GREThA
Groupe de Recherche en
Économie Théorique et Appliquée

Clustering of categorical variables around latent variables

Jérome SARACCO

Université de Bordeaux

GREThA UMR CNRS 5113

Marie CHAVENT

Vanessa KUENTZ

Université de Bordeaux

Institut de Mathématiques (IMB), UMR CNRS 5251

Cahiers du GREThA
n° 2010-02

GRETHA UMR CNRS 5113
Université Montesquieu Bordeaux IV
Avenue Léon Duguit - 33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75 - Fax : +33 (0)5.56.84.86.47 - www.gretha.fr

Classification de variables qualitatives autour de variables latentes

Résumé

En classification, on s'intéresse habituellement à classifier les observations et non les variables. Cependant la classification de variables trouve tout son sens en réduction de dimension, pour la sélection de variables ou encore dans certaines applications (analyse sensorielle, biochimie, marketing, etc.). L'idée est alors de chercher des groupes de variables liées c'est-à-dire porteuses de la même information. Une fois que les variables sont organisées en groupes homogènes telles que les variables au sein d'une même classe sont similaires, il est alors possible de sélectionner dans chaque classe une variable ou de résumer chaque classe de variables par une variable synthétique, encore appelée variable latente. Plusieurs approches ont été spécifiquement développées pour la classification de variables quantitatives. Cependant, pour des données qualitatives, peu de méthodes ont été proposées. Dans cet article, nous étendons le critère proposé par Vigneau et Qannari (2003) dans leur méthode CLV (« Clustering around Latent Variables ») pour la classification de variables quantitatives au cas de données qualitatives. La variable latente d'une classe maximise l'homogénéité de la classe, définie comme la somme des rapports de corrélation entre les variables qualitatives de la classe et cette variable latente quantitative. Nous montrons que cette variable latente peut être obtenue par une Analyse des Correspondances Multiples des variables de la classe. Plusieurs algorithmes de classification utilisant le même critère d'homogénéité sont alors définis : algorithme de type nuées dynamiques, classification hiérarchique ascendante et descendante. Enfin ces différentes approches sont utilisées dans une étude de cas réelle concernant la satisfaction de navigants plaisanciers.

Mots-clés : classification de variables qualitatives, rapport de corrélation, algorithme des nuées dynamiques, classification hiérarchique

Clustering of categorical variables around latent variables

Abstract

In the framework of clustering, the usual aim is to cluster observations and not variables. However the issue of variable clustering clearly appears for dimension reduction, selection of variables or in some case studies (sensory analysis, biochemistry, marketing, etc.). Clustering of variables is then studied as a way to arrange variables into homogeneous clusters, thereby organizing data into meaningful structures. Once the variables are clustered into groups such that variables are similar to the other variables belonging to their cluster, the selection of a subset of variables is possible. Several specific methods have been developed for the clustering of numerical variables. However concerning categorical variables, much less methods have been proposed. In this paper we extend the criterion used by Vigneau and Qannari (2003) in their Clustering around Latent Variables approach for numerical variables to the case of categorical data. The homogeneity criterion of a cluster of categorical variables is defined as the sum of the correlation ratio between the categorical variables and a latent variable, which is in this case a numerical variable. We show that the latent variable maximizing the homogeneity of a cluster can be obtained with Multiple Correspondence Analysis. Different algorithms for the clustering of categorical variables are proposed: iterative relocation algorithm, ascendant and divisive hierarchical clustering. The proposed methodology is illustrated by a real data application to satisfaction of pleasure craft operators.

Keywords: clustering of categorical variables, correlation ratio, iterative relocation algorithm, hierarchical clustering

JEL : C49 ; C69

Reference to this paper: Jérome SARACCO, Marie CHAVENT, Vanessa KUENTZ, *"Clustering of categorical variables around latent variables"*, *Working Papers of GREThA*, n°2010-02, http://ideas.repec.org/p/grt/wpegrt/2010-02.html.

Clustering of categorical variables around latent variables

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

[1] Université de Bordeaux, IMB, CNRS, UMR 5251, France

[2] INRIA Bordeaux Sud-Ouest, CQFD team, France

[3] Université Montesquieu - Bordeaux IV, GREThA, CNRS, UMR 5113, France

Abstract

Clustering of variables is studied as a way to arrange variables into homogeneous clusters, thereby organizing data into meaningful structures. Once the variables are clustered into groups such that variables are similar to the other variables belonging to their cluster, the selection of a subset of variables is possible. Several specific methods have been developed for the clustering of numerical variables. However concerning categorical variables, much less methods have been proposed. In this paper we extend the criterion used by Vigneau and Qannari (2003) in their Clustering around Latent Variables approach for numerical variables to the case of categorical data. The homogeneity criterion of a cluster of categorical variables is defined as the sum of the correlation ratio between the categorical variables and a latent variable, which is in this case a numerical variable. We show that the latent variable maximizing the homogeneity of a cluster can be obtained with Multiple Correspondence Analysis. Different algorithms for the clustering of categorical variables are proposed: iterative relocation algorithm, ascendant and divisive hierarchical clustering. The proposed methodology is illustrated by a real data application to satisfaction of pleasure craft operators.

Keywords: clustering of categorical variables, correlation ratio, iterative relocation algorithm, hierarchical clustering.

1 Introduction

Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) are appealing statistical tools for multivariate description of respectively numerical and categorical data. Rotated principal components fulfill the need to get more interpretable components. Clustering of variables is an alternative since it makes it possible to arrange variables into homogeneous clusters and thus to obtain meaningful structures. From a general point of view, variable clustering lumps together variables which are strongly related to each other and thus bring the same information. Once the variables are clustered into groups such that attributes in each group reflect the same aspect, the practicioner may be spurred on to select one variable from each group. One may also want to construct a synthetic variable. For instance in the case of quantitative variables, a solution is to realize a PCA (see Jolliffe, 2002) in each cluster and to retain the first principal component as the synthetic variable of the cluster. Another advantage that may be gained from the clustering of variables relates to the selection of a subset of variables. It is an alternative to procedures for discarding or selecting variables based on a statistical criterion that have been proposed by Jolliffe (1972), Mc Cabe (1984), Krzanowski (1987), Al-Kandari and Jolliffe (2001) or Guo et al. (2002) among others. The selection of a subset of variables is the aim of a lot of research in several areas of application. For instance in descriptive sensory profiling, this strategy of analysis can be used to reduce a list of attributes by selecting relevant and non redundant attributes. In biochemistry clustering genes based upon their expression patterns allows to predict gene function. For preference studies when putting on the market new products, clustering of variables is also helpful to detect the existence of segments among the panel of consumers. Variable clustering can also be useful for association rules mining. Plasse et al. (2007) illustrate on an industrial application from the automotive industry the help of building homogeneous clusters of binary attributes for the discovering of relevant association rules mining. A conjoint use of variable clustering and Partial Least Squares (PLS) structural equations modeling is presented in Stan and Saporta (2005) in which clustering of variables is used to fulfill at best the underlying hypothesis in PLS approach of unidimensionality of the blocks of variables.

A simple and frequently used approach for variable clustering is to construct first a matrix of dissimilarities between the variables and then to apply classical cluster analysis methodology devoted to objects (units) which are able to deal with dissimilarity matrices (single, complete, average linkage hierarchical clustering or distance-based k-means). Partitioning Around Medoids can also deal with dissimilarity as input data (see Kaufman and Rousseeuw, 1990). Methods dealing only with numerical data like Ward or k-means among others can also be applied on the numerical coordinates obtained from Multidimensional Scaling of a previously built dissimilarity matrix.

Concerning quantitative variables, many authors have proposed different dissimilarity measures. Let us remind here some of these coefficients. Correlation coefficients (parametric or nonparametric) can be converted to different dissimilarities depending if the aim is to lump together correlated variables regardless of the sign of the correlation or if a negative correlation coeffcient between two variables shows disagreement between them. Soffritti (1999) defines a monotonous multivariate association measure that takes into account the within correlation and the number of variables of each group. A distance based on Escoufier's operator which takes the correlations as well as the variances of the variables into consideration has also been developped by Qannari et al. (1998). Note that this distance is also extended to the case of categorical variables and to a mixture of both types of data.

For categorical variables, many association measures can be used as χ^2, Rand, Belson, Jaccard, Sokal and Jordan among others. Some transformations are then in order to bring the coefficients into dissimilarity or distance measures. We can cite for instance the work of Abdallah and Saporta (1998) who consider various association measures and give the definition of a threshold beyond which two variables can be considered as linked.

Some specific approaches have also been developed for the clustering of variables. Once again for quantitative data, several specific methods have been proposed. We can cite among others the approach of Hastie et al. (2000) in genome biology or the recent work of Vichi and Saporta (2009), which aims at a simultaneous clustering of objects and a partitioning of variables. However the most famous one remains the VARCLUS procedure of SAS software. Two other interesting approaches that were independently proposed are Clustering around Latent Variables (CLV), introduced by Vigneau and Qannari (2003), and Diametrical Clustering of Dhillon et al. (2003). When the aim is to lump together correlated variables regardless of the sign of the correlation, both methods aim at maximizing the sum over all clusters of the squared correlations between the variables and a latent variable.

Let us now tackle the issue of specific methods developed in view of clustering of categorical variables. Surprisingly, it has received much less attention than the numerical case. As far as we know, only Likelihood Linkage Analysis proposed by Lerman (1993) is a specific method devoted to clustering of variables that can deal with both numerical and categorical data.

In this paper we propose specific methods for the clustering of categorical variables. The homogeneity criterion of a cluster is not simply a distance based criterion but an extension of that used in CLV (Vigneau and Qannari, 2003). It is equal to the sum of the correlation ratio between the categorical variables and a latent variable, which is in this case a numerical variable. We show that the latent variable maximizing the homogeneity of a cluster is the first principal component obtained by MCA (see Greenacre and Blasius, 2006) of the data of the cluster.

The overview of the paper is as follows. In Section 2, a specific measure of the homogeneity of a cluster of categorical variables is given and a partitioning criterion is defined. Section 3 is devoted to different clustering algorithms optimizing this specific criterion: iterative relocation algorithm, ascendant and divisive hierarchical clustering. In Section 4, a real data application relative to satisfactory of pleasure craft operators is treated. First the proposed hierarchical clustering algorithm is applied on a real data set. Then an empirical comparison of the performances of the different proposed algorithms is presented. Finally in Section 5, some concluding remarks and perspectives are given.

2 A correlation ratio based partitioning criterion for 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. Let $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ be the set of the p columns of $\mathbf{X}$, called for seek of simplicity categorical variables.

Homogeneity criterion of a cluster. Let $\mathcal{C} \subset \mathcal{V}$ be a cluster of categorical variables and $\mathbf{y}$ be a vector of $\mathbb{R}^n$ called latent variable. The homogeneity criterion of $\mathcal{C}$ measures the adequacy between the variables in $\mathcal{C}$ and $\mathbf{y}$:

$$S(\mathcal{C}) = \sum_{\mathbf{x}_j \in \mathcal{C}} \eta^2(\mathbf{x}_j, \mathbf{y}), \tag{1}$$

where $\eta^2(\mathbf{x}_j, \mathbf{y})$ stands for the correlation ratio between the categorical variable $\mathbf{x}_j$ and a numerical latent variable $\mathbf{y}$. This ratio is equal to the between group sum of squares of $\mathbf{y}$ in the groups defined by the categories of $\mathbf{x}_j$, divided by the total sum of squares of $\mathbf{y}$: $\eta^2(\mathbf{x}_j, \mathbf{y}) = \dfrac{\sum_{s \in \mathcal{M}_j} n_s (\bar{y}_s - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$, with n_s the frequency of category s, $\mathcal{M}_j$ the set of categories of $\mathbf{x}_j$ and $\bar{y}_s$ the mean value of $\mathbf{y}$ calculated on the objects belonging to category s. The correlation ratio belongs to $[0, 1]$ and measures the link between the categorical variable $\mathbf{x}_j$ and a numerical latent variable $\mathbf{y}$.

Definition of the latent variable of a cluster. In cluster $\mathcal{C}$, the latent variable $\mathbf{y}$ is defined to maximize the homogeneity criterion $S(\mathcal{C})$:

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

Result 1. The latent variable $\mathbf{y}$ of $\mathcal{C}$ is the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$, with $\widetilde{\mathbf{F}}$ defined in (3).

Proof. As $\eta^2(\mathbf{x}_j, \mathbf{u}) = \eta^2(\mathbf{x}_j, \alpha\mathbf{u})$, for any nonnull real α, the optimization problem (2) has an infinite set of solutions. We choose here to add the constraint $\mathbf{u}^t\mathbf{u} = 1$. To define the matrix $\widetilde{\mathbf{F}}$ we need to introduce usual notations from the theory of MCA. We can code the data of cluster $\mathcal{C}$ using indicator matrix $\mathbf{G}$ of dimension $n \times q$, with q the number of categories of the variables in $\mathcal{C}$, in which each category level is given a separate column and an entry of 1 indicates the relevant level of the category. The indicator matrix $\mathbf{G}$ is divided by its grand total $np_{\mathcal{C}}$, where $p_{\mathcal{C}}$ designates the number of variables in $\mathcal{C}$, to obtain the so-called "correspondence matrix" $\mathbf{F} = \frac{1}{np_{\mathcal{C}}}\mathbf{G}$, so that $\mathbf{1}_n^t\mathbf{F}\mathbf{1}_q = 1$, where, generically, $\mathbf{1}_i$ is an $i \times 1$ vector of ones. Furthermore, the row and column marginals define respectively the vectors of row and column masses $\mathbf{r} = \mathbf{F}\mathbf{1}_q$ and $\mathbf{c} = \mathbf{F}^t\mathbf{1}_n$. Let $\mathbf{D}_r = \text{diag}(\mathbf{r})$ and $\mathbf{D}_c = \text{diag}(\mathbf{c})$ be the diagonal matrices of these masses. In this particular case, the ith element of $\mathbf{r}$ is $f_{i.} = \frac{1}{n}$ and the sth element of $\mathbf{c}$ is $f_{.s} = \frac{n_s}{np_{\mathcal{C}}}$. We can now define the matrix

$$\widetilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D}_c^{-1/2}. \tag{3}$$

Let us first show that if $\bar{u} = 0$ and $\text{var}(\mathbf{u}) = \frac{1}{n}$, we have $\mathbf{u}^t\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t\mathbf{u} = \frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}} \eta^2(\mathbf{x}_j, \mathbf{u})$. Remembering from the definition of $\mathbf{F}$ that $f_{is} = \frac{g_{is}}{np_{\mathcal{C}}}$, the general term of $\mathbf{D}_r^{-1/2}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D}_c^{-1/2}$ is then $\tilde{f}_{is} = \frac{\sqrt{n_s}\sqrt{p_{\mathcal{C}}}}{n_s}(\frac{g_{is}}{p_{\mathcal{C}}} - \frac{n_s}{np_{\mathcal{C}}})$. It follows that $\sum_{i=1}^n \tilde{f}_{is}u_i = \frac{\sqrt{n_s}}{\sqrt{p_{\mathcal{C}}}}\bar{u}_s$, where $\bar{u}_s$ is the mean value of $\mathbf{u}$ calculated on the objects belonging to category s. Then we get

$$\mathbf{u}^t\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t\mathbf{u} = \frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}}\sum_{s\in\mathcal{M}_j} n_s\bar{u}_s^2 = \frac{\frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}}\sum_{s\in\mathcal{M}_j}\frac{n_s}{n}(\bar{u}_s - 0)^2}{\frac{1}{n}} = \frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}} \eta^2(\mathbf{x}_j, \mathbf{u}). \tag{4}$$

Moreover as the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$ maximizes $\mathbf{u}^t\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t\mathbf{u}$ with respect to $\mathbf{u} \in \mathbb{R}^n$ under the constraint $\mathbf{u}^t\mathbf{u} = 1$, it is a solution of (2). Since it is normalized, its variance is equal to $\frac{1}{n}$. Then we have to check that it is centered. If $\widetilde{\mathbf{F}}$ is supposed to be of rank r, the Singular Value Decomposition (SVD) of $\widetilde{\mathbf{F}}$ is $\widetilde{\mathbf{F}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t$, where $\mathbf{\Lambda}$ contains the r nonnull singular values of $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$ and $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$ sorted in decreasing order, $\mathbf{U}$ (resp. $\mathbf{V}$) is the matrix whose columns are the normalized eigenvectors of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$ (resp. $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$) associated with the nonnull eigenvalues. Thus $\mathbf{U} = \widetilde{\mathbf{F}}\mathbf{V}\mathbf{\Lambda}^{-1}$ and then the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$, as a linear combination of the columns of $\widetilde{\mathbf{F}}$ which are centered, is in turn centered, which completes the proof.

Result 2. The latent variable $\mathbf{y}$ is colinear with the first principal component issued from MCA of the row profiles of the data matrix of $\mathcal{C}$.

Proof. MCA is defined here as the application of weighted PCA to the centered row profiles matrix $\mathbf{D}_r^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)$ with distances between profiles measured by the chi-squared metric defined by $\mathbf{D}_c^{-1}$. The $n \times r$ matrix Ψ of row principal coordinates is then defined by $\Psi = \mathbf{D}_r^{-1/2}\widetilde{\mathbf{F}}\mathbf{V}$, with the expression of $\widetilde{\mathbf{F}}$ given in (3). From the SVD of $\widetilde{\mathbf{F}}$, we get $\Psi = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{\Lambda}$, thereby implying that the latent variable, defined as the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$, is colinear with the first principal component obtained with MCA.

Partitioning criterion. We denote by $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ a partition of $\mathcal{V}$ into K clusters and by $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ a set of K latent variables. The paper addresses the problem of partitioning a set of p variables into K disjoint clusters in which variables are similar to the other variables belonging to their cluster and dissimilar to variables that belong to different clusters. The partitioning criterion concentrates on maximizing the cohesion (homogeneity) of the clusters in the partition:

$$H(\mathcal{P}_K) = \sum_{k=1}^{K} S(\mathcal{C}_k). \tag{5}$$

with $S(C_k)$ defined in (1). In the next section, we propose different clustering algorithms using this criterion.

3 Different clustering algorithms

Given criterion (5) measuring the homogeneity of a partition of a set of variables into K disjoint clusters, there are different possible clustering algorithms for maximizing this criterion. First we describe an iterative relocation algorithm, then two hierarchical algorithms are proposed: ascendant and divisive.

Iterative relocation algorithm. A first solution to search for optimal partitions of the variables is given by an iterative algorithm in the course of which the variables are allowed to move in and out of the groups at the different stages of the algorithm achieving at each stage an increase of criterion (5). This partitioning algorithm runs as follows:

(a) *Initialization step*: The specification of this step may be reached by different ways. The first solution consists in computing the first K principal components issued from MCA of the centered row profiles matrix of $\mathbf{X}$. As has been described in Section 2, each component can play the role of the latent variable of a cluster with itself as single member. Then we go to step (c) for the allocation step. This initialization can be coupled with a rotation to start with a better partition as in the VARCLUS procedure. We can use for instance the planar rotation iterative procedure for rotation in MCA proposed by Chavent et al. (2009). By doing this, the values of the correlation ratio between the variables and the latent variables are either large or small and the allocation is easier and then may be better. Another solution is to select randomly K variables of $\mathcal{V}$ and to apply MCA on the row profiles obtained with the data provided by each single variable in order to get K latent variables. These latent variables define at the beginning K clusters each containing only one member. Then we go to step (c). As it is well-known that iterative relocation algorithms provide a local optimum, the proposed iterative relocation algorithm is run several times, with multiple random initializations and we retain the best partition in sense of our partitioning criterion (5).

(b) *Representation step*: For all k in $1, ..., K$, we compute the latent variable $\mathbf{y}_k$ of $\mathcal{C}_k$ as the first normalized eigenvector of $\widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t$, where $\widetilde{\mathbf{F}}_k$ is defined in (3) for a generic cluster.

(c) *Allocation step*: Each variable is then assigned to the cluster which latent variable is closest to it in sense of correlation ratio. For all j in $1, ..., p$, find ℓ such that $\ell = \arg \max_{k=1,...,K} \eta^2(\mathbf{x}_j, \mathbf{y}_k)$. 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 step (c) then *stop*, else return to step (b).

An empirical comparison of the efficiency of the iterative relocation algorithm according to the initialization step (a) is provided in Section 4.

Ascendant hierarchical approach. We propose herein a hierarchical clustering strategy based on the same criterion (5). First from Result 1, this criterion can be rewritten as follows:

$$H(\mathcal{P}_K) = \sum_{k=1}^{K} p_k \lambda_k, \tag{6}$$

where p_k is the number of variables in $\mathcal{C}_k$ and λ_k is the largest eigenvalue of matrix $\widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t$, with $\widetilde{\mathbf{F}}_k$ defined in (3) for a generic cluster.

In the ascendant hierarchical clustering algorithm, one recursively merges two clusters, starting from the stage in which each variable is considered to form a cluster by itself to the stage where there is a single cluster containing all variables. Given the current partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, two clusters are merged in order to find a partition $\mathcal{P}_{K-1}$ which contains $K-1$ clusters and optimizes the chosen cohesion measure (6). More precisely because

$$H(\mathcal{P}_{K-1}) = H(\mathcal{P}_K) - \underbrace{(S(\mathcal{C}_l) + S(\mathcal{C}_m) - S(\mathcal{C}_l \cup \mathcal{C}_m))}_{h(\mathcal{C}_l \cup \mathcal{C}_m)}, \tag{7}$$

the merging of two clusters $\mathcal{C}_l$ and $\mathcal{C}_m$ results in a variation of criterion (6) given by:

$$h(\mathcal{C}_l \cup \mathcal{C}_m) = \lambda_l + \lambda_m - \lambda_{l \cup m}. \tag{8}$$

We can prove (see Appendix) that:

$$\lambda_{l \cup m} \leqslant \lambda_l + \lambda_m, \tag{9}$$

which implies that the merging of two clusters at each step results in a decrease in criterion (6). Therefore the strategy consists in merging the two clusters that result in the smallest decrease in the cohesion measure.

Divisive hierarchical approach. Divisive hierarchical clustering reverses the process of agglomerative hierarchical clustering, by starting with all variables in one cluster, and successively dividing each cluster into two sub-clusters. Given the current partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, one cluster $\mathcal{C}_l$ is split in order to find a partition $\mathcal{P}_{K+1}$ which contains $K+1$ clusters and optimizes the chosen adequacy measure (6). More precisely, at each stage, the divisive hierarchical clustering method

- splits a cluster $\mathcal{C}_l$ into a bipartition $(\mathcal{A}_l, \bar{\mathcal{A}}_l)$;
- chooses in the partition $\mathcal{P}_K$ the cluster $\mathcal{C}_l$ to be split in such a way that the new partition $\mathcal{P}_{K+1}$ has a maximum cohesion measure.

The problem of how to split a cluster. In order to split optimally a cluster $\mathcal{C}_l$ one has to choose the bipartition $(\mathcal{A}_l, \bar{\mathcal{A}}_l)$, amongst the $2^{p_l - 1} - 1$ possible bipartitions of this cluster of p_l variables (with p_l the number of variables in $\mathcal{C}_l$), which maximizes criterion (6). It is clear that such complete enumeration provides a global optimum but is computationally prohibitive. The iterative relocation algorithm proposed above can then be used to get a partition into two clusters which is locally optimal for criterion (6).

Selecting the cluster to be split. In divisive clustering, the set of clusters obtained after $K-1$ divisions is a hierarchy $\mathcal{H}_K$ whose singletons are the K clusters of the partition $\mathcal{P}_K$ obtained in the last stage of the procedure. Because the resulting hierarchy can be considered as a partial hierarchy halfway between the top and bottom levels, it is referred to as an upper hierarchy (Mirkin, 2005). This upper hierarchy is then indexed by h so that in the dendrogram the height of a cluster $\mathcal{C}_l$ split into two sub-clusters $\mathcal{A}_l$ and $\bar{\mathcal{A}}_l$ is:

$$h(\mathcal{C}_l) = S(\bar{\mathcal{A}}_l) + S(\mathcal{A}_l) - S(\mathcal{C}_l).$$

When the divisions are continued until giving singleton clusters, all of the clusters can be systematically split and the full hierarchy $\mathcal{H}_n$ can be indexed by h. When the divisions are not continued down to $\mathcal{H}_n$, the clusters are not systematically split: in order to have the dendrogram of the upper hierarchy $\mathcal{H}_K$ built at the "top" (the $K-1$ largest) levels of the dendrogram of $\mathcal{H}_n$, a cluster represented higher in the dendrogram of $\mathcal{H}_n$ has to be split before the others. The proposed procedure then chooses to split the cluster $\mathcal{C}_l$ with the maximum value $h(\mathcal{C}_l)$. Consequently because

$$\mathcal{H}(\mathcal{P}_{K+1}) = \mathcal{H}(\mathcal{P}_K) + h(\mathcal{C}_l)$$

maximizing $h(\mathcal{C}_l)$ ensures that the new partition $\mathcal{P}_{K+1} = \mathcal{P}_K \cup \{\mathcal{A}_l, \bar{\mathcal{A}}_l\} - \{\mathcal{C}_l\}$ has a maximum cohesion measure.

Remark. The index h of the hierarchy in (8) is well positive (see Appendix for the proof) but we have not yet demonstrated that it is a monotone increasing function, that is $\forall \mathcal{A}, \mathcal{B} \in \mathcal{H}$, if $\mathcal{A} \subset \mathcal{B}$, then $h(\mathcal{A}) \leq h(\mathcal{B})$. Note that in practice, we have never observed inversion phenomenom.

4 Real data application

In the subsequent clustering of categorical variables is applied to a real data set. A user satisfaction survey of pleasure craft operators on the "Canal des Deux Mers", located in South of France, was carried out by the public corporation "Voies Navigables de France" responsible for managing and developing the largest network of navigable waterways in Europe. This study was realized from June to December 2008. Pleasure craft operators were asked their opinion about numerous questions with categorical answers, thus providing $p = 85$ categorical variables, each having two or three categories of response. The objective of the present case study is to examine the redundancy among variables in order to select a subset of attributes to be used in further studies saving time for the respondents, money for the edition of the questionnaires and the statistical treatment of the data.

First an application is reached on a reduced[1] data set to illustrate the interpretation of the results obtained with the proposed ascendant hierarchical clustering algorithm. Then the different algorithms of clustering (iterative relocation algorithm and its various initializations, ascendant and divisive hierarchical clustering) are applied on the complete data set to compare empirically the advantages of each approach.

4.1 Illustration on a reduced data set

We focus here on fourteen categorical variables described in Table 1. After removal of individuals with missing values for some of the questions, the sample size is $n = 709$ pleasure craft operators.

[1] We only consider here a subset of 14 variables over the 85 categorical variables.

Name of the variable	**Description of the variable**	**Categories**
$\mathbf{x}_1$=“sites worth visiting”	*What do you think about information you were provided with concerning sites worth visiting?*	satisfactory, unsatisfactory, no opinion
$\mathbf{x}_2$=“leisure activity”	*How would you rate the information given on leisure activity?*	
$\mathbf{x}_3$=“historical canal sites”	*What is your opinion concerning tourist information on historical canal sites (locks, bridges, etc.)?*	
$\mathbf{x}_4$=“manoeuvres”	*At the start of your cruise, were you sufficiently aware of manoeuvres at locks?*	yes, no
$\mathbf{x}_5$=“authorized mooring”	*At the start of your cruise, were you sufficiently aware of authorized mooring?*	
$\mathbf{x}_6$=“safety regulations”	*At the start of your cruise, were you sufficiently aware of safety regulations?*	
$\mathbf{x}_7$=“information on services”	*Please give us your opinion about signs you encountered along the way concerning information regarding services.*	satisfactory, unsatisfactory
$\mathbf{x}_8$=“number of taps”	*What do you think about number of taps on your trip?*	sufficient, unsufficient
$\mathbf{x}_9$=“cost of water”	*The general cost of water is ...*	inexpensive, average, expensive
$\mathbf{x}_{10}$=“cost of electricity”	*The general cost of electricity is ...*	
$\mathbf{x}_{11}$=“visibility of electrical outlets”	*What is your opinion of visibility of electrical outlets?*	sufficient, unsufficient
$\mathbf{x}_{12}$=“number of electrical outlets”	*What do you think about number of electrical outlets on your trip?*	
$\mathbf{x}_{13}$=“cleanliness”	*How would you describe the canal’s degree of cleanliness?*	clean, average, dirty
$\mathbf{x}_{14}$=“unpleasant odours”	*Were there unpleasant odours on the canal?*	none, occasional, frequent

Table 1: Description of the 14 categorical variables.

The ascendant hierarchical approach described in Section 3 is applied. Figure 1 shows the resulting dendrogram. The evolution of the aggregation criterion h is given in Figure 2. This figure should be read as a scree-graph. The aggregation criterion jumped when passing from 5 clusters to 4 clusters. This should suggest that “different” clusters are being merged and therefore the partition into 5 clusters is retained. The choice of the number of clusters can also be based on practical considerations such as the easiness of interpretation. Here the partition into 5 clusters provides satisfactory interpretable results. In a subsequent stage, the iterative relocation algorithm is performed with $K = 5$ clusters with as initial partition the one derived from the hierarchical procedure. In this case study, this complement stage leads to no improvement of criterion (5) as no variable changes membership.

Table 2 describes the 5-clusters partition of the 14 categorical variables. For instance cluster $\mathcal{C}_4$ contains variables dealing with the information on the use of the canal: sites worth visiting, leisure activity and historical canal sites. The value in brackets shows the correlation ratio between a variable of the cluster and the corresponding latent variable. We see that the variables in a cluster are highly related with their latent variable. Table 3 gives the values of the Tschuprow coefficient between the variables of cluster $\mathcal{C}_4 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ and the remaining ones. We see that the variables are more related to the other variables belonging to their cluster than to variables that belong to different clusters. Then an advantage which may be gained from the clustering of variables relates to the selection of a subset of variables. For instance in this case study we could reduce the number of questions in the survey by selecting one variable in each cluster using the correlation ratio values given in Table 2.

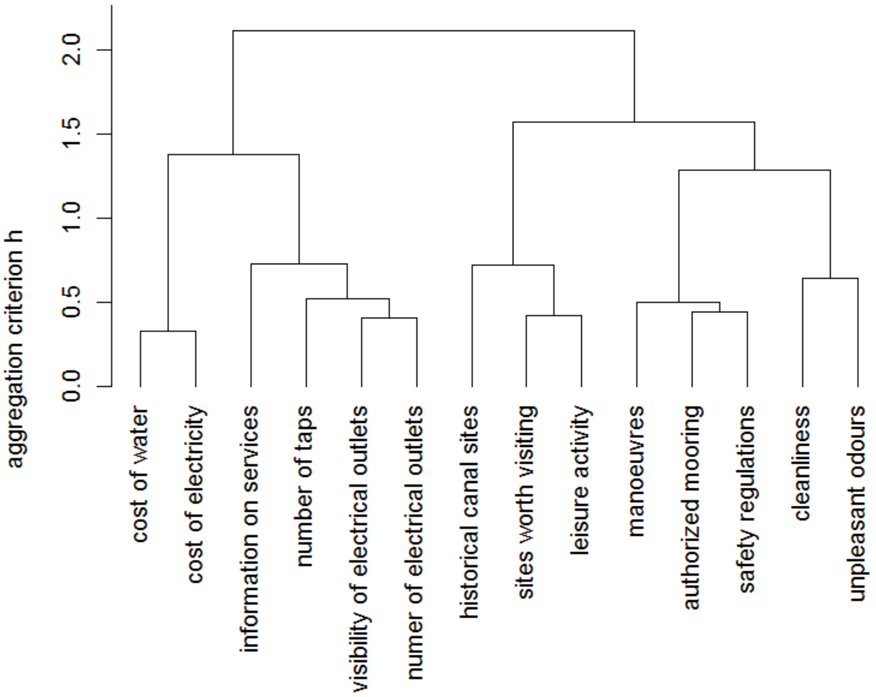

Figure 1: Dendrogram of the ascendant hierarchical clustering of the 14 categorical variables.

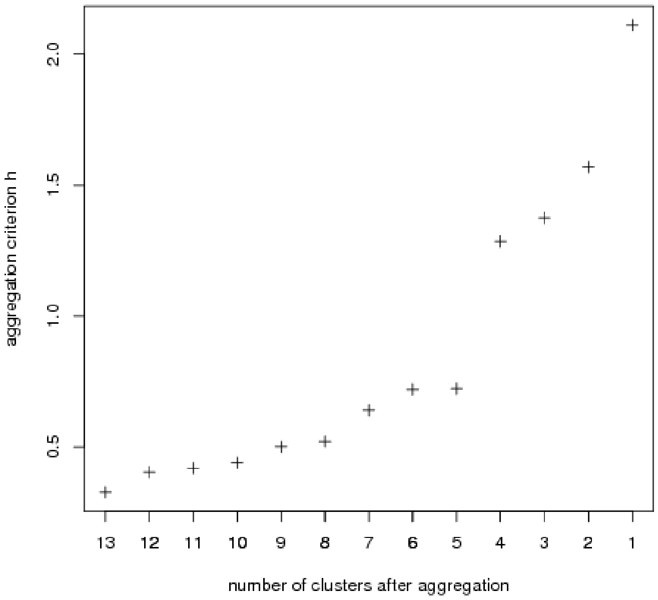

Figure 2: Evolution of the aggregation criterion h of the ascendant hierarchical clustering of the 14 categorical variables.

4.2 Empirical study and comparison of the different proposed clustering algorithms

We focus here on all the $p = 85$ categorical variables from the survey.

<table>
<tr><th>$\mathcal{C}_1$: <b>environment</b></th><th>$\mathcal{C}_2$: <b>navigation rules</b></th><th>$\mathcal{C}_3$: <b>cost of services</b></th></tr>
<tr><td>cleanliness (0.68)
unpleasant odours (0.68)</td><td>manoeuvres (0.66)
authorized mooring (0.71)
safety regulations (0.69)</td><td>cost of water (0.84)
cost of electicity (0.84)</td></tr>
<tr><th>$\mathcal{C}_4$: <b>use of the canal</b></th><th colspan="2">$\mathcal{C}_5$: <b>available services</b></th></tr>
<tr><td>sites worth visiting (0.71)
leisure activity (0.69)
historical canal sites (0.46)</td><td colspan="2">information on services (0.40)
number of taps (0.59)
visibility of electrical outlets (0.65)
number of electrical outlets (0.71)</td></tr>
</table>

Table 2: Partition of the 14 categorical variables into 5 clusters (correlation ratio between a variable of the cluster and the corresponding latent variable).

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	...	$\mathbf{x}_{14}$
$\mathbf{x}_1$	**1.00**	**0.36**	**0.24**	0.09	0.10	0.11	0.08	0.06	...	0.05
$\mathbf{x}_2$	**0.36**	**1.00**	**0.20**	0.10	0.11	0.13	0.11	0.07	...	0.03
$\mathbf{x}_3$	**0.24**	**0.20**	**1.00**	0.02	0.04	0.05	0.11	0.08	...	0.05

Table 3: Values of the Tschuprow coefficient between the variables of $\mathcal{C}_4$ and the remaining ones.

The proportion of explained cohesion. The clustering objective is formally expressed as the maximization of criterion (5) which can be perceived as a cohesion measure of the clusters in the partition. The cohesion criterion of a given partition $\mathcal{P}_K$ is given by $\mathcal{H}(\mathcal{P}_K) = \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{y}_k)$, with $\mathbf{y}_k$ the latent variable of cluster $\mathcal{C}_k$. Similarly the total cohesion of a set $\mathcal{V}$ of p variables can be measured by $\mathcal{H}(\mathcal{V}) = \sum_{j=1}^{p} \eta^2(\mathbf{x}_j, \mathbf{y})$ with $\mathbf{y}$ the latent variable (or total representative) of $\mathcal{V}$. The cohesion measure is equal to $\mathcal{H}(\mathcal{V})$ for the single cluster $(\mathcal{V})$ and to p for the singleton partition. Hence the quality of the partitions $\mathcal{P}_K$ built by the three methods from the same set of variables $\mathcal{V}$, can be ranked using the proportion of gain in cohesion, that is the ratio of the gain obtained with $\mathcal{P}_K$ to the maximum gain that can be reached with the singleton partition:

$$E(\mathcal{P}_K) = \frac{\mathcal{H}(\mathcal{P}_K) - \mathcal{H}(\mathcal{V})}{p - \mathcal{H}(\mathcal{V})}.$$

This lies between 0% for the single cluster $(\mathcal{V})$ and 100% for the singleton partition. Because E increases with the number K of clusters of the partition, it can be used only to compare partitions having the same number of clusters. In the following, we assume that a partition $\mathcal{P}_K$ is better than a partition $\mathcal{P}'_K$ if $E(\mathcal{P}_K) > E(\mathcal{P}'_K)$. We will call $E(\mathcal{P}_K)$ the proportion of explained cohesion by the partition $\mathcal{P}_K$.

Different initializations of the iterative relocation algorithm. As has already been pointed, the iterative relocation algorithm involves an initialization step that can be specified for instance by the three techniques proposed in Section 3. The aim of the following is to study the impact of the initialization on the quality of the obtained partition. Table 4 gives the proportion $E(\mathcal{P}_K)$ of explained cohesion for partitions from $K = 2$ to 20 clusters. Each column displays this proportion obtained respectively with the initialization via the first K principal components, the first K rotated principal components and the best of $N = 30$ random initializations.

The partitions obtained with the initialization via the rotated principal components are always better (except for $K = 5$ where it is almost equal) than those obtained with the principal components. Thus the complement step of rotation seems to be efficient. For the third column, the iterative relocation algorithm is

K	K principal components	K rotated principal components	$N = 30$ random initializations
2	3.19	**3.39**	1.48
3	5.95	**6.40**	5.25
4	8.03	**8.87**	8.57
5	10.55	10.13	**11.60**
6	11.86	12.48	**14.62**
7	14.29	14.94	**17.13**
8	15.70	17.74	**18.87**
9	17.85	18.24	**21.22**
10	19.67	20.87	**23.83**
11	21.26	22.18	**25.80**
12	22.46	24.66	**27.76**
13	23.69	26.31	**29.16**
14	24.89	27.68	**31.41**
15	26.47	28.21	**33.51**
16	27.66	29.71	**35.33**
17	29.46	31.16	**37.05**
18	29.92	32.56	**38.21**
19	31.46	34.16	**40.53**
20	32.68	35.74	**42.39**

Table 4: Iterative relocation algorithm: comparison of the proportion $E(\mathcal{P}_K)$ of explained cohesion with various initializations.

executed $N = 30$ times with different random initial seeds and the best solution in sense of the partitioning criterion (5) is retained. The partitions obtained with the rotated principal components are better up to 4 clusters and the iterative relocation algorithm with random initializations takes the lead from 5 clusters onwards. Moreover the gain in the proportion of explained cohesion increases as the number of clusters increases (18.6%=(42.39-35.74)/35.74 for 20 clusters versus 14.5%=(11.60-10.13)/10.13 for 5 clusters). Note that one possible explanation for the worse results of the multiple random initializations is probably that there is no strong structure in the data for a small number K of clusters so that the draw of some random initial seeds does not provide good partitions. As a rule concerning the iterative relocation methodology, running the algorithm several times with different initial partition in each run seems to be a satisfactory strategy.

Comparison of the different approaches. Now, we compare the results of the iterative relocation algorithm with multiple random initializations, which provides the best partitions in sense of $E(\mathcal{P}_K)$, with ascendant and divisive hierarchical clustering.

Comparing the first two columns of Table 5, we see that the ascendant hierarchical clustering is more efficient than the divisive one. A possible explanation is that the agglomerative algorithm is “stepwise optimal”: at each step, the amalgamation chosen is the best (in terms of the specified clustering criterion) that can be made at that time. However one reason for having worse results for the divisive approach is probably the way of splitting a cluster into two sub-clusters. This is reached by iterative relocation algorithm (with $N = 30$ multiple random initializations) and thus the bipartition obtained may not be optimal, thus altering the quality of the hierarchy built with the divisive clustering.

Then we compare the results obtained with the ascendant hierarchical procedure with those reached with the iterative relocation algorithm (with $N = 30$ random initial seeds). The latter always provides better

K	ascendant hierarchical clustering	divisive hierarchical clustering	iterative relocation algorithm	ascendant hierarchical algorithm
			($N = 30$ random initializations)	+ iterative relocation
2	3.01	2.58	1.48	**3.26**
3	5.73	4.51	5.25	**6.18**
4	8.19	7.31	8.57	**9.05**
5	10.63	9.31	11.60	**11.62**
6	12.92	10.95	**14.62**	13.99
7	15.13	12.36	**17.13**	15.99
8	17.19	13.61	**18.87**	17.98
9	19.23	14.92	**21.22**	19.83
10	21.24	16.62	**23.83**	21.88
11	23.09	18.62	**25.80**	23.67
12	24.93	19.72	**27.76**	25.45
13	26.72	21.14	**29.16**	27.35
14	28.48	22.61	**31.41**	29.07
15	30.16	23.87	**33.51**	30.73
16	31.78	25.40	**35.33**	32.03
17	33.38	26.73	**37.05**	33.63
18	34.92	28.09	**38.21**	35.05
19	36.45	29.38	**40.53**	36.54
20	37.94	30.95	**42.39**	38.03

Table 5: Comparison of the proportion $E(\mathcal{P}_K)$ of explained cohesion with different algorithms of clustering.

partitions in sense of the cohesion measure (5), except as seen previously for a small number of clusters ($K = 2, 3$). Once again the gain in the proportion of explained cohesion increases as the number of clusters increases (11.2% for 20 clusters versus 4.6% for 4 clusters). However one may prefer the hierarchical technique which has the advantage to build a hierarchy of nested partitions of the variables and then may be beneficial for the interpretation of the results and the choice of a number K of clusters.

We also propose in the fourth column of Table 5 to complement the ascendant hierarchical clustering by the iterative relocation algorithm with as initial partition the one derived from the hierarchical procedure. For a given partition $\mathcal{P}_K$ this step aims at improving criterion (5) by allowing variables to change membership. Thus for each number of clusters $K = 2, \ldots, 20$, we see that the new partitions obtained are better than the initial ones (first column). However the iterative relocation algorithm (with $N = 30$ random initializations) takes the lead from $K = 6$ clusters onwards.

5 Concluding remarks

This paper proposes an extension of an existing criterion for the clustering of numerical variables (Vigneau and Qannari, 2003) to the case of categorical data. The partitioning criterion measuring the cohesion of the clusters in the partition is based on correlation ratio between the categorical variables of the cluster and a numerical latent variable. The latent variable of a cluster which optimizes the homogeneity criterion of a cluster is computed from MCA. Several algorithms for the clustering of categorical variables using the proposed partitioning criterion are described (iterative relocation algorithm, ascendant and divisive hierarchical clustering).

The results obtained with the proposed approach are illustrated and interpretated on a real data set. An empirical comparison of the different clustering approaches is also derived on this data set. We see on the proposed case study that the partitioning criterion may have several local optima. Then concerning the iterative relocation algorithm, the multiple random intitializations provides the best partitions in sense of proportion of explained cohesion. The divisive hierarchical clustering suffers from multiple local optima of the iterative relocation algorithm when splitting a cluster into two sub-clusters and then provides worse results than the ascendant hierarchial clustering or iterative relocation algorithm. Surprisingly the iterative relocation algorithm provides better results than the ascendant hierarchical clustering complemented by an iterative relocation of the variables. However one advantage of the hierarchical procedure is the easier interpretability of the results since it produces a hierarchy of nested partitions of the variables. The proposed algorithms have been implemented in $\mathcal{R}$ and source codes are available from the authors.

Furthermore a classical approach in data mining consists in carrying out a MCA and subsequently applying a clustering algorithm on the component scores of the objects, thereby using the first few components only. However DeSarbo et al. (1990), De Soete and Caroll (1994) and Vichi and Kiers (2001) warn against this approach, called “tandem analysis”, because MCA may identify dimensions that do not necessarily contribute much to perceiving the clustering structure in the data and that, on the contrary, may obscure or mask the taxonomic information. Cluster analysis of variables is then an alternative technique as it makes it possible to organize the data into meaningful structures. Therefore the construction of latent variables may be more efficient that the classical MCA step.

One remaining point to study is the monotony of the proposed partitioning criterion. Another interesting aspect would be to compare the computational complexity of the different proposed algorithms. Concerning future prospects, the choice of the number of clusters with a bootstrap approach, consisting in generating multiple data replications of the data set and examining if the partition is stable, is currently under study. Research will also be undertaken on the treatment of missing values to avoid, as has been made in the presented real data application, deleting individuals who have returned questionnaires with the answers to some questions not completed.

Acknowledgements

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

Appendix: Proof of inequality (9)

We have

$$\begin{aligned}
\lambda_{l\cup m} &= \max_{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_{l\cup m}\widetilde{\mathbf{F}}^t_{l\cup m}\mathbf{u}\}\\
&= \max_{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_{l}\widetilde{\mathbf{F}}^t_{l}\mathbf{u}+\mathbf{u}^t\widetilde{\mathbf{F}}_{m}\widetilde{\mathbf{F}}^t_{m}\mathbf{u}\}\\
&\leq \max_{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_{l}\widetilde{\mathbf{F}}^t_{l}\mathbf{u}\}+\max_{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_{m}\widetilde{\mathbf{F}}^t_{m}\mathbf{u}\}\\
&= \lambda_l+\lambda_m.
\end{aligned}$$

where the definition of $\widetilde{\mathbf{F}}$ is given in (3) for a generic cluster.

References

Abdallah, H. and Saporta, G., (1998), Classification d'un ensemble de variables qualitatives, *Revue de Statistique Appliquée*, **46**(4), 5-26.

Al-Kandari, N.M., Jolliffe, I.T., (2001), Variable selection and interpretation of covariance principal components, *Communications in Statistics - Simulation and Computation*, **30**, 339-354.

Chavent, M., Kuentz, V., Saracco, J., (2009), Rotation in Multiple Correspondence Analysis: a planar rotation iterative procedure, *Submitted paper.*

DeSarbo, W.S., Jedidi, K., Cool, K., Schendel, D., (1990), Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups, *Marketing Letters*, **2**, 129-146.

De Soete, G. and Carroll, J.D., (1994), K-means clustering in a low-dimensional Euclidean space, *In: Diday, E., et al. (Eds.), New Approaches in Classification and Data Analysis. Springer, Heidelberg*, 212-219.

Dhillon, I.S, Marcotte, E.M., Roshan, U., (2003), Diametrical clustering for identifying anti-correlated gene clusters, *Bioinformatics*, **19**(13), 1612-1619.

Greenacre, M.J., Blasius, J., (2006), *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London.

Guo, Q., Wu, W., Massart, D.L., Boucon, C., de Jong, S., (2002), Feature selection in principal component analysis of analytical data, *Chemometrics and Intelligent Laboratory Systems*, **61**, 123-132.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P., (2000), 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology*, **1**(2), 1-21.

Jolliffe, I.T., (1972), Discarding variables in a principal component analysis. I. Artificial data, *Journal of the Royal Statistical Society. Series C. Applied Statistics* , **21**, 160-173.

Jolliffe, I.T., (2002), *Principal Component Analysis*, Second Edition, Springer-Verlag, New York.

Kaufman, L. and Rousseeuw P.J., (1990), *Finding groups in data: an introduction to cluster analysis*, Wiley Series in probability and mathematical statistics, New York.

Krzanowski, W.J., (1987), Selection of variables to preserve multivariate data structure, using principal components, *Journal of the Royal Statistical Society. Series C. Applied Statistics*, **36**, 22-33.

Lerman, I.C., (1993), Likelihood linkage analysis (LLA) classification method: An example treated by hand, *Biochimie*, **75**,(5) 379-397.

McCabe, G.P., (1984), Principal variables, *Technometrics*, **26**(2), 137-144.

Mirkin, B., (2005), *Clustering for Data Mining. A Data Recovery Approach.*, Chapman & Hall, CRC Press, London, Boca Raton, FL.

Qannari, E.M., Vigneau, E., Courcoux PH., (1998), Une nouvelle distance entre variables. Application en classification, *Revue de Statistique Appliquée*, **46**(2), 21-32.

Plasse, M., Niang, N., Saporta, G., Villeminot, A., Leblond, L., (2007), 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**, 596-613.

Soffritti, G., (1999), Hierarchical clustering of variables: a comparison among strategies of analysis, *Communications in Statistics - Simulation and Computation*, **28**(4), 977-999.

Stan, V. and Saporta, G., (2005), Conjoint use of variables clustering and PLS structural equations modelling. *In PLS05, 2005. 4th International Symposium on PLS and related methods, Barcelone, 7-9 septembre 2005.*

Vichi, M. and Kiers, H.A.L., (2001), Factorial k-means analysis for two way data, *Computational Statistics and Data Analysis*, **37**, 49-64.

Vichi, M. and Saporta, G., (2009), Clustering and disjoint principal component analysis, *Computational Statistics and Data Analysis*, **53**, 3194-3208.

Vigneau, E. and Qannari, E.M., (2003), Clustering of Variables Around Latent Components, *Communications in Statistics - Simulation and Computation*, **32**(4), 1131-1150.

Cahiers du GREThA
Working papers of GREThA

GREThA UMR CNRS 5113

Université Montesquieu Bordeaux IV
Avenue Léon Duguit
33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75
Fax : +33 (0)5.56.84.86.47

www.gretha.fr

Cahiers du GREThA (derniers numéros)

2009-09 : ROUILLON Sébastien, *Un nouveau mécanisme décentralisant les équilibres de Lindahl*

2009-10 : PETIT Emmanuel, *Does indignation lead to generosity? An experimental investigation*

2009-11 : KECHIDI Med, TALBOT Damien, *Réseau de proximité et gestion des interactions techniques et organisationnelles : les firmes pivots de l'aéronautique*

2009-12 : DOUAI Ali, MONTALBAN Matthieu, *Institutions and the environment: the case for a historical political economy*

2009-13 : NICET-CHENAF Dalila, ROUGIER Eric, *FDI and growth: A new look at a still puzzling issue*

2009-14 : NICET-CHENAF Dalila, ROUGIER Eric, *Human capital and structural change: how do they interact with each other in growth?*

2009-15 : DOYEN Luc, PERREAU Jean-Christophe, *Sustainable coalitions in the commons*

2009-16 : YILDIZOGLU Murat, *Approche évolutionniste de la dynamique économique*

2009-17 : JULLIEN Bernard, *Approche institutionnaliste de la dynamique industrielle*

2009-18 : BELIS-BERGOUIGNAN Marie-Claude, *Analyse évolutionniste de la dynamique sectorielle*

2009-19 : JULLIEN Bernard, *L'analyse sectorielle institutionnaliste : projet et méthodes*

2009-20 : CORIS Marie, FRIGANT Vincent, LAYAN Jean-Bernard, TALBOT Damien, *Les dynamiques spatiales des activités productives*

2009-21 : CARRINCAZEAUX Christophe, *Les dynamiques spatiales de l'innovation*

2009-22 : OLTRA Vanessa, SAINT JEAN Maïder, *Innovations environnementales et dynamique industrielle*

2009-23 : CORIS Marie, FRIGANT Vincent, LUNG Yannick, Changements *organisationnels et diversité des formes institutionnelles*

2009-24 : DUPUY Claude, MONTALBAN Matthieu, MOURA Sylvain, *Finance et dynamiques des industries*

2009-25 : CLEMENT Matthieu, *Amartya Sen et l'analyse socioéconomique des famines : portée, limites et prolongements de l'approche par les entitlements*

2010-01 : ZUMPE Martin, *Règles de politique monétaire, apprentissage et stabilité: une revue de la littérature récente*

2010-02 : SARACCO Jérome, CHAVENT Marie, KUENTZ Vanessa, *Clustering of categorical variables around latent variables*

La coordination scientifique des Cahiers du GREThA est assurée par Sylvie FERRARI et Vincent FRIGANT. La mise en page est assurée par Dominique REBOLLO.