note*: $\sum_{k \in K_j} w_{kj} = Q_j$);

- giving the weight w_{kj}/Q_j to column k of set j;
- giving the weight p_i to row i.

Both the distances between units and those between columns induced by this PCA are equal to the distances usually considered in MCA. In particular, the square distance between units i and l is:

$$d^2(i,l) = \sum_{k\in K_j} \frac{Q_j}{w_{kj}} \left[\frac{p_i z_{ikj}}{p_i Q_j} - \frac{p_l z_{lkj}}{p_l Q_j} \right]^2 = \sum_{k\in K_j} \frac{1}{Q_j w_{kj}} [z_{ikj} - z_{lkj}]^2. \tag{1}$$

The introduction of frequency tables as sets of variables (1 table = 1 set) induces a specific problem. The reference principal axes method for such a table is CA, which imposes as row weights the coefficients of the row margin. If the row margins of the frequency tables are equal (or proportional) to one another, MFA can be directly extended to this type of data (Abdessemed and Escofier, 1996). When the row margins are different, the extension of MFA is complicated by the fact that the row weights vary according to the set of variables. A solution to this problem is found under the name of MFACT (Bécue-Bertaut and Pagès, 1999, 2001, 2004). We recall here the four steps of the reasoning leading to MFACT.

- *Equivalence between CA and a particular PCA*. The results of classical CA of the J_f frequency tables (or contingency tables) juxtaposed row-wise can be obtained by performing a non-standardized PCA on the table that has the general term (Escofier and Pagès, 1988–1998, p. 96):

$$\frac{f_{ikj} - f_{i..} \cdot f_{.kj}}{f_{i..} \cdot f_{.kj}}, \tag{2}$$

  using $\{f_{i..};\, i=1,\ldots,I\}$ as row weights (and as metric in the column space) and $\{f_{.kj};\, k=1,\ldots,K_j;\, j=1,\ldots,J_\mathrm{f}\}$ as column weights (and as metric in the row space). This analysis decomposes the discrepancy between the data (f_{ikj}) and the independence model corresponding to the overall table $(f_{i..} \cdot f_{.kj})$.
- *Generalization of CA*. CA can be generalized (Escofier, 1984) to any model $\{m_{ikj};\, i=1,\ldots,I;\, k=1,\ldots,K_j;\, j=1,\ldots,J_\mathrm{f}\}$ with the same margins as the data table (i.e. $m_{i..} = f_{i..};\, m_{.kj} = f_{.kj}$). The outputs of this particular CA can be obtained by performing a non-standardized PCA on the table with the general term:

$$\frac{f_{ikj} - m_{ikj}}{f_{i..} f_{.kj}}, \tag{3}$$

  using $\{f_{i..};\, i=1,\ldots,I\}$ as row weights (and as metric in the column space) and $\{f_{.kj};\, k=1,\ldots,K_j;\, j=1,\ldots,J_\mathrm{f}\}$ as column weights (and as metric in the row space).
- *Internal correspondence analysis* (*ICA*). ICA, or within-tables CA, (Benzécri, 1983; Escofier and Drouet, 1983; Cazes and Moreau, 1991, 2000) can be seen as a CA that refers to a particular model—the intra-tables independence model—whose general term is

$$m_{ikj} = \left(\frac{f_{i.j}}{f_{..j}} \right) \cdot f_{.kj}. \tag{4}$$

  The results can be obtained by performing a non-standardized PCA on the table with the following general term:

$$\frac{f_{ikj} - m_{ikj}}{f_{i..} f_{.kj}} = \frac{f_{ikj} - \left(\frac{f_{i.j}}{f_{..j}} \right) \cdot f_{.kj}}{f_{i..} f_{.kj}} = \frac{1}{f_{i..}} \left[\frac{f_{ikj}}{f_{.kj}} - \frac{f_{i.j}}{f_{..j}} \right], \tag{5}$$

  using $\{f_{i..};\, i=1,\ldots,I\}$ as row weights (and as metric in the column space) and $\{f_{.kj};\, k=1,\ldots,K_j;\, j=1,\ldots,J_\mathrm{f}\}$ as column weights (and as metric in the row space). As seen in Section 2, $f_{i..}$ is the sum of row i as computed across all the frequency sets.

Combining ICA and MFA. MFACT combines ICA (which solves the problem of the different row margins) and MFA (which balances the influence of the sets). Thus, MFACT performs a non-standardized PCA on the global table whose general term is given in (5) (as ICA) using $\{f_{i..}; i=1,\dots,I\}$ as row weights (as ICA) but $\{f_{.kj}/\lambda_1^j; k=1,\dots,K_j; j=1,\dots,J_{\mathrm{f}}\}$ as column weights (i.e. the weights used in ICA divided by λ_1^j). λ_1^j is the first eigenvalue issued from the separate PCA of table j, with the weights used in ICA ($f_{i..}$ for row i; $f_{.kj}$ for column k of set j). These weights result from compromise that is necessary when different frequency tables with different margins are compared. The separate analyses, using the row weights $f_{i..}$, are termed “pseudo-separate CA” since the actual separate CA uses the weight $f_{i.j}$. The deformation may be disregarded if the row margins differ only slightly from one table to another.

4. Mixture of quantitative, categorical and frequency data

The introduction of a mixture of quantitative, categorical and frequency tables in a single analysis poses the problem of unit weighting. In the case of quantitative and categorical sets, the unit weights (which are usually uniform) are fixed by the user, whereas in the case of frequency data, they are imposed by the table margins. Compromise weights are necessary since the unit weights have to be identical across all the tables.

A first solution consists in adopting the weights issued from the multiple frequency table, that is, $p_i = f_{i..}$. In this case, the extended MFA is based on a non-standardized weighted PCA performed on the multiple table presented in Table 1, using:

- $\{p_i = f_{i..}; i = 1,\dots,I\}$ as row-unit weights (and as metric in the column space);
- the initial weights of the columns (belonging to set j) but divided by λ_1^j *as* column weights (and as metric in the row space), that is, $(1/\lambda_1^j)$ in the case of a quantitative set, $((w_{kj}/Q_j)/\lambda_1^j)$ in the case of a categorical set, $(f_{.kj}/\lambda_1^j)$ in the case of a frequency set. λ_1^j denotes the first eigenvalue issued from the separate PCA of table j.

The choice of these unit weights has an effect on

- the quantitative sets, through the computation of the mean and standard deviation of every variable;
- the categorical sets, through the computation of the coefficients w_{kj};
- the frequency sets (when there are several), through the use of $f_{i..}$ as row weights.

Another solution consists in adopting the weights issued from the categorical or quantitative tables, usually uniform weights. In this case, the generalized MFA is equivalent to a non-standardized weighted PCA performed on the table presented in Table 1, using $\{p_i = 1/I; i = 1,\dots,I\}$ as row-unit weights (and as metric in the column space), and the column weights used in Section 4.2 as column weights. In this case, the contribution of the quantitative and categorical sets to the global distance corresponds exactly—except for the overweighting by $1/\lambda_1^j$—to the distances issued from classical PCA and MCA. However, these unit weights modify the contribution of the frequency tables to the global distance, which no longer corresponds exactly to the distance issued from MFACT.

In practice, the user will choose one or another unit weighting depending on the characteristics of the application. In the data sets that we are used to handling, that is, survey data with frequency tables issued from open-ended questions, we adopt the weights imposed by the frequency sets. This choice favours the respondents giving the longest answers, those who generally use a richer and more varied vocabulary. In the following section, we take up this framework.

Table 1
Multiple table issued from the original table by the appropriate transformations

Units	Variable k in quantitative set j	Indicator variable k (=category) in categorical set j	Variable k in frequency set j	Unit weights
1				
i	$\dfrac{x_{ikj}-\bar{x}_{kj}}{s_{kj}}$	$\dfrac{(z_{ijk}-w_{kj})}{w_{kj}}$	$\dfrac{f_{ikj}-(f_{i.j}/f_{..j})\cdot f_{.kj}}{p_i\cdot f_{.kj}}$	p_i
I				
Column weights	$\dfrac{1}{\lambda_1^j}$	$\dfrac{w_{kj}}{Q_j\lambda_1^j}$	$\dfrac{f_{.kj}}{\lambda_1^j}$	

5. Main properties of the extended MFA

This extension of MFA remains within the framework of the principal axes methods and classical results are obtained: coordinates, contributions and squared cosines of rows and columns. An important output consists in characterizing every unit by principal coordinate vectors, thus converting the original variables into quantitative variables. Considering the units in this new reference space (using the classical Euclidean distance) is the equivalent of adopting the global distance induced by MFA.

We describe the main properties of this extended MFA below and focus on the particularities derived from the heterogeneity of the data.

5.1. Unit and column clouds

The unit clouds, as described separately by each set of columns and globally by the combination of all the sets, are centred.

MFA induces a distance between units corresponding to a weighted sum of the separate distances induced by every set of variables. The square distance between units i and l, which is computed from the coordinates given in Table 1 the column weights (and metric in the unit space) indicated in Section 4.2, is

$$d^2(i,l) = \sum_{j\in J_\mathrm{q}} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \left[\frac{x_{ikj}-x_{lkj}}{s_{kj}}\right]^2 + \sum_{j\in J_\mathrm{c}} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \frac{1}{Q_j w_{kj}} [z_{ikj}-z_{lkj}]^2$$

$$+ \sum_{j\in J_\mathrm{f}} \frac{1}{\lambda_1^j} \sum_{k\in K_j} \frac{1}{f_{.kj}} \left[\left(\frac{f_{ikj}}{f_{i..}} - \frac{f_{lkj}}{f_{l..}}\right) - \frac{f_{.kj}}{f_{..j}}\left(\frac{f_{i.j}}{f_{i..}} - \frac{f_{l.j}}{f_{l..}}\right)\right]^2. \tag{6}$$

Formula (6) highlights the contribution of each set of variables to the global distance. Disregarding the overweighting by the reverse of the first eigenvalue:

- the quantitative set j ($j \in J_\mathrm{q}$) contributes to the distance (between units i and l) as computed in PCA performed on table j;
- the categorical set j ($j \in J_\mathrm{c}$) contributes to the distance (between units i and l) as computed in MCA performed on table j;
- the frequency table j ($j \in J_\mathrm{f}$) contributes to the distance (between units i and l) as computed in MFACT performed on the table that juxtaposes all the frequency tables.

Moreover, the overweighting by $1/\lambda_1^j$ balances the influence of the different tables.

As in MCA and CA, the clouds of columns corresponding to each set of categorical and frequency variables are centred for the corresponding weights. The proximities between columns of the same type are computed as in the separate analyses. In practice, quantitative variables are represented on a separate graphic as in PCA; however, it is useful to represent categories and frequencies on a single graphic. The relative positions of categories and frequencies are interpreted using transition formulae (see below).

5.2. Transition formulae

The relation that gives the coordinate $F_s(i)$ of unit i (along axis s) from the coordinates of the columns $\{G_s(kj);\ k=1,\ldots,K_j;\ j=1,\ldots,J\}$ is obtained by applying the general transition formula (Pagès, 2002)

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j\in J_\mathrm{q}} \frac{1}{\lambda_1^j} \left[\sum_{k\in K_j} x_{ikj} G_s(kj)\right] + \frac{1}{\sqrt{\lambda_s}} \sum_{j\in J_\mathrm{c}} \frac{1}{\lambda_1^j Q_j} \left[\sum_{k\in K_j} z_{ikj} G_s(kj)\right]$$

$$+ \frac{1}{\sqrt{\lambda_s}} \sum_{j\in J_\mathrm{f}} \frac{1}{\lambda_1^j} \frac{f_{i.j}}{f_{i..}} \left[\sum_{k\in K_j} \frac{f_{ikj}}{f_{i.j}} G_s(kj)\right]. \tag{7}$$

This transition formula is similar:

- for the quantitative sets, to the one used in PCA;
- for the categorical sets, to the one used in MCA;
- for the frequency sets, to the one used in MFACT.

Similarly, by using the general transition formulae, we can express the coordinates of a column depending on the coordinates of the units. These formulae are similar to the formulae obtained:

- in PCA, in the case of quantitative sets: the coordinate of a quantitative variable on axis s is its correlation coefficient with the principal coordinate vector corresponding to this axis;
- in MCA, in the case of categorical sets: a category lies—except for a coefficient—at the centroid of the units presenting it;
- in MFACT, in the case of frequency sets: a frequency column is attracted (or repelled) by the rows that are more (or less) associated with it than if there were independence between rows and columns in table j.

These transition formulae allow supplementary units and/or variables to be considered.

5.3. *Superimposed representation of the partial clouds*

As in MFA or MFACT, it is possible to obtain a superimposed representation of the J clouds of units, named partial clouds, corresponding to the J separate analyses (Bécue-Bertaut and Pagès, 2001). The representation of the jth partial cloud is issued from formula (7) but restricted to the columns of set j only. This superimposed representation can also be extended to the categories which are located in the centroid of the units that present them. We use this possibility to represent the centroids of the clusters.

6. Clustering step

6.1. *Clustering method*

Clustering can be performed from the principal coordinates without any loss of information: by computing the classical Euclidean distance from the principal coordinates, we obtain the same results as we do when we compute the distance given in formula (6) from the initial variables. Therefore, we have to deal with the classical problem of clustering from quantitative variables, for which different methods can be used: either partitioning clustering—classical k-means or recent methods such as CLUES (Wang et al., 2007)—or hierarchical clustering. In this case, we use the latter, in conjunction with the generalized Ward's criterion, a method that is appropriate for operating from coordinates issued from a principal axis method (Lebart, 1994; Lebart et al., 2000, pp. 167–168; Nakache and Confais, 2005, pp. 130–131). In the computation of Ward's criterion, the units are weighted as in the principal axes method.

It can also be useful to compute the distances between units just from their coordinates on the first principal axes, those considered to be the most significant in a statistical sense—which involves filtering the "noise" conveyed by the last axes (Lebart et al., 2000, pp. 187–188).

6.2. *Association between partition and variables*

In order to interpret the partition, we measure the association between the partition (considered as a categorical variable) and

- each quantitative variable, using the usual square correlation ratio η^2 (between-cluster variance/total variance);
- each categorical variable, using Cramer's coefficient V (Saporta, 2006, p. 150);
- each frequency set, using Cramer's coefficient. In this case, we consider the set as a whole; χ^2 statistic and Cramer's coefficient V are computed on the table crossing the frequency set and the partition.

Cramer's coefficient standardizes the χ^2 statistic by dividing it by its maximum: it does not depend on the grand total of the table and varies from 0 (independence) to 1 (functional relationship); it can be compared between tables having different grand total and different dimensions.

$$V = \sqrt{\frac{\chi^2}{n \cdot min(r-1, c-1)}}, \tag{8}$$

where χ^2 is the chi-square statistic, n the grand total of the table (i.e. the number of units, in the case of categorical variables, or the grand total of the corresponding crossed table in the case of a frequency set) and r and c are the number of rows and columns of the table, respectively.

6.3. Description of each cluster

For each quantitative variable, the global and cluster means are compared using a classical statistic (Lebart et al., 2000, pp. 181–182).

For each categorical variable, we compare the proportions of units belonging to every category within the cluster, on the one hand, and in the overall sample, on the other; this comparison is achieved by a permutation test using a hypergeometric model (Lebart et al., 2000, pp. 182–184). Thus, for every cluster, we associate each category with a significance level (or p-value) that allows the categories to be ranked from the most over-represented to the most under-represented.

A very similar process allows the characterization of a cluster by the frequency variables that are over- or under-represented in all the answers belonging to that cluster (Lebart et al., 1998, Chapter 6). Here again, for every cluster, we can rank the different events associated with the frequency variables (which are the different words in our example) from the most over-represented to the most under-represented.

7. Example: children's reading habits

7.1. Data

The application is extracted from a large study carried out in a town near Barcelona. A closed questionnaire concerning reading habits was answered by 895 fifth-grade children (10- and 11-years-olds). They had to also complete the two following statements:

1. *Para mí leer es*…(For me, to read means…).
2. *Creo que leer es importante porque*…(I believe that reading is important because…).

We keep only the answers of the 817 children who completed almost all the questionnaire. The eight closed questions concerning reading habits (Table 2) correspond to the first set (categorical variables) and the two open-ended questions make up sets 2 and 3. The columns of the first set (indicator variables) are the categories of the closed questions; the columns of the second and third sets contain the frequency of the words used in the answers to the corresponding open-ended question, counted for every child. We keep only the words used at least 10 times by the children.

Table 2
The eight closed questions

1. *At school*, *we read* (*amount of school reading*)	(very little, enough, a lot)
2. *At home*, *we have* (*amount of books*)	(few, enough, a lot of books)
3. *I read* (*amount of reading*)	(very little, enough, a lot)
4. *I read* (*ease of reading*)	(very easily, easily, with difficulty)
5. *The books given by the teacher*	(I like them, I do not like them)
6. *I read when* (*context of reading*)	(I feel like reading, I do school work, both)
7. *I prefer reading* (*way of reading*)	(silently, aloud, both)
8. *Reading school books*	(I enjoy it, I do not enjoy it, it depends)

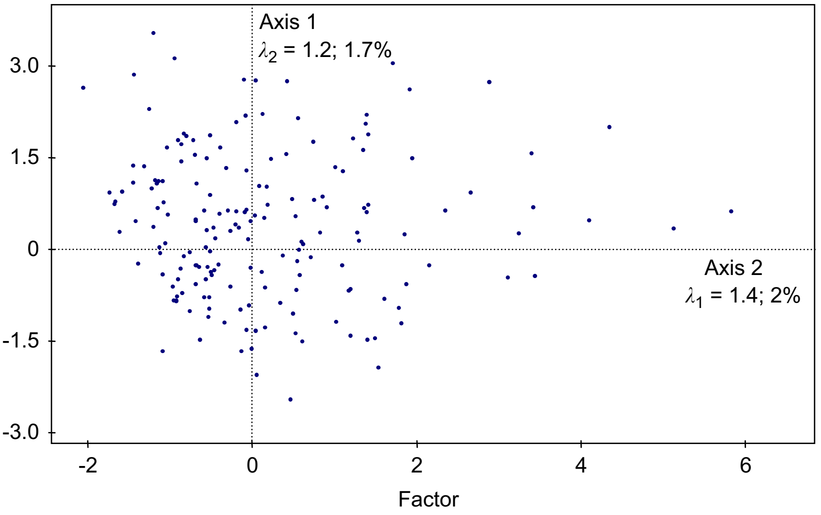

Fig. 2. First principal plane: representation of children.

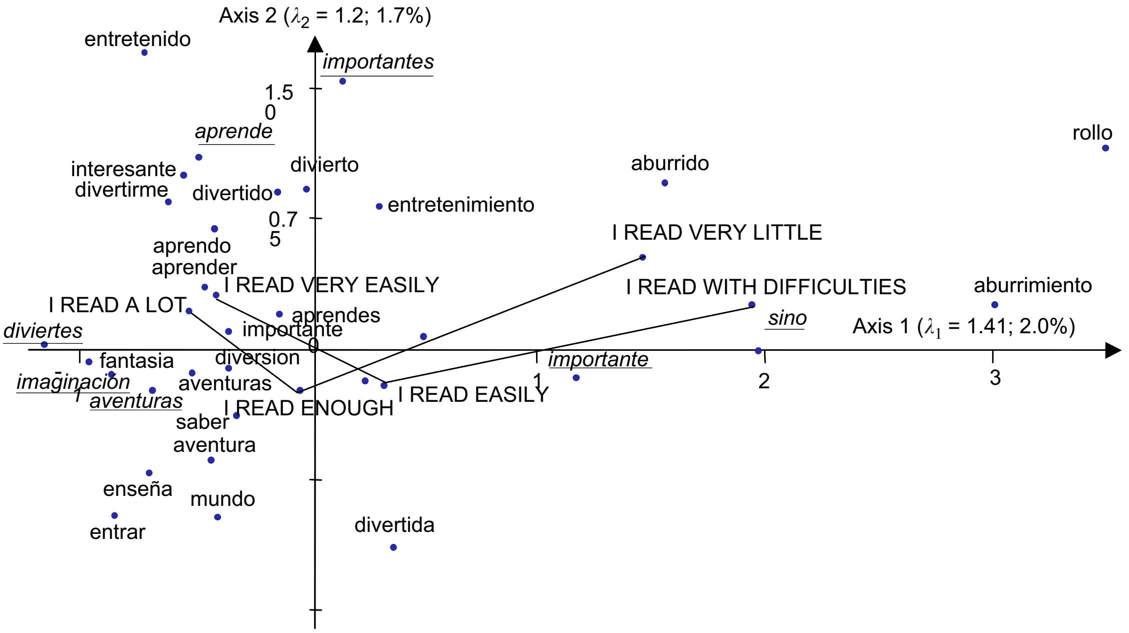

Fig. 3. First principal plane: extract of the column representation. The underlined words correspond to the second open-ended question.

7.2. MFA: main features of the first two axes

In the separate analyses, the first eigenvalues are, respectively 0.22 (closed questions), 0.51 (first open-ended question) and 0.49 (second open-ended question); these different values illustrate the need to balance the influence of the sets using MFA column overweighting.

Figs. 2 and 3 show the representation of children and columns (categories and words) on the first principal plane. The interpretation rules are those used in MCA (categories) or CA (frequencies).

The first axis, to which the three sets of columns contribute 47%, 31% and 22% of the inertia, respectively, can be interpreted as a general reading "level" axis. This dispersion direction is common to the three sets but differs from

Table 3
Measure of the association between partition and questions (partition built up from closed and open-ended questions)

Question	Cramer's coefficient
At home, we have (amount of books)	0.52
Reading school books	0.44
I read (amount of reading)	0.41
For me, to read means (open-ended)	0.38
I read (ease of reading)	0.35
Reading is important because (open-ended)	0.27
The books given by the teacher	0.26
I prefer reading (way of reading)	0.20
I read when (context of reading)	0.20
At school, we read (amount of school reading)	0.14

the first principal axes obtained in the separate analyses: the first eigenvalue of the global analysis (1.4) is far from its maximum (here, 3), which is reached when the axes of the separate analyses are identical (Escofier and Pagès, 1988–1998, p. 161).

This axis contrasts categories and words indicating an enjoyment of reading with those that betray a dislike. Thus, on the left of the axis we find *I read a lot*, *I read very easily*, *diviertes* (*you enjoy yourself*) and *imaginación* (*imagination*), whilst on the right we find *I read very little*, *I read with difficulty*, *rollo* (*drag*, used in the expression *reading is a drag*) and *aburrimiento* (*boredom*). As regards the open-ended questions, the children who "*read a lot*" or only "*read enough*" do not differ greatly and are close to one another.

The second axis, mainly due to the second open-ended question, contrasts reading as a school activity (*importantes* = *important*, *aprende* = *you learn*) with reading for pleasure (*divertida* = *fun*).

7.3. Hierarchical clustering

For the clustering step, we wish to take into account most of the information considered as relevant in the separate analyses. We therefore select a subspace that ensures a good quality of representation for the axes interpreted in the separate analyses, projected as illustrative on the axes issued from the global analysis. This approach leads to the selection of the first seven global axes. Furthermore, this selection is confirmed by applying the empirical criterion called Catell's scree-test (Catell, 1966Lebart et al., 2000, p. 374; Saporta, 2006, p. 173).

The first seven axes explain only 11% of the total inertia but, as Lebart et al. (2000, pp. 368–369) point out, in MCA as well as in CA applied to units × words tables, and also in MFACT applied to juxtaposed tables, the inertia rates corresponding to the first axes are necessarily weak due to the coding (in the case of categorical variables) and to the specificity of textual data (in the case of frequency variables). Thus, these rates can only give a pessimistic idea of the information conveyed by the axes.

The hierarchical tree (not reproduced here) suggests keeping a partition in seven clusters. The ratio of the between-class inertia to the total inertia is 43%, which shows that the clusters do not have clear edges, which might be expected given the great homogeneity in the socio-economic conditions and age of the children.

7.4. Characterization of the partition

Table 3 shows the measure of the association between the partition and the questions (closed and open-ended) through Cramer's coefficient: we can see, for example, how the first open-ended question (*For me*, *to read means*…) and the amount of reading present very similar levels of association with the partition.

7.5. Description of the clusters

For example, clusters 2 and 5, both of which contain good or very good readers according to the closed questions (partial points corresponding to set 1, Fig. 4), give very different answers to the open-ended questions (as indicated by the position of the partial points corresponding to sets 2 and 3 in Fig. 4). Children in cluster 2 definitely read for educational purposes whereas those in cluster 5 read for fun.

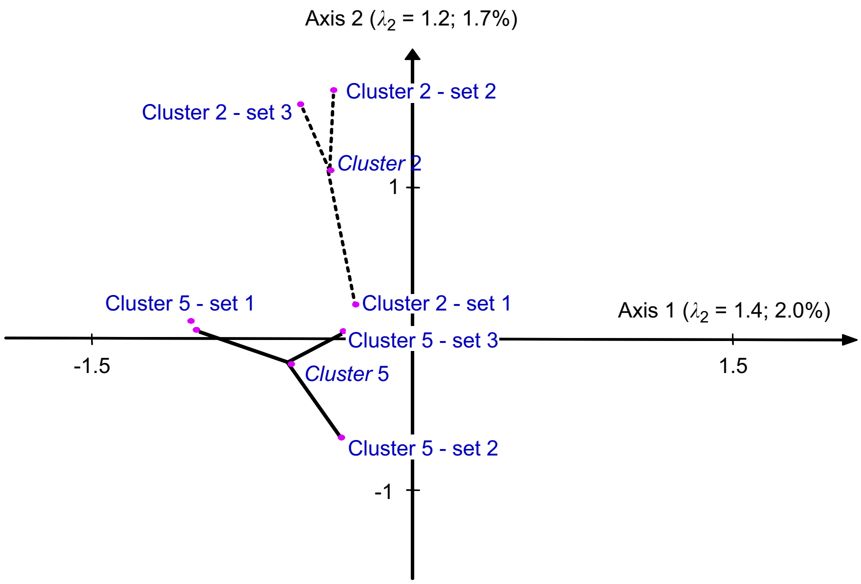

Fig. 4. Superimposed representation of partial and global centroids of clusters 2 and 5. For each cluster, there are one global point and three partial points corresponding to the three sets.

The interpretation of the clustering results can be illustrated by directly considering the answers of the children belonging to the clusters. Table 4 presents a detailed description of the "fond of reading" cluster (cluster 5) according to the over-represented categories and words. In particular, the responses "*read a lot*" and "*very easily*" are given more often by the children in this cluster than in the sample as a whole.

However, we can see that in this cluster only 50% of the children "*read a lot*" (versus 28% in the whole sample) while 45% "*read enough*" (a category not over-represented in the cluster since it is shared by 58% of the children in the whole sample). In fact, the children belong to this cluster not only because of their declared reading habits but also as a result of their free answers. These answers show an enthusiastic attitude to reading, which they consider to be a hobby and not a school requirement. They have a wide vocabulary, considering their young age (10–11 years). These features are clearly highlighted by the characteristic words and answers (Table 4); the latter, also called modal answers, are given by the children closest to the cluster centroid (Lebart et al., 1998, pp. 136–145). We also note that the mean length of the answers is the highest in comparison with the other clusters, which shows that the children are interested in the topic and like to speak about it.

7.6. *Comparison with the results obtained when the clustering step is performed from closed questions only*

To underline the contribution of the open-ended questions, we compare the former results with those obtained when clustering is performed from the answers to the closed questions only. An equivalent strategy is used: first, MCA is performed on the set of closed questions; then, a partition into five clusters is obtained through hierarchical clustering starting from the first three axes (those that are actually interpretable).

As Table 5 shows, in this second analysis, the declared amount of reading and lack of difficulty play a basic role. With only two exceptions, the closed questions have a higher level of association with the partition than was found in the former case. Specifically, these children read a lot and do so easily. The open-ended questions, in this case treated as illustrative elements, present a low level of association with the partition: in a single cluster, the children present a large variability in their free answers (which definitely shows that the differences revealed by open-ended questions are absent in closed questions).

The description of the most "fond of reading" cluster (Table 6) is very instructive. The children belonging to this cluster have chosen, in almost every case, the "convenient" closed categories: 82% "*read a lot*" (and only 18% "*read enough*"), 93% "*read very easily*" and 63% "*read a lot at school*", when in the former "fond of reading class" reading as a school activity was not over-represented.

Table 4
Description of the "fond of reading" cluster: over-represented categories and words (p-value < 0.01) and modal answers (partition built up from open-ended and closed questions)

Set of questions	"Fond of reading" cluster 220 children
Reading habits (closed questions; active)	Over-represented categories: I read a lot (50%; 28%) I read very easily (81%; 58%) I read silently (85%; 73%) I read when I feel like reading (64%; 51%)
For me, to read means… (first open-ended question; active)	Over-represented words: *pasar*, as in *pasar un buen rato* (to have a good time; 35 of 41), *diversión* (fun; 30 of 42), *aventura* (adventure; 44 of 58), *rato* (time; 21 of 28), *tiempo* (time; 14 of 16), *divertirme* (to have fun; 32 of 45), *mundo* (world; 13 of 16), *libro* (book; 24 of 38), *entrar* (to get into; 9 of 10), *fantasia* (fantasy; 9 of 11), *forma* (way; 8 of 12)
Mean length of the answers: in the whole sample, 6.8 words	Mean length of the answers: in the cluster, 8.8 words Modal answers: • *Entrar en el libro que estoy leyendo y pasar las aventuras que hay en el libro* (to get into the plot of the book that I am reading and to live out the adventures that it contains) • *Entrar en el libro, ser el protagonista y pasar aventuras leyendo* (to get into the book, to be the main character and to live out adventures whilst reading)
I believe that reading is important because (second open-ended question; active)	Over-represented words: *imaginación* (imagination; 18 of 19), *hace* (to do; 8 of 11), *aprende* (to learn; 25 of 53), *vocabulario* (vocabulary; 10 of 16), *divertido* (fun; 9 of 15), *ayuda* (help; 15 of 30), *aventura* (adventure; 8 of 13)
Mean length of the answers: in the whole sample, 7.4 words	Mean length of the answers: in the cluster, 8.7 words Modal answers: • *Te enseña palabras nuevas. Viajas a paises con la imaginación* (it teaches you new words. You travel to other countries with your imagination.) • *Aprendo ortografía. Se me abre la imaginación* (I learn spelling. It stimulates my imagination.)

For each category: percentages of children presenting this category within the cluster and across the whole sample.
For each word: frequency within the cluster and across the whole sample.

Table 5
Measure of the association between partition and questions (partition built up from closed questions only)

Question	Cramer's coefficient
I read (*amount of reading*)	0.62
I read with (*ease of reading*)	0.50
I read when (*context of reading*)	0.45
The books given by the teacher	0.43
Reading school books	0.39
At school, *we read* (*amount of school reading*)	0.39
I prefer reading (*way of reading*)	0.32
At home, *we have* (*amount of books*)	0.30
For me, *to read means* (*open-ended*)	0.17
Reading is important because (*open-ended*)	0.15

In fact, these children present a large variability in their free answers, from enthusiastic comments to short, laconic answers that mention reading as a way of learning. Thus, for the first open-ended question, no word is over-represented; for the second open-ended question, only two words are over-represented.

Table 6
Description of the "fond of reading" cluster: over-represented categories and words (p-value < 0.01) and modal answers (partition built up from closed questions only)

Set of questions	"Fond of reading" cluster 168 children
Attitude about reading (closed questions; active)	Over-represented categories: I read a lot (82%; 28%) I read very easily (93%; 58%) At school, we read a lot (63%; 32%) At home, we have a lot of books (93%; 67%) I enjoy reading school books (96%; 78%) I enjoy reading books given by the teacher (98%; 85%) I read silently (85%; 73%) I read when I feel like reading (62.5%; 51%)
For me, to read means… (first open-ended question; illustrative) Mean length of the answers: in the whole sample, 6.8 words	Over-represented words: no word is over-represented in this cluster Mean length of the answers: in the cluster, 7.6 words
I believe that reading is important because (second open-ended question; illustrative) Mean length of the answers: in the whole sample, 7.4 words	Over-represented words: aprende (to learn; 19 of 53), cosas (things; 79 of 295) Mean length of the answers: in the cluster, 7.8 words Modal answers: • *Se aprende* (you learn) • *Se aprende* (you learn)

For each category: percentages of children presenting this category within the cluster and across the whole sample.
For each word: frequency within the cluster and across the whole sample.

8. Conclusion

The need to simultaneously consider mixed data, composed of sets of quantitative, categorical and frequency variables, arises in a wide range of applications. This can be done by combining MFACT, an extension of MFA for dealing with multiple frequency tables, and classical MFA so that quantitative and/or categorical variables may be taken into account. This extended MFA induces a global distance between units that balances the influence of the sets of variables, allowing the units to be clustered based on all the variables. Two solutions are proposed to the unit weighting problem that arises when frequency sets are merged with categorical or quantitative sets.

This technique can be very useful in survey data analysis: it is a natural requirement for a researcher to be able to consider the answers to open-ended and closed questions for clustering the respondents to a questionnaire. In a clustering of this type,

- the closed questions provide a solid framework, but they tend to reproduce the a priori assumptions that underlie the construction of the questionnaire when they are analysed separately;
- the open-ended questions bring variety, but may be puzzling the researcher when they are examined separately.

Taking into account the two types of questions simultaneously seems to provide a means of combining their respective advantages. Up to now, this approach has not been used much, most likely due to the inherent technical problems.

Acknowledgements

This work has been partially supported by Spanish Ministry of Education and Science and FEDER (Grant SEJ2005-00741/ECON). We thank Claude Langrand for his advices.

References

Abdessemed, L., Escofier, B., 1996. Analyse factorielle multiple de tableaux de fréquences; comparaison avec l'analyse canonique des correspondances. Journal de la Société de Statistique de Paris 137 (2), 3–18.

Bécue-Bertaut, M., Pagès, J., 1999. Intra-sets multiple factor analysis. Application to textual data. In: Bacelar-Nicolau, H., Costa Nicolau, F., Janssen, J. (Eds.), Applied Stochastic Models and Data Analysis. INE, Lisbon, pp. 72–79.

Bécue-Bertaut, M., Pagès, J., 2001. Analyse simultanée de questions ouvertes et de questions fermées. Méthodologie, exemple. Journal de la Société Française de Statistique 42 (4), 91–104.

Bécue-Bertaut, M., Pagès, J., 2004. A principal axes method for comparing contingency tables: MFACT. Comput. Statist. Data Anal. 45 (3), 481–503.

Benzécri, J.P., 1983. Analyse de l'inertie intraclasse par l'analyse d'un tableau de contingence. Les Cahiers de l'Analyse des Données 8 (3), 351–358.

Catell, R.B., 1966. The scree test for the number of factors. Mult. Behav. Res. 1, 245–276.

Cazes, P., Moreau, J., 1991. Analysis of a contingency table in which the rows and the columns have a graph structure. In: Diday, E., Lechevallier, Y. (Eds.), Symbolic-Numeric Data Analysis and Learning. Nova Science Publishers, New York, pp. 271–280.

Cazes, P., Moreau, J., 2000. Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphe bistochastique. In: Moreau, J., Doudin, P.A., Cazes, P. (Eds.), L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données. Springer, Berlin-Heidelberg, pp. 87–103.

Chae, S.S., Warde, W.D., 2006. Effect of using principal coordinates and principal components on retrieval of clusters. Comput. Statist. Data Anal. 50 (6), 1407–1417.

Escofier, B., 1984. Analyse factorielle en référence à un modèle: application à l'analyse d'un tableau d'échanges. Rev. Statist. Appl. 32, 25–36.

Escofier, B., Drouet, D., 1983. Analyse des différences entre plusieurs tableaux de fréquence. Les Cahiers de l'Analyse des Données 8 (4), 491–499.

Escofier, B., Pagès, J., 1988–1998. Analyses factorielles simples et multiples; objectifs, méthodes et interprétation. Dunod, Paris.

Escofier, B., Pagès, J., 1994. Multiple factor analysis: AFMULT package. Comput. Statist. Data Anal. 18, 121–140.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27 (4), 857–871.

Grabmeier, J., Rudolph, A., 2002. Techniques of cluster algorithms in data mining. Data Mining and Knowledge Discovery 6, 303–360.

Lebart, L., 1994. Complementary use of correspondence analysis and cluster analysis. In: Greenacre, M., Blasius, J. (Eds.), Correspondence Analysis in the Social Sciences. Academic Press, San Diego, pp. 162–178.

Lebart, L., Salem, A., Berry, L., 1998. Exploring Textual Data. Kluwer Academic Publishers, Dordrecht.

Lebart, L., Morineau, A., Piron, M., 2000. Statistique exploratoire multidimensionnelle. Dunod, París.

Nakache, J.P., Confais, J., 2005. Approche pragmatique de la classification. Technip, Paris.

Pagès, J., 2002. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. Rev. de Statist. Appl. 50 (4), 5–37.

Podani, J., 1999. Extending Gower's general coefficient of similarity to ordinal characters. Taxon 48 (2), 331–340.

Saporta, G., 2006. Probabilités, analyse des données et statistique. Technip, Paris.

Wang, X., Qiu, W., Zamar, R.H., 2007. CLUES: a non-parametric clustering method based on local shrinking. Comput. Statist. Data Anal. 52 (1), 286–298.