A Partitioning Method for the Clustering of Categorical Variables

Marie Chavent[1,2], Vanessa Kuentz[1,2], and Jérôme Saracco[1,2,3]

[1] Université de Bordeaux, IMB, CNRS, UMR 5251, France
chavent@math.u-bordeaux1.fr;kuentz@math.u-bordeaux1.fr
[2] INRIA Bordeaux Sud-Ouest, CQFD team, France
[3] Université Montesquieu - Bordeaux IV, GREThA, CNRS, UMR 5113, France
jerome.saracco@u-bordeaux4.fr

Summary. In the framework of clustering, the usual aim is to cluster observations and not variables. However the issue of clustering variables clearly appears for dimension reduction, selection of variables or in some case studies. A simple approach for the clustering of variables could be to construct a dissimilarity matrix between the variables and to apply classical clustering methods. But specific methods have been developed for the clustering of variables. In this context center-based clustering algorithms have been proposed for the clustering of quantitative variables. In this article we extend this approach to categorical variables. The homogeneity criterion of a cluster of categorical variables is based on correlation ratios and Multiple Correspondence Analysis is used to determine the latent variable of each cluster. A simulation study shows that the method recovers well the underlying simulated clusters of variables. Finally an application on a real data set also highlights the practical benefits of the proposed approach.

Key words: clustering of variables, center-based clustering algorithm, latent variable, Multiple Correspondence Analysis.

1 Introduction

From a general point of view, variable clustering lumps together variables which are strongly related to each other and thus bring the same information. It is a possible solution for selection of variables or dimension reduction which are current problems with the emergency of larger and larger data bases. In some case studies, the main objective is to cluster variables and not units, such as sensory analysis (identification of groups of descriptors), biochemistry (gene clustering), etc. Techniques of variable clustering can also be useful for association rules mining (see for instance Plasse and al. [4]).

A simple approach for the clustering of variables could be to calculate first the matrix of the dissimilarities between the variables and then to apply

classical clustering methods which are able to deal with dissimilarity matrices (complete or average linkage hierarchical clustering among others). Other methods like Ward or k-means (dealing only with quantitative data) could also be applied on the numerical coordinates obtained from Multidimensional Scaling of this dissimilarity matrix. But specific methods have also been developed for the clustering of variables. In this context Cluster Analysis of Variables Around Latent Components (Vigneau and Qannari [5]) and Diametrical clustering (Dhillon and al. [3]) are two independently proposed center-based clustering methods for the clustering of quantitative variables. These methods are iterative two steps relocation algorithms involving at each iteration the identification of a cluster centroid by optimization of an homogeneity criterion and the allocation of each variable to the "nearest" cluster. The cluster centroid is a synthetic component, called latent variable, which summarizes the variables belonging to the cluster. When high absolute correlations imply agreement, both methods aim at maximizing the same homogeneity criterion (based on squared correlations). In this case, the latent variable of a cluster is the first principal component issued from Principal Component Analysis (PCA) of the matrix containing the variables of the cluster.

In this paper we extend this relocation partitioning method to the case of categorical variables. The homogeneity criterion is now based on correlation ratios between the categorical variables and the cluster centroids which are numerical variables, defined by optimization of this homogeneity criterion.

Sect. 2 presents the center-based clustering algorithm for the clustering of categorical variables. A simulation study is carried out in Sect. 3 to show the numerical performance of the approach and a real data application illustrates its practical benefits. Finally some concluding remarks are given in Sect. 4.

2 A Center-based Partitioning Method for the Clustering of Categorical Variables

Let $\mathbf{X} = (x_{ij})$ be a data matrix of dimension (n, p) where a set of n objects are described on a set of p categorical variables, that is, $x_{ij} \in \mathcal{M}_j$ where $\mathcal{M}_j$ is the set of categories of the jth variable. Let $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p\}$ be the set of the p columns of $\mathbf{X}$, called for sake of simplicity categorical variables. We denote by $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots, \mathcal{C}_K\}$ a partition of $\mathcal{V}$ into K clusters and by $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_k, \ldots, \mathbf{y}_K\}$ a set of K vectors of $\mathbb{R}^n$ called latent variables.

The aim is to find a couple $(\mathcal{P}, \mathcal{Y})$, optimum with respect to the following homogeneity criterion:

$$H(\mathcal{P}, \mathcal{Y}) = \sum_{k=1}^{K} S(\mathcal{C}_k, \mathbf{y}_k), \tag{1}$$

where S measures the adequacy between $\mathcal{C}_k$ and the latent variable $\mathbf{y}_k$:

$$S(\mathcal{C}_k, \mathbf{y}_k) = \sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{y}_k), \tag{2}$$

with $\eta^2(\mathbf{x}_j, \mathbf{y}_k)$ the correlation ratio measuring the link between $\mathbf{x}_j$ and $\mathbf{y}_k$.

Definition 1 *The correlation ratio* $\eta^2(\mathbf{x}_j, \mathbf{y}_k) \in [0,1]$ *is equal to the between group sum of squares of* $\mathbf{y}_k$ *in the groups defined by the categories of* $\mathbf{x}_j$*, divided by the total sum of squares of* $\mathbf{y}_k$*. We have with* $\mathbf{y}_k = (y_{k,1}, \ldots, y_{k,i}, \ldots, y_{k,n}) \in \mathbb{R}^n$*,* $\eta^2(\mathbf{x}_j, \mathbf{y}_k) = \dfrac{\sum_{s \in \mathcal{M}_j} n_s (\bar{\mathbf{y}}_{ks} - \bar{\mathbf{y}}_k)^2}{\sum_{i=1}^{n} (y_{k,i} - \bar{\mathbf{y}}_k)^2}$*, with* n_s *the frequency of category* s*,* $\mathcal{M}_j$ *the set of categories of* $\mathbf{x}_j$ *and* $\bar{\mathbf{y}}_{ks}$ *the mean value of* $\mathbf{y}_k$ *calculated on the objects belonging to category* s*.*

2.1 Definition of the Latent Variable

The latent variable $\mathbf{y}_k$ of a cluster $\mathcal{C}_k$ is defined by maximization of the adequacy criterion S:

$$\mathbf{y}_k = \arg\max_{\mathbf{u} \in \mathbb{R}^n} \sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{u}). \tag{3}$$

Proposition 1 *The first principal component obtained with Multiple Correspondence Analysis of* $\mathbf{X}_k$*, the matrix containing the variables of* $\mathcal{C}_k$*, is a solution of* (3) *and is then a latent variable* $\mathbf{y}_k$ *of* $\mathcal{C}_k$*.*

PROOF. Let us introduce some notations. Let $\mathbf{G} = (g_{is})_{n \times q_k}$, with $g_{is} = 1$ if i belongs to category s and 0 otherwise, be the indicator matrix of the q_k categories of the p_k variables in $\mathcal{C}_k$. We note $\mathbf{F}_k = (f_{is})_{n \times q_k}$ the frequency matrix built from $\mathbf{G}$. The row and column marginals define respectively the vectors of row and column masses $\mathbf{r}_k$ and $\mathbf{c}_k$. The ith element of $\mathbf{r}_k$ is $f_{i.} = \frac{1}{n}$ and the sth element of $\mathbf{c}_k$ is $f_{.s} = \frac{n_s}{np_k}$. Let us consider the two following diagonal matrices $\mathbf{D}_n = \text{diag}(\mathbf{r}_k)$ and $\mathbf{D}_{q_k} = \text{diag}(\mathbf{c}_k)$. We introduce the matrix $\widetilde{\mathbf{F}}_k = \mathbf{D}_n^{-1/2}(\mathbf{F}_k - \mathbf{r}_k \mathbf{c}_k^t)\mathbf{D}_{q_k}^{-1/2}$ which general term writes:

$$\tilde{f}_{is} = \frac{\sqrt{n_s p_k}}{n_s}\left(\frac{g_{is}}{p_k} - \frac{n_s}{np_k}\right) = \begin{cases} \frac{\sqrt{n_s p_k}}{n_s}\left(\frac{1}{p_k} - \frac{n_s}{np_k}\right) & \text{if } i \text{ belongs to category } s, \\ 0 & \text{otherwise.} \end{cases}$$

First we show that if $\mathbf{u}^t\mathbf{u} = 1$ and $\bar{\mathbf{u}} = 0$, then $\frac{1}{p_k}\sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{u}) = \mathbf{u}^t \widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t \mathbf{u}$. If $\bar{\mathbf{u}} = 0$, $\sum_{i=1}^{n} \tilde{f}_{is}\mathbf{u}_i = \frac{\sqrt{n_s}}{\sqrt{p_k}}\bar{\mathbf{u}}_s$, where $\bar{\mathbf{u}}_s$ is the mean value of $\mathbf{u}$ calculated on the objects belonging to category s. Thus we have:

$$\begin{aligned} \mathbf{u}^t \widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t \mathbf{u} &= \frac{1}{p_k} \sum_{\mathbf{x}_j \in \mathcal{C}_k} \sum_{s \in \mathcal{M}_j} n_s \bar{\mathbf{u}}_s^2 = \frac{\frac{1}{p_k}\sum_{\mathbf{x}_j \in \mathcal{C}_k} \sum_{s \in \mathcal{M}_j} \frac{n_s}{n}(\bar{\mathbf{u}}_s - 0)^2}{\frac{1}{n}} \\ &= \frac{1}{p_k} \sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{u}). \end{aligned}$$

As the first normalized eigenvector $\mathbf{u}_1$ of $\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t$ maximizes $\mathbf{u}^t\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t\mathbf{u}$, it is a solution of (3).

Finally, as $\eta^2(\mathbf{x}_j, \mathbf{u}) = \eta^2(\mathbf{x}_j, \alpha\mathbf{u})$, for any nonnull real α, $\alpha\mathbf{u}_1$ is also a solution of (3). The proof is then completed by showing that $\mathbf{u}_1$ is colinear to the first principal component issued from MCA on the centered row profiles matrix $\mathbf{R}_k$ of $\mathbf{X}_k$. MCA can be viewed as a weighted PCA applied to $\mathbf{R}_k = \mathbf{D}_n^{-1}(\mathbf{F}_k - \mathbf{r}_k\mathbf{c}_k^t)$. The first principal component is then $\boldsymbol{\psi}_1 = \mathbf{R}_k\mathbf{D}_{q_k}^{-1/2}\mathbf{v}_1$, where $\mathbf{v}_1$ is the eigenvector associated with the largest eigenvalue λ_1 of $\widetilde{\mathbf{F}}_k^t\widetilde{\mathbf{F}}_k$. Then we use the SVD of $\widetilde{\mathbf{F}}_k$ to write $\boldsymbol{\psi}_1 = \sqrt{\lambda_1}\sqrt{n}\mathbf{u}_1$, and the proof is complete. □

2.2 The Center-based Clustering Algorithm

The corresponding center-based algorithm is the following:

(a) *Initialization step*: We compute the first K principal components issued from MCA of $\mathbf{X}$. Then we assign each variable to the nearest component, that is to the component with which its correlation ratio is the highest. Thus we get an initial partition $\{C_1, \ldots, C_k, \ldots, C_K\}$ of $\mathcal{V}$.

(b) *Representation step*: $\forall k = 1, ..., K$, compute the latent variable $\mathbf{y}_k$ of $\mathcal{C}_k$ as the first principal component $\boldsymbol{\psi}_1$ of $\mathbf{X}_k$ (or as the first normalized eigenvector $\mathbf{u}_1$ of $\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t$).

(c) *Allocation step*: $\forall j = 1, ..., p$, find ℓ such that $\ell = \arg\max_{k=1,\ldots,K} \eta^2(\mathbf{x}_j, \mathbf{y}_k)$. Let $\mathcal{C}_k$ be the previous cluster of $\mathbf{x}_j$. Then if $\ell \neq k$, $\mathcal{C}_\ell \leftarrow \mathcal{C}_\ell \cup \{\mathbf{x}_j\}$ and $\mathcal{C}_k \leftarrow \mathcal{C}_k \backslash \{\mathbf{x}_j\}$.

(d) If nothing changes in (c) then *stop*, else return to step (b).

Proposition 2 *The center-based algorithm converges to a local optimum of the homogeneity criterion H.*

PROOF. We show that the homogeneity criterion H increases until convergence. For that we have to prove that $H(\mathcal{P}^n, \mathcal{Y}^n) \leq H(\mathcal{P}^n, \mathcal{Y}^{n+1}) \leq H(\mathcal{P}^{n+1}, \mathcal{Y}^{n+1})$, where the superscript n denotes the nth iteration of the algorithm.

The first inequality is verified since the latent variable of a cluster $\mathcal{C}_k^n$ is defined to maximize S and then $S(\mathcal{C}_k^n, \mathbf{y}_k^n) \leq S(\mathcal{C}_k^n, \mathbf{y}_k^{n+1})$. Then by summing up on k, we get $H(\mathcal{P}^n, \mathcal{Y}^n) \leq H(\mathcal{P}^n, \mathcal{Y}^{n+1})$.

Finally according to the definition of the allocation step, we have $\sum_{k=1}^K \sum_{\mathbf{x}_j \in \mathcal{C}_k^n} \eta^2(\mathbf{x}_j, \mathbf{y}_k^{n+1}) \leq \sum_{k=1}^K \sum_{\mathbf{x}_j \in \mathcal{C}_k^{n+1}} \eta^2(\mathbf{x}_j, \mathbf{y}_k^{n+1})$, which proves the second inequality. □

3 Applications

In this section we present some applications of the center-based clustering algorithm for the clustering of categorical variables. In the first one we con-

sider a simulated example in order to show the numerical performance of the proposed approach. Then we apply it on a real categorical data set to show the potential of the approach.

3.1 Simulation Study

In this simulation study we consider six binary variables $x_1, \ldots, x_6$ and we study four different states of relationship between them. The idea is to simulate at first three groups of variables which are well defined, that is the variables within each cluster are strongly linked to each other and they are weakly related to variables belonging to other clusters. They form the partition $\mathcal{Q} = (\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3)$ with $\mathcal{Q}_1 = \{x_1, x_2\}$, $\mathcal{Q}_2 = \{x_3, x_4\}$ and $\mathcal{Q}_3 = \{x_5, x_6\}$. Then we increasingly disrupt the underlying structure. Let a (resp. b, c, d, e, f) denote a category of x_1 (resp. x_2, x_3, x_4, x_5, x_6) and $\mathbb{P}$ denote a probability measure. To generate a contingency table, the following log-linear model (see for instance Agresti [1]) is simulated:

$$\begin{aligned}\log(\mathbb{P}(x_1 = a, \ldots, x_6 = f)) =&(\lambda_a^{x_1} + \lambda_b^{x_2} + \beta_{ab}^{x_1 x_2}) + \\ &(\lambda_c^{x_3} + \lambda_d^{x_4} + \beta_{cd}^{x_3 x_4}) + + (\lambda_e^{x_5} + \lambda_f^{x_6} + \beta_{ef}^{x_5 x_6}) + \beta_{ad}^{x_1 x_4} + \beta_{cf}^{x_3 x_6} \qquad (4)\end{aligned}$$

where $a, b, c, d, e, f \in \{0, 1\}$. The parameters $\lambda_a^{x_1}, \lambda_b^{x_2}, \lambda_c^{x_3}, \lambda_d^{x_4}, \lambda_e^{x_5}, \lambda_f^{x_6}$ represent the effect of each variable and the parameters $\beta_{ab}^{x_1 x_2}, \beta_{cd}^{x_3 x_4}, \beta_{ef}^{x_5 x_6}$ are interactions corresponding with cohesion terms in each group. The parameter $\beta_{ad}^{x_1 x_4}$ (resp. $\beta_{cf}^{x_3 x_6}$) is used to add some interactions between categories of variables belonging to different groups $\mathcal{Q}_1$ and $\mathcal{Q}_2$ (resp. $\mathcal{Q}_2$ and $\mathcal{Q}_3$). The first state of mixing corresponds to the initial partition and is called "no mixing". Then we moderately mix the two groups by increasing the value of $\beta_{00}^{x_1 x_4}$, it will be refered as "moderate mixing". In the third case named "strong mixing", the value of $\beta_{00}^{x_1 x_4}$ is high. In the last state called "very strong mixing", the values of $\beta_{00}^{x_1 x_4}$ and $\beta_{00}^{x_3 x_6}$ are high. Thus there is no more structure in the data.

For each state of mixing we simulate $N = 50$ contingency tables, each corresponding to a global sample size $n = 2000$ using log-linear model (4), where the values of the parameters are given in Table 1. Only the nonnull parameter values are specified, all the remaining ones are set to zero. In this table the value h of the effect parameters $\lambda_0^{x_2}, \lambda_0^{x_4}, \lambda_0^{x_6}$ is generated with the univariate uniform distribution on $[1, 1.5]$ to get N slightly different contingency tables.

We apply the proposed algorithm on the generated categorical data.

- When there is no mixing between the groups, the proposed approach always recovers the underlying clusters.
- When the mixing between the groups is moderate, the algorithm misclassifies one variable. We always obtain the partition $\{\{x_1, x_2, x_4\}, \{x_3\}, \{x_5, x_6\}\}$.
- When two groups are strongly mixed, the algorithm always misclassifies two variables. The corresponding partition is $\{\{x_1, x_4\}, \{x_2, x_3\}, \{x_5, x_6\}\}$.

Table 1. Values of the parameters of model (4) used in the simulations

State of mixing	no mixing	moderate mixing	strong mixing	very strong
Effect of each variable	$\lambda_0^{x_1} = \lambda_0^{x_3} = \lambda_0^{x_5} = 1$ $\lambda_0^{x_2} = \lambda_0^{x_4} = \lambda_0^{x_6} = h \in [1, 1.5]$			
Cohesion terms	$\beta_{00}^{x_1x_2} = -1.5$ $\beta_{00}^{x_3x_4} = -1.1$ $\beta_{00}^{x_5x_6} = -0.9$	$\beta_{00}^{x_1x_2} = -1.5$ $\beta_{00}^{x_3x_4} = -1.2$ $\beta_{00}^{x_5x_6} = -1$	$\beta_{00}^{x_1x_2} = -0.8$ $\beta_{00}^{x_3x_4} = -0.7$ $\beta_{00}^{x_5x_6} = -0.9$	
Interaction terms	0	$\beta_{00}^{x_1x_4} = -0.9$	$\beta_{00}^{x_1x_4} = -1.5$	$\beta_{00}^{x_1x_4} = 0.9$ $\beta_{00}^{x_3x_6} = -1.5$

- When the mixing is very strong, not surprisingly the algorithm misclassifies three variables since there is no more visible structure in the data. The obtained partition is always $\{\{x_1\}, \{x_2, x_4, x_5\}, \{x_3, x_6\}\}$.

3.2 Real Data Application

We consider a real data set on a user satisfaction survey of pleasure craft operators on the "Canal des Deux Mers" located in South of France which contains numerous questions with numerical or categorical answers. This study has been realized from June to September 2008. In this application we only focus on fourteen categorical variables described in Table 2. The sample size is $n = 709$ pleasure craft operators.

Table 2. Description of the 14 categorical variables

Name of the variable	**Description of the variable**	**Categories**
x_1="sites worth visiting"	*What do you think about information you were provided with concerning sites worth visiting?*	satisfactory, unsatisfactory, no opinion
x_2="leisure activity"	*How would you rate the information given on leisure activity?*	
x_3="historical canal sites"	*What is your opinion concerning tourist information on historical canal sites (locks, bridges, etc.)?*	
x_4="manoeuvres"	*At the start of your cruise, were you sufficiently aware of manoeuvres at locks?*	yes, no
x_5="authorized mooring"	*At the start of your cruise, were you sufficiently aware of authorized mooring?*	
x_6="safety regulations"	*At the start of your cruise, were you sufficiently aware of safety regulations?*	
x_7="services"	*Please give us your opinion about signs you encountered along the way concerning information regarding services.*	satisfactory, unsatisfactory
x_8="number of taps"	*What do you think about number of taps on your trip?*	sufficient, unsufficient
x_9="cost of water"	*The general cost of water is ...*	inexpensive, average, expensive
x_{10}="cost of electricity"	*The general cost of electricity is ...*	
x_{11}="visibility of electrical outlets"	*What is your opinion of visibility of electrical outlets?*	sufficient, unsufficient
x_{12}="number of electrical outlets"	*What do you think about number of electrical outlets on your trip?*	
x_{13}="cleanliness"	*How would you describe the canal's degree of cleanliness?*	clean, average, dirty
x_{14}="unpleasant odours"	*Were there unpleasant odours on the canal?*	none, occasional, frequent

In this case study, we have chosen to retain $K = 5$ clusters because it provides a satisfactory mean correlation ratio value (0.68), that is the mean

of the correlation ratio between the variables in each cluster and the corresponding latent variable. Moreover the interpretation of the clusters seems to be sound. This choice has also been confirmed by a bootstrap approach which consists in generating multiple data replications of the data set and examining if the partition is stable. Table 3 describes the 5-clusters partition of the variables. For instance cluster 4 contains variables dealing with the use of the canal. As has already been pointed, MCA is used to have a first solution to start the algorithm. Comparing the obtained solution with the MCA solution shows that cluster 1 and 4 are merged and that only one iteration is needed to obtain convergence to a local optimum corresponding to the partition given in Table 3. The value in brackets of this table corresponds to the correlation ratio between the variable and the latent variable representing the cluster it belongs to. We see that the variables in a cluster are highly related with their latent variable. Table 4 gives the values of the Tschuprow coefficient between the variables of cluster 4 $\{x_1, x_2, x_3\}$ and the remaining ones. We see that the variables are more related with variables in the same cluster than with the variables in the other clusters. This means that dimension reduction is possible. For instance in this case study we could reduce the number of the questions in the survey by selecting one question in each cluster. Furthermore we could replace the classical previous step of MCA for the clustering of the individuals by the construction of the latent variables.

Table 3. Partition of the 14 categorical variables into 5 clusters (correlation ratio between the variable and the latent variable of the cluster)

$\mathcal{C}_1$: **environment**	$\mathcal{C}_2$: **navigation rules**	$\mathcal{C}_3$: **cost of services**
cleanliness (0.68)	manoeuvres (0.66)	cost of water (0.84)
unpleasant odours (0.68)	authorized mooring (0.71)	cost of electricity (0.84)
	safety regulations (0.69)	
$\mathcal{C}_4$: **use of the canal**	$\mathcal{C}_5$: **available services**	
sites worth visiting (0.71)	services (0.40)	
leisure activity (0.69)	number of taps (0.59)	
historical canal sites (0.46)	visibility of electrical outlets (0.65)	
	number of electrical outlets (0.71)	

Table 4. Values of the Tschuprow coefficient between the variables of cluster 4 and the remaining ones

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	...	x_{14}
x_1	**1.00**	**0.36**	**0.24**	0.09	0.10	0.11	0.08	0.06	...	0.05
x_2	**0.36**	**1.00**	**0.20**	0.10	0.11	0.13	0.11	0.07	...	0.03
x_3	**0.24**	**0.20**	**1.00**	0.02	0.04	0.05	0.11	0.08	...	0.05

4 Concluding Remarks

In this paper we propose an extension of an existing center-based algorithm to the case of categorical variables. For numerical variables the homogeneity criterion is calculated with squared correlations between the variables of the cluster and its latent variable, which is defined as the first principal component issued from PCA. For categorical variables correlation ratios and MCA are then used respectively in place of squared correlations and PCA. The originality of the proposed approach lies in the fact that the center of a cluster of categorical variables is a numerical variable. A simulation study shows that the proposed method is efficient to recover simulated clusters of variables and a real data application illustrates the practical benefits of the approach.

The initialization of the algorithm is actually reached by computing the first K principal components issued from MCA. Another solution is to run several times the algorithm with multiple random initializations and to retain the best partition in sense of the homogeneity criterion. The initialization with MCA can also be coupled with a rotation to start with a better partition. For instance, the planar iterative rotation procedure proposed for MCA by Chavent and al. [2] can be used. Another interesting perspective would be to use this partitioning method in a divisive hierarchical approach to divide at best a cluster into two sub-clusters. Both research on ascendant and divisive hierarchical algorithms and a comparison of the different types of initialization for the partitioning method are currently under investigation.

Source codes of the implementation in `R` are available from the authors.

Acknowledgements. The authors are greatful to the public corporation "Voies Navigables de France" and the private firm Enform for providing the real data set.

References

1. A. Agresti. *Categorical data analysis*, Second Edition, Wiley Series in Probability and Statistics, 2002.
2. M. Chavent, V. Kuentz, and J. Saracco. Rotation in Multiple Correspondence Analyis: a planar rotation iterative procedure, *Submitted paper*, 2009.
3. I.S. Dhillon, E.M. Marcotte, and U. Roshan. Diametrical clustering for identifying anti-correlated gene clusters, *Bioinformatics*, **19**(13), pages 1612-1619, 2003.
4. M. Plasse, N. Nianga, G. Saporta, A. Villeminot, and L. Leblond. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set, *Computational Statistics and Data Analysis*, **52**(1), pages 596-613, 2007.
5. E. Vigneau, and E.M. Qannari. Clustering of variables around latent components, *Communications in statistics Simulation and Computation*, **32**(4), pages 1131-1150, 2003.