Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 46 (2004) 269–294
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Agglomerative hierarchical clustering of continuous variables based on mutual information

Ivan Kojadinovic*

IREMIA, Université de La Réunion, 15 avenue René Cassin-BP 7151, 97715 Saint-Denis messag cedex 9, Ile de La Réunion, France

Received 1 April 2002

Abstract

In order to study interdependencies among continuous variables in the framework of a data analysis problem, an agglomerative hierarchical clustering of the set of variables is performed. The similarity measure used within the clustering algorithm is based on the notion of *mutual information*. Recent results on the estimation of this measure of stochastic dependence are presented and the behavior of the clustering algorithm is studied on several artificial problems, i.e., which "structure" is known.

Keywords: Agglomerative hierarchical clustering; Continuous variables; Mutual information; Shannon entropy; Redundancy; Adaptive kernel density estimation

1. Introduction

The study of functional dependencies among variables is one of the first fundamental steps when dealing with a data analysis problem. The most frequently adopted approach consists in estimating dependencies for pairs of variables using a sample version of a *measure of stochastic dependence*. In the continuous case, the most widely used measure is probably Pearson's linear correlation coefficient, which unfortunately only enables the detection of linear dependencies among variables.

A more ambitious approach consists in clustering the set of variables under study into groups according to the functional dependencies among variables. In order to do so, it is necessary to define a *similarity* or *dissimilarity* measure, generally on pairs of

* Tel.: +33-262938327; fax: +33-262938260.
E-mail address: ivan.kojadinovic@univ-reunion.fr (I. Kojadinovic).

0167-9473/$ - see front matter
doi:10.1016/S0167-9473(03)00153-1

variables. Such a measure can then be used within a clustering algorithm which aim is to identify *classes* of variables (i.e. non void subsets of variables) as *homogeneous* and as *separated* from each other as possible. In the framework of variable clustering, *homogeneous* means that variables within one cluster should be as functionally dependent as possible while *separated* means that the clusters of variables should be as "mutually stochastically independent" as possible. By highlighting possibly nonlinear interdependencies among variables, such a clustering algorithm can prove very useful as a preprocessing step, for example, in the framework of *subset variable selection* for discrimination or regression.

In this work, we address the problem of continuous variable clustering using a classical *hierarchical* approach and the notion of *mutual information* as similarity measure.

This paper is organized as follows. In Section 1, we recall the elements at the root of the classical hierarchical clustering model and the classical agglomerative hierarchical clustering algorithm. Then, we review the approaches to the hierarchical clustering of variables encountered in the literature for continuous as well as for discrete variables. After, the notion of *mutual information* is introduced through its links with the Kullback and Leibler (1951) divergence and the Shannon (1948) entropy and recent results on its estimation are presented. In Section 6, we present our approach to continuous variables clustering: the mutual information is used as similarity measure, indices of *homogeneity* and *separation* of partitions are proposed and an empirical study of the resulting clustering algorithm is performed on several artificial problems, i.e. which "structure" is known.

2. The classical agglomerative hierarchical clustering algorithm

The classical agglomerative hierarchical clustering algorithm is grounded on two elements: a *similarity* or *dissimilarity* measure between objects and an *aggregation criterion* or *linkage rule* between *classes* of objects (Chandon and Pinson, 1981; Saporta, 1990; Gordon, 1999).We have chosen to present the algorithm when it is based on a *similarity* measure. Let Θ be a finite set of objects. A similarity measure on Θ is generally defined as a symmetric mapping from $\Theta \times \Theta$ to $\mathbb{R}^+$ such that, for all $\{X, Y\} \subseteq \Theta$,

$$s(X,X) \geqslant s(X,Y),$$

i.e., the similarity between an object and itself is always greater than the similarity between the object and any other object of Θ.

Starting from the finest partition of Θ in which each object constitutes a class, the classical agglomerative hierarchical clustering algorithm successively forms new classes by merging at each step two most similar classes. Now, generally, the similarity measure is defined only on $\Theta \times \Theta$ and therefore does not enable the comparison of two classes of Θ. This is where the *aggregation criterion* intervenes since its role is to determine how the similarity between two classes should be computed, generally, from the similarities on pairs of objects. The best known aggregation criteria are probably the

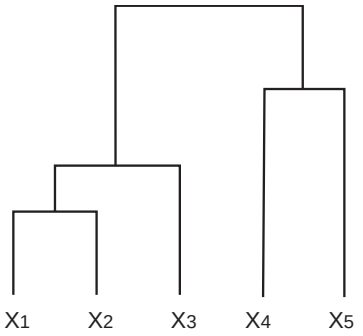

Fig. 1. Dendrogram representing a complete hierarchy.

single linkage, the *complete linkage* and the *average linkage* (Benzécri, 1976; Hansen and Jaumard, 1997; Gordon, 1999). Choosing the single linkage (resp. the complete linkage) as aggregation criterion is equivalent to taking as similarity between two classes the highest inter-class similarity (resp. the smallest inter-class similarity). The average linkage is simply defined as the arithmetic mean of the inter-class similarities.

Most of the agglomerative hierarchical clustering algorithms build a *complete hierarchy* of classes. A *hierarchy* $\mathscr{H}$ on Θ is a subset of the power set of Θ such that:

- $\mathscr{H}$ contains Θ,
- the minimal elements of $\mathscr{H}$ (with respect to inclusion) cover Θ,
- for all subsets $\mathbb{X}, \mathbb{Y} \in \mathscr{H}$, $\mathbb{X} \cap \mathbb{Y} \in \{\mathbb{X}, \mathbb{Y}, \emptyset\}$; in other terms, two classes of $\mathscr{H}$ are either disjoint, or contained in each other.

The different steps of the classical agglomerative hierarchical clustering algorithm based on a similarity measure are given below:

(1) Each object of Θ forms a class.
(2) The similarity associated with each pair of classes is calculated by means of the aggregation criterion.
(3) A pair of classes having the highest similarity degree, say $\{\mathbb{X}, \mathbb{Y}\}$, is identified, the composite class $\mathbb{X} \cup \mathbb{Y}$ is formed and the number of classes is decremented.
(4) Steps 2, 3 and 4 are repeated until the number of classes is equal to 1.

The hierarchy of classes built by the clustering algorithm can be represented by means of a *dendrogram*. For example, the dendrogram given in Fig. 1 represents the hierarchy

$$\{\{X_1\},\{X_2\},\{X_3\},\{X_4\},\{X_5\},\{X_1,X_2\},\{X_4,X_5\},$$
$$\{X_1,X_2,X_3\},\{X_1,X_2,X_3,X_4,X_5\}\}$$

on the set $\{X_1,X_2,X_3,X_4,X_5\}$. More generally when $|\Theta| = m$, the hierarchy of classes built by the above clustering algorithm contains $2m-1$ classes amongst which are the singletons of Θ. Such a hierarchy is said to be *complete* (Hansen and Jaumard, 1997).

The representation of a complete hierarchy by means of a dendrogram suggests to equivalently see the result of a hierarchical clustering as a sequence of m coarser and coarser partitions (Fig. 1). These partitions, obtained by "cutting" the dendrogram according to a horizontal line, are called *compatible* with the hierarchy (Saporta, 1990, Chapter 12). Now, if the clustering is performed as a preprocessing task, it is generally more appropriate to present the result of the clustering in the form of a partition than in the form of a hierarchy of classes. The most natural next step is then to choose one partition among the m partitions compatible with the complete hierarchy. In order to guide this choice, indices of *separation* and *homogeneity* of partitions based on the similarity measure can be used (Hansen and Jaumard, 1997). Recall that the more the objects within a class are similar, the more *homogeneous* it is and that the more dissimilar the classes of a partition, the more *separated* the partition.

The most natural and best known measure of *homogeneity* of a class is probably the *diameter*, which is defined as the smallest intra-class similarity. The notion of *split* of a class, defined as the highest degree of similarity between an object within the class and an object outside the class, is a natural measure of *separation* (Hansen and Jaumard, 1997). More formally, given a similarity measure s defined on $\Theta \times \Theta$, the diameter of a class $\mathbb{X} \subseteq \Theta$ is defined by

$$\mathscr{D}(\mathbb{X}) := \begin{cases} \min\limits_{\{X,Y\}\subseteq\mathbb{X}} s(X,Y) & \text{if } |\mathbb{X}| > 1, \\ s(X,X) & \text{if } |\mathbb{X}| = \{X\}. \end{cases} \tag{1}$$

The split of a class $\mathbb{X} \subsetneq \Theta$ is given by

$$\mathscr{S}(\mathbb{X}) := \max_{\substack{X\in\mathbb{X} \\ Y\in\aleph\setminus\mathbb{X}}} s(X,Y).$$

The degree of homogeneity of a partition $\{\mathbb{X}_1,\ldots,\mathbb{X}_r\}$ of Θ can then be measured by aggregating the diameters $\mathscr{D}(\mathbb{X}_1),\ldots,\mathscr{D}(\mathbb{X}_r)$, which leads for example to the notions of *average diameter* or *minimum diameter*. Similarly, for $r \geqslant 2$, the degree of separation of the partition can be measured by aggregating the splits $\mathscr{S}(\mathbb{X}_1),\ldots,\mathscr{S}(\mathbb{X}_r)$, which leads for example to the notions of *average split* or *maximum split* (Hansen and Jaumard, 1997).

It is easy to verify that the above measures, for instance the average diameter and the maximum split, increase with the number of clusters of the m partitions compatible with the constructed complete hierarchy. In other words, as the clustering algorithm proceeds, the degree of homogeneity of the coarsest partition compatible with the hierarchy under construction decreases and its degree of separation increases.

Now, should Θ be structured into r $(1 \leqslant r \leqslant m)$ very homogeneous groups of objects, it is easy to see that the $m - r + 1$ finest compatible partitions should have almost the same degree of homogeneity. Similarly, if there are r $(2 \leqslant r \leqslant m)$ very separated groups of objects, the $r - 1$ compatible partitions containing between 2 and r clusters should have almost the same degree of separation. Such situations can be identified simply by plotting the degrees of separation and homogeneity against the number of clusters and by considering their variations. In such a way, one can select the coarsest "most homogenuous" compatible partition or the finest "most separated"

partition. In certain cases, when the set of objects Θ happens to be structured into r ($2 \leqslant r \leqslant m$) very separated *and* very homogeneous clusters, the coarsest "most homogenuous" partition will coincide with the finest "most separated" partition, which could thus be considered as the optimal partition, in terms of separation *and* homogeneity, among the m compatible partitions. The practical use of measures of homogeneity and separation as means to chose a compatible partition is illustrated in Section 5.3.

3. Approaches to the hierarchical clustering of variables encountered in the literature

Most of the approaches to the clustering of variables encountered in the literature are of hierarchical type (Nicolau and Bacelar-Nicolau, 1998) and vary, on one hand, according to the nature of the variables under study, and, on the other hand, according to the choice of the similarity measure. In any case, the fundamental notion of *similarity* coincides with that of *functional dependence*. We shall present these approaches in a probabilistic setting.

In the case of discrete random variables taking a finite number of values, the similarity between variables is generally measured in terms of *distance from independence* using *divergence measures* (Kus, 1999), that is, as the *divergence* of the estimated joint distribution with respect to the tensor product of the estimated marginal distributions, which represents stochastic independence. The most frequently used divergence measures are Pearson's χ^2 and the Kullback and Leibler (1951) divergence. However, under their brut form, such indices are not always appropriate as similarity measures since their estimated values are not necessarily comparable from one pair of variables to another. This is the case when all the random variables under study do not take the same number of values (Saporta, 1990, Chapter 12). The use of normalized versions of these indices, such as the Cramer coefficient (Saporta, 1990, Chapter 7) or the normalized mutual information proposed by Joe (1989b) constitutes a first solution to this problem. Another approach, known as the *analysis of the link likelihood*, has been proposed by Lerman (1981) and mainly consists in replacing the value of the estimated similarity measure by the probability of finding a lower value under the hypothesis of stochastic independence, called *absence of link* in that context (Lerman et al., 1993; Nicolau and Bacelar-Nicolau, 1998). More formally, given a similarity measure s, the similarity between two random variables X and Y is replaced by $Pr(s(X,Y) \leqslant \hat{s}(X,Y))$ under the hypothesis of stochastic independence between X and Y, where $\hat{s}(X,Y)$ is the similarity between X and Y estimated from the available observations. Applying this approach thus clearly requires the knowledge of the probability distribution of the similarity measure under the hypothesis of absence of link.

In the case of continuous random variables, the existing approaches are mainly based on the linear correlation coefficient (Saporta, 1990, Chapter 12) or on Spearman's or Kendall's rank correlations (SAS, 1995). The main inconvenient of these approaches is that they do not enable the detection of all types of functional dependencies among variables. Indeed, it is clear that similarity indices based on the linear correlation coefficient can only highlight linear dependencies. Those based on Spearman's or Kendall's rank correlation enable only the detection of monotonic functional dependencies. To

our opinion, the use of more general measures of stochastic dependence has not been suggested thus far because of the difficulty of their estimation. Hence, to our knowledge, there are no approaches to the clustering of variables really satisfying in the continuous case.

4. Mutual information

The most natural approach to variable clustering is clearly the *analysis of the link likelihood* proposed by Lerman (1981). Unfortunately, in the continuous case, there are no similarity measure enabling the detection of all types of functional dependencies having an estimator which probability distribution is known under the hypothesis of stochastic independence. It seems therefore necessary to use a more "empirical" approach (Nicolau and Bacelar-Nicolau, 1998). We thus propose to adopt a classical hierarchical approach based on the measure of stochastic dependence known as *mutual information*. This choice is mainly due to the fact that the estimation of the mutual information is grounded on that of the Shannon (1948) entropy which has been recently studied in the continuous case (Joe, 1989a; Hall and Morton, 1993).

4.1. Mutual information, Shannon entropy and redundancy

Let us consider two random vectors $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ which joint density is absolutely continuous with respect to the Lebesgue measure. The *mutual information* between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ is generally defined as the *distance from independence* between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ measured by the Kullback and Leibler (1951) divergence (Cover and Thomas, 1991; Ullah, 1996).

Divergence measures can be seen as dissimilarity measures between probability densities (Kus, 1999). The Kullback and Leibler divergence, playing a central role in probability theory and statistics, is defined, for two Lebesgue densities p and q with same support, by

$$KL(p,q) := \int p \log\left(\frac{p}{q}\right) \tag{2}$$

with the convention that $0 \log \frac{0}{0} := 0$.

Let us denote by $p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}$ the joint density of $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ and by $p_{\vec{\mathbb{X}}}$ and $p_{\vec{\mathbb{Y}}}$ respectively the marginal densities of $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ respectively. The mutual information between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ is then defined by

$$I(\vec{\mathbb{X}};\vec{\mathbb{Y}}) := KL(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}, p_{\vec{\mathbb{X}}} \otimes p_{\vec{\mathbb{Y}}}), \tag{3}$$

where $p_{\vec{\mathbb{X}}} \otimes p_{\vec{\mathbb{Y}}}$ denotes the tensor product of $p_{\vec{\mathbb{X}}}$ and $p_{\vec{\mathbb{Y}}}$. From this definition, we see that the mutual information is symmetric and, from the Jensen inequality applied to the Kullback and Leibler divergence, we have that it is always non negative and zero if and only if $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ are stochastically independent.

The mutual information can also be interpreted as the *uncertainty reduction measure* (DeGroot, 1962) obtained from the Shannon (1948) entropy. The Shannon entropy of a Lebesgue density p, when it exists, is defined by

$$H(p) := -\int p \log p \tag{4}$$

with the convention that $0 \log 0 := 0$. It can be interpreted as a measure of the *uncertainty* or the *information* contained in the density p.

With respect to the Shannon entropy, the mutual information between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ can be easily rewritten as

$$I(\vec{\mathbb{X}}; \vec{\mathbb{Y}}) = H(p_{\vec{\mathbb{X}}}) - E_{p_{\vec{\mathbb{Y}}}}[H(p_{\vec{\mathbb{X}}|\vec{\mathbb{Y}}=y})], \tag{5}$$

where $p_{\vec{\mathbb{X}}|\vec{\mathbb{Y}}=y}(x) := (p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(x,y)/p_{\vec{\mathbb{Y}}}(y)$. By symmetry, we also have

$$I(\vec{\mathbb{X}}; \vec{\mathbb{Y}}) = H(p_{\vec{\mathbb{Y}}}) - E_{p_{\vec{\mathbb{X}}}}[H(p_{\vec{\mathbb{Y}}|\vec{\mathbb{X}}=x})].$$

Hence, the mutual information can be interpreted as the reduction in the uncertainty of $\vec{\mathbb{X}}$ (resp. $\vec{\mathbb{Y}}$) due to the knowledge of $\vec{\mathbb{Y}}$ (resp. $\vec{\mathbb{X}}$) (Ullah, 1996). Rewriting the expectation in Expression (5) as

$$E_{p_{\vec{\mathbb{Y}}}}[H(p_{\vec{\mathbb{X}}|\vec{\mathbb{Y}}=y})] = H(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}) - H(p_{\vec{\mathbb{Y}}}),$$

we also obtain

$$I(\vec{\mathbb{X}}; \vec{\mathbb{Y}}) = H(p_{\vec{\mathbb{X}}}) + H(p_{\vec{\mathbb{Y}}}) - H(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}). \tag{6}$$

As we can see from Expression (3), the mutual information between two continuous random vectors is defined only if their joint density exists. Hence, the mutual information between two functionally dependent random vectors is not defined. Following Joe (1989b), we shall consider that the higher the mutual information between two random vectors having a Lebesgue density, the "stronger" their functional dependence, and conversely.

The notion of mutual information can be straightforwardly generalized to more than two random vectors, in which case it is called *redundancy* (Wienholt and Sendhoff, 1996). The redundancy among $r \geqslant 2$ random vectors $\vec{\mathbb{X}}_1, \ldots, \vec{\mathbb{X}}_r$ having a joint Lebesgue density is defined by

$$R(\vec{\mathbb{X}}_1; \ldots; \vec{\mathbb{X}}_r) := KL(p_{(\vec{\mathbb{X}}_1,\ldots,\vec{\mathbb{X}}_r)}, p_{\vec{\mathbb{X}}_1} \otimes \cdots \otimes p_{\vec{\mathbb{X}}_r}),$$

which, in terms of the Shannon entropy, can be easily rewritten as

$$R(\vec{\mathbb{X}}_1; \ldots; \vec{\mathbb{X}}_r) = \sum_{i=1}^{r} H(p_{\vec{\mathbb{X}}_i}) - H(p_{(\vec{\mathbb{X}}_1,\ldots,\vec{\mathbb{X}}_r)}). \tag{7}$$

As previously, it is easy to verify that the redundancy is always positive and equal to zero if and only $\vec{\mathbb{X}}_1, \ldots, \vec{\mathbb{X}}_r$ are stochastically mutually independent. As for the mutual information, we shall consider that the higher the redundancy among the random vectors, the "stronger" their functional dependency (Joe, 1989b).

4.2. Estimation

In real situations, instead of the probability densities of the random vectors, we generally only have n independent realizations of these random vectors. These realizations can however be used to estimate the unknown densities and thus the quantities presented in the previous subsection.

From Expressions (6) and (7), we see that estimating the mutual information and the redundancy amounts to estimating Shannon entropies. The estimation of the Shannon entropy in the continuous case has been recently studied by Joe (1989a) and Hall and Morton (1993).

Consider a random vector $\vec{\mathbb{X}}$ having a Lebesgue density. A pointwise estimation of the entropy of its density can be obtained in two steps: first, by substituting the density of $\vec{\mathbb{X}}$ in the expression of the Shannon entropy by an estimate computed from available independent realizations; then, by computing the remaining integral by numerical quadrature (Silverman, 1986; Joe, 1989b).

The difficulties linked to numerical integration can however be avoided. Let $F_{\vec{\mathbb{X}}}$ be the cumulative distribution function of $\vec{\mathbb{X}}$ and let $\vec{\mathbb{X}}_1, \ldots, \vec{\mathbb{X}}_n$ be a random sample drawn from $p_{\vec{\mathbb{X}}}$. The Shannon entropy of $p_{\vec{\mathbb{X}}}$ can then be rewritten as

$$H(p_{\vec{\mathbb{X}}}) = -\int \log p_{\vec{\mathbb{X}}} \, \mathrm{d}F_{\vec{\mathbb{X}}}.$$

Substituting $F_{\vec{\mathbb{X}}}$ by the empirical cumulative distribution function and $p_{\vec{\mathbb{X}}}$ by an estimate, we obtain a natural estimator of the Shannon entropy given by

$$\hat{H}(p_{\vec{\mathbb{X}}}) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\vec{\mathbb{X}}}(\vec{\mathbb{X}}_i). \tag{8}$$

The above estimator has been studied by Joe (1989a) and Hall and Morton (1993) in the case where $p_{\vec{\mathbb{X}}}(\vec{\mathbb{X}}_i)$ is estimated by *kernel density estimation* (see Appendix A). In that context, Hall and Morton (1993) have shown that the estimator $\hat{H}(p_{\vec{\mathbb{X}}})$ is consistent if the dimension of $\vec{\mathbb{X}}$ is strictly inferior to 4 and if the density of $\vec{\mathbb{X}}$ satisfies certain regularity conditions. A synthesis on the estimation of the Shannon entropy in the continuous case can be found in Beirlant et al. (1997).

From a more practical perspective, as density estimation technique, we propose the use of the *adaptive kernel method* (see Appendix A) which can be considered as an improved version of the classical kernel density estimation technique (Silverman, 1986; Fukunaga, 1990; Scott, 1992). From the expression of the adaptive kernel density estimator given in (A.4), it is easy to verify that the complexity of the estimation of the Shannon entropy by means of estimator (8) is then at least $O(rn^2)$, where r is the dimension of the data.

From the estimator of the Shannon entropy given in (8), an estimator of the mutual information between two random vectors $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ having a Lebesgue density can be straightforwardly derived. Indeed, given a random sample $(\vec{\mathbb{X}}_1, \vec{\mathbb{Y}}_1), \ldots, (\vec{\mathbb{X}}_n, \vec{\mathbb{Y}}_n)$ drawn

from $p_{(\vec{\mathbb{X}},\vec{\mathbb{X}})}$, an estimator of $I(\vec{\mathbb{X}};\vec{\mathbb{Y}})$ is simply

$$\hat{I}(\vec{\mathbb{X}};\vec{\mathbb{X}}) := \hat{H}(p_{\vec{\mathbb{X}}}) + \hat{H}(p_{\vec{\mathbb{Y}}}) - \hat{H}(p_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})})$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\hat{p}_{(\vec{\mathbb{X}},\vec{\mathbb{Y}})}(\vec{\mathbb{X}}_i, \vec{\mathbb{Y}}_i)}{\hat{p}_{\vec{\mathbb{X}}}(\vec{\mathbb{X}}_i)\hat{p}_{\vec{\mathbb{Y}}}(\vec{\mathbb{Y}}_i))}\right). \tag{9}$$

The consistency of the above estimator follows immediately from that of the Shannon entropy when its consistency conditions are satisfied. An estimator of the redundancy can be derived in a similar way.

5. Agglomerative hierarchical clustering of continuous variables based on mutual information

We present our approach in a probabilistic setting. The set of m continuous random variables to be clustered will be denoted $\aleph := \{X_1, \ldots, X_m\}$. In the sequel, the subsets of $\aleph$ will be denoted by upper-case *black-board* letters, e.g. $\mathbb{X}$. Given a subset $\mathbb{X} \subseteq \aleph$ composed of r variables, $\vec{\mathbb{X}}$ will denote a r-dimensional random vector which coordinates are distinct elements from $\mathbb{X}$. We shall also assume that the density of the random vector $\vec{\aleph}$ is absolutely continuous with respect to the Lebesgue measure.

This section is organized as follows. The two first subsections are devoted to the definition of the similarity measure from the notion of mutual information and to the study of measures of homogeneity and separation of partitions. An empirical study of the resulting agglomerative hierarchical clustering algorithm is then performed on four artificial problems, i.e., which "structure" is known.

5.1. Definition of the similarity measure

As mentioned in the introduction, we have chosen to define the similarity between two disjoint classes $\mathbb{X}$ and $\mathbb{Y}$ of $\aleph$ by

$$s(\mathbb{X},\mathbb{Y}) := I(\vec{\mathbb{X}};\vec{\mathbb{Y}}).$$

The existence of the Lebesgue density of $\vec{\aleph}$ ensures that the similarity measure s is defined for all pairs of disjoint classes of $\aleph$. According to the results on the mutual information given in Section 4.1, we know that $s(\mathbb{X},\mathbb{Y})$ is always positive and that it is zero if and only if $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ are stochastically independent. Following Joe (1989b), we shall consider that the higher $s(\mathbb{X},\mathbb{Y})$, the "stronger" the functional dependence between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$, and conversely. When the classes under consideration are singletons $\{X\}$ and $\{Y\}$, we shall write $s(X,Y)$ instead of $s(\{X\},\{Y\})$ in order to avoid a heavy notation.

As it can be deduced from the definitions given in Section 4.1, the mutual information between two random vectors having a joint Lebesgue density can be arbitrarily high. In order to avoid the manipulation of arbitrarily high quantities, the similarity measure s can be normalized using a strictly increasing transformation from $\mathbb{R}^+$ to a bounded real interval, typically $[0,1]$. Joe (1989b) proposed to use the transformation

$x \mapsto \sqrt{1-\exp(-2x)}$, which gives as normalized similarity degree between two disjoint classes $\mathbb{X}$ and $\mathbb{Y}$

$$s_*(\mathbb{X},\mathbb{Y}) := \sqrt{1-\exp[-2s(\mathbb{X},\mathbb{Y})]}. \tag{10}$$

This normalized similarity is clearly between 0 and 1 and equals 0 if and only if $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ are stochastically independent. Furthermore, for all random variables X and Y having a joint normal distribution with correlation $\rho, |\rho| < 1$, we have $s_*(X,Y) = |\rho|$ since in this case, $I(X;Y) = 1/2\log(1-\rho^2)$, as shown in (Cover and Thomas, 1991). Given this normalization transformation, it seems natural to consider that the closer $s_*(\mathbb{X},\mathbb{Y})$ is to 1, the "stronger" the functional dependence between $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$, and conversely.

The measure of similarity s and its normalized version s_* enable not only the comparison of variables of $\aleph$ but also the comparison of disjoint classes of $\aleph$. It follows that the use of s or s_* in the framework of the agglomerative hierarchical clustering algorithm presented in Section 2 does not require the use of an aggregation criterion. Furthermore, s_* being obtained from s using a strictly increasing transformation, it is important to notice that the use of s or s_* necessarily leads to the same hierarchy of classes on $\aleph$ since s and s_* induce the same preorder on the set of pairs of disjoint classes of $\aleph$.

5.2. Measures of homogeneity and separation

The result of a hierarchical clustering is a complete hierarchy of classes. If the hierarchical clustering is performed as a preprocessing step, its result is generally easier to interpret and to use if it is given in the form of a partition than in the form of a hierarchy of classes. The most natural way of proceeding then consists in selecting one partition among the m partitions compatible with the obtained hierarchy on $\aleph$. As discussed in Section 2, in order to guide the choice of a partition, measures of separation and homogeneity, based for example of the notions of split and diameter, can be used.

As one can see from Expression (1), computing the diameter of a class may require the comparison of a variable with itself, which is not possible if s or s_* are used as similarity measures since random vectors of the form (X,X), $X \in \aleph$, do not have a Lebesgue density. Following Joe (1989b), who conjectured that the "stronger" the functional dependence between two random vectors, the higher their mutual information, we propose to naturally extend the definition of s by

$$s(X,X) := \infty, \quad X \in \aleph,$$

which, by continuity, leads to the following extension of s_*:

$$s_*(X,X) := 1, \quad X \in \aleph.$$

Now that s and s_* have been extended to enable the comparison of a variable with itself, the notions of split and diameter can be defined from either of the two similarity measures. However, if the aim is to define measures of separation and homogeneity of partitions, the use of s may be problematic since the computation of the degree of

homogeneity of a partition may require the aggregation of infinite diameters. Therefore, the use of a normalized similarity measure such as s_* seems more appropriate.

In the framework of variable clustering, it seems also natural to measure the degree of separation of a partition $\mathscr{P} = \{\mathbb{X}_1, \ldots, \mathbb{X}_r\}$, $2 \leqslant r \leqslant m$, using the notion *redundancy* defined by Expression (7). Indeed, as seen in Section 4.1, $R(\vec{\mathbb{X}}_1; \ldots; \vec{\mathbb{X}}_r)$ can be interpreted as measure of the *distance from mutual independence* of the random vectors $\vec{\mathbb{X}}_1, \ldots, \vec{\mathbb{X}}_r$. Thus, the lower $R(\vec{\mathbb{X}}_1; \ldots; \vec{\mathbb{X}}_r)$, the "less" functionally dependent are $\vec{\mathbb{X}}_1, \ldots, \vec{\mathbb{X}}_r$ and thus, the more separated is $\mathscr{P}$. When $R(\vec{\mathbb{X}}_1; \ldots; \vec{\mathbb{X}}_r) = 0$, we know, from Section 4.1, that the random vectors $\vec{\mathbb{X}}_1, \ldots, \vec{\mathbb{X}}_r$ are mutually stochastically independent. We shall then say that the partition $\{\mathbb{X}_1, \ldots, \mathbb{X}_r\}$ is *well separated.*

The redundancy satisfies the following property.

Proposition 1. *Consider a partition* $\{\mathbb{X}_1, \ldots, \mathbb{X}_r\}$, $2 \leqslant r \leqslant m$, *of* $\aleph$. *Then, we have*

$$R(\vec{\mathbb{X}}_1; \ldots, \vec{\mathbb{X}}_{r-2}; (\vec{\mathbb{X}}_{r-1}, \vec{\mathbb{X}}_r)) \leqslant R(\vec{\mathbb{X}}_1; \ldots; \vec{\mathbb{X}}_r)$$

with equality if and only if $\vec{\mathbb{X}}_{r-1}$ *and* $\vec{\mathbb{X}}_r$ *are stochastically independent.*

Proof. Using the positivity of the mutual information, we can write

$$\begin{aligned} R(\vec{\mathbb{X}}_1; \ldots, \vec{\mathbb{X}}_{r-2}; (\vec{\mathbb{X}}_{r-1}, \vec{\mathbb{X}}_r)) &= \sum_{i=1}^{r-2} H(p_{\vec{\mathbb{X}}_i}) + H(p_{(\vec{\mathbb{X}}_{r-1}, \vec{\mathbb{X}}_r)}) - H(p_{\vec{\aleph}}) \\ &\leqslant \sum_{i=1}^{r-2} H(p_{\vec{\mathbb{X}}_i}) + H(p_{\vec{\mathbb{X}}_{r-1}}) + H(p_{\vec{\mathbb{X}}_r}) - H(p_{\vec{\aleph}}) \\ &\leqslant R(\vec{\mathbb{X}}_1; \ldots; \vec{\mathbb{X}}_r). \end{aligned}$$

Furthermore, from the properties of the mutual information, we have that $H(p_{(\vec{\mathbb{X}}_{r-1}, \vec{\mathbb{X}}_r)}) = H(p_{\vec{\mathbb{X}}_{r-1}}) + H(p_{\vec{\mathbb{X}}_r})$ if and only if $\vec{\mathbb{X}}_{r-1}$ and $\vec{\mathbb{X}}_r$ are stochastically independent. □

Hence, given a partition $\mathscr{P}$, a partition $\mathscr{P}'$, obtained by merging two classes $\mathbb{X}$ and $\mathbb{Y}$ of $\mathscr{P}$, is more separated than $\mathscr{P}$ if and only if the random vectors $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ are not stochastically independent. If $\vec{\mathbb{X}}$ and $\vec{\mathbb{Y}}$ are stochastically independent, the partitions $\mathscr{P}$ and $\mathscr{P}'$ are equivalent in terms of separation.

The redundancy can thus be used exactly as more traditional measures of separation based on the split (cf. Section 2). Indeed, should there be $r (2 \leqslant r \leqslant m)$ "almost" mutually stochastically independent groups of variables in $\aleph$, the $r-1$ compatible partitions of the constructed hierarchy containing between 2 and r clusters should have almost the same redundancy. Such a situation can be identified simply by plotting the redundancy against the number of clusters of the compatible partition, which, by observing the redundancy variations, enables to select the finest "less redundant" partition, that is the "most separated" partition.

5.3. Empirical study of the clustering algorithm

In real variable clustering problems, instead of the Lebesgue density of $\vec{\aleph}$, we generally only have n independent realizations of this random vector. However, as seen in Section 4.2, these realizations can be used to estimate the similarities between disjoint classes of variables by means of the estimator given in (9). In the sequel, the similarity measure estimated from the available data will be denoted by $\hat{s}$ and its normalized version by $\hat{s}_*$. A theoretical study of $\hat{s}$ (or equivalently $\hat{s}_*$) in the framework of the hierarchical clustering algorithm described in Section 2 appearing very difficult even if strong regularity constraints are imposed on the underlying densities (Joe, 1989b; Hall and Morton 1993), as a next step, it seemed important to us to study its properties in that context at least empirically. In order to do so, we have generated four clustering problems. For each of the problems, we first study the influence of the sample size on the preorder induced by $\hat{s}$ on the set of pairs of disjoint classes, that is, its influence on the hierarchy of classes built by the clustering algorithm (based on the theoretical results given in Section 4.2 concerning the estimation of the mutual information, we shall consider, in the sequel, that, for large sample sizes, $\hat{s}$ leads to the same hierarchy of classes as s).

As a second study, for each problem, we verify that the sample versions of the measures of homogeneity and separation presented in Section 5.2 enable to detect (at least approximately) the "structure" of the set of variables to be clustered. By "structure", we refer to a *partition* or a *covering* that appropriately describes the interdependencies among variables.

5.3.1. A detailed study of the clustering alogrithm on a first simple problem

In the framework of the first problem, we study the set of continuous random variables $\aleph_1 := \{X_1, \ldots, X_9\}$. The variable X_1 has a uniform density on $[-1, 1]$, the variables X_4 and X_7 are standard normal, these three variables being stochastically mutually independent. The other variables are obtained from X_1, X_4 and X_7 using the following transformations:

$$X_2 := \tanh(X_1) + X_1^2 + \varepsilon_2,$$

$$X_3 := 2\sin(|X_1|) + \varepsilon_3,$$

$$X_5 := \sin(X_4) + \tanh(X_4) + \varepsilon_5,$$

$$X_6 := X_4^2 + \varepsilon_6,$$

$$X_8 := |X_7| + \varepsilon_8,$$

$$X_9 := \sin(|X_7|) + \varepsilon_9.$$

The random variables ε_i in the expressions above are white noises of variance 0.01 and have been introduced to make this artificial problem more realistic. Despite the disturbance caused by the ε_i, here, it seems natural to consider that the partition

$$\{\{X_1, X_2, X_3\}, \{X_4, X_5, X_6\}, \{X_7, X_8, X_9\}\} \tag{11}$$

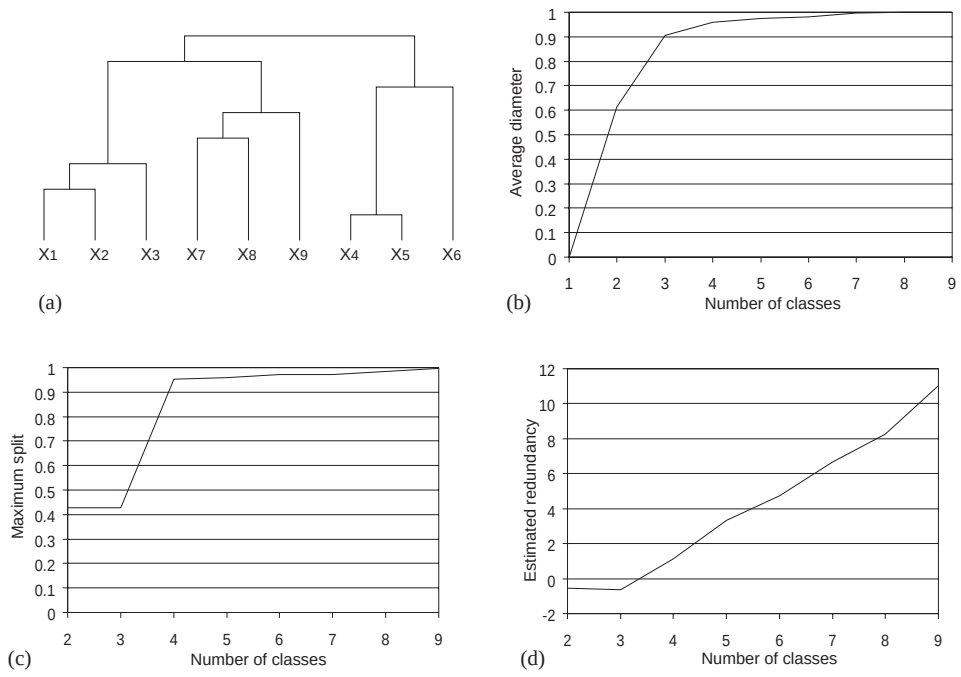

Fig. 2. Results of the hierarchical clustering of $\aleph_1$ from $n = 1600$ realizations.

is the optimal partition of $\aleph_1$, in terms of separation *and* homogeneity. This *well separated* partition can be considered as an appropriate representation of the "structure" of $\aleph_1$.

We have first randomly generated $n = 1600$ realizations of each of the three variables X_1, X_4 and X_7 using the `ran2` routine given in (Press et al., 1992, Chapter 7) to generate uniform deviates and the Box–Muller method to generate normal deviates from uniform ones. The corresponding realizations of the six other variables have then been computed using the transformations given above. The hierarchy built by the clustering algorithm from the $n = 1600$ observations is represented by the dendrogram given in Fig. 2(a). As we can see, the 3-class partition compatible with the obtained hierarchy is Partition (11), which is the optimal partition for this problem in terms of separation *and* homogeneity. The conjectured consistency of the estimator of the mutual information given in (9) and the rather large sample size from which the estimations have been performed suggest to consider that the same hierarchy of classes would have been obtained using s instead of $\hat{s}$, should the density of $\vec{\aleph}_1$ had have been known.

Let us now verify that the sample versions of the measures of homogeneity and separation previously proposed enable to designate Partition (11) as the optimal partition in terms of separation *and* homogeneity among the nine partitions compatible with the obtained hierarchy.

Table 1
Upper triangle: estimated similarity matrix ($\hat{s}$, $n = 1600$). Lower triangle: estimated normalized similarity matrix ($\hat{s}_{**}$, $n = 1600$)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1		1.61	1.23	0.00	−0.04	0.07	0.00	−0.01	−0.01
X_2	0.98		0.72	−0.01	−0.04	−0.09	−0.01	−0.03	−0.02
X_3	0.96	0.90		0.00	−0.03	−0.06	0.00	−0.01	0.01
X_4	0.39	0.38	0.41		2.75	1.11	0.01	0.01	0.00
X_5	0.32	0.31	0.35	0.99		0.64	−0.02	−0.03	−0.02
X_6	0.21	0.00	0.24	0.95	0.88		−0.05	−0.06	−0.05
X_7	0.41	0.39	0.41	0.43	0.36	0.28		1.37	1.03
X_8	0.38	0.25	0.39	0.42	0.35	0.23	0.97		1.19
X_9	0.39	0.36	0.42	0.41	0.36	0.26	0.95	0.96	

We first proceed to the estimation of the average diameter and the maximum split of the nine compatible partitions. To do so, it is necessary to estimate the similarity s on all pairs of variables, and then to normalize the obtained values by means of Transformation (10). The similarity matrix of the variables of $\aleph_1$, estimated from the $n = 1600$ generated realizations, is given in the upper triangle of Table 1. As we can notice, certain coefficients of the matrix, corresponding to pairs of stochastically independent variables, are slightly negative, whereas they should be zero. This observation suggests that the estimator of the mutual information given in (9) has a slightly negative bias under the hypothesis of stochastic independence. It follows that the estimated measure of similarity $\hat{s}$ is not necessarily positive and therefore that its normalized version $\hat{s}_*$ is not always defined. In the case where certain coefficients $\hat{s}(X,Y), \{X,Y\} \subseteq \aleph$, are strictly negative, we propose to use a slightly modified normalization of the similarity measure s defined by

$$\hat{s}_{**}(X,Y) := \sqrt{1 - 2\exp\left[-2\left(\hat{s}(X,Y) - \min_{\{X,Y\}\in\aleph} \hat{s}(X,Y)\right)\right]}.$$

Note however that any other strictly increasing transformation from $\mathbb{R}$ to a real bounded interval could have been used.

The estimated similarity matrix normalized using the above transformation is given in the lower triangle of Table 1. We then define the notions of diameter and split from the coefficients $\hat{s}_{**}(X;Y), \{X,Y\} \subseteq \aleph$, i.e., the diameter of a class $\mathbb{X} \subseteq \aleph$ is computed as

$$\mathscr{D}(\mathbb{X}) = \begin{cases} \min\limits_{\{X,Y\}\subseteq\mathbb{X}} \hat{s}_{**}(X,Y) & \text{if } |\mathbb{X}| > 1, \\ 1, & \text{if } |\mathbb{X}| = \{X\}, \end{cases}$$

and its split is computed, provided $\mathbb{X} \neq \aleph$, as

$$\mathscr{S}(\mathbb{X}) = \max_{\substack{X\in\mathbb{X} \\ Y\in\aleph\setminus\mathbb{X}}} \hat{s}_{**}(X,Y).$$

Table 2
Upper triangle: estimated similarity matrix ($\hat{s}_{**}$, $n = 1600$). Lower triangle: estimated correlation matrix in absolute value ($n = 1600$)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1		0.98	0.96	0.39	0.32	0.22	0.41	0.38	0.39
X_2	0.83		0.89	0.38	0.32	0.00	0.39	0.35	0.36
X_3	0.02	0.47		0.41	0.35	0.24	0.41	0.39	0.42
X_4	0.01	0.02	0.05		0.99	0.95	0.43	0.42	0.41
X_5	0.02	0.00	0.02	0.95		0.87	0.36	0.35	0.35
X_6	0.04	0.03	0.02	0.02	0.02		0.27	0.23	0.26
X_7	0.01	0.02	0.04	0.02	0.02	0.01		0.97	0.95
X_8	0.00	0.04	0.07	0.00	0.01	0.00	0.04		0.96
X_9	0.02	0.05	0.06	0.02	0.01	0.02	0.01	0.72	

In order to compare the nine partitions of $\aleph_1$ compatible with the hierarchy represented in Fig. 2(a) in terms of homogeneity and separation, we have computed the average diameter and the maximum split of these partitions. These measures of homogeneity and separation are represented with respect to the number of classes of the partitions in Figs. 2(b) and (c). As we can see, the average diameter of the partitions decreases strongly as soon as the number of classes becomes strictly smaller than 3 and the maximum split reaches its optimal value as soon as the number of classes becomes lower than 3. It seems therefore natural to designate the 3-class partition given in (11) as the optimal partition compatible with the obtained hierarchy in terms of homogeneity *and* separation.

As a next step, we have also estimated the redundancy of each partition compatible with the obtained hierarchy. This measure of separation is represented with respect to the number of classes of the compatible partitions in Fig. 2(d). As we can notice, the estimated redundancy decreases with the number of classes but does not remain positive. This last observation suggests that the estimator of the redundancy based on (8) has also a slightly negative bias under the hypothesis of mutual stochastic independence. However, the estimated redundancies of the 2 and 3-class partitions are approximately equal, which seems to designate the 3-class partition (11) as the finest most separated partition. Thus, in this statistical context, it seems natural to solely take into account the variation of the estimated redundancy. Indeed, only the knowledge of the probability distribution of the estimated redundancy under the hypothesis of stochastic mutual independence could have enabled us to draw conclusions with respect to the redundancy values.

To illustrate the necessity of using similarity measures able to detect all types of functional dependencies, in Table 2, we compare the normalized similarity matrix and the linear correlation matrix (in absolute value) of the variables of $\aleph_1$, both estimated from the same $n = 1600$ realizations. While the normalized estimated similarity matrix clearly highlights very strong functional dependencies among the variables X_1, X_2, X_3, the variables X_4, X_5, X_6 and the variables X_7, X_8, X_9, the coefficients of the correlation matrix are globally very small, thereby indicating an almost complete absence of linear

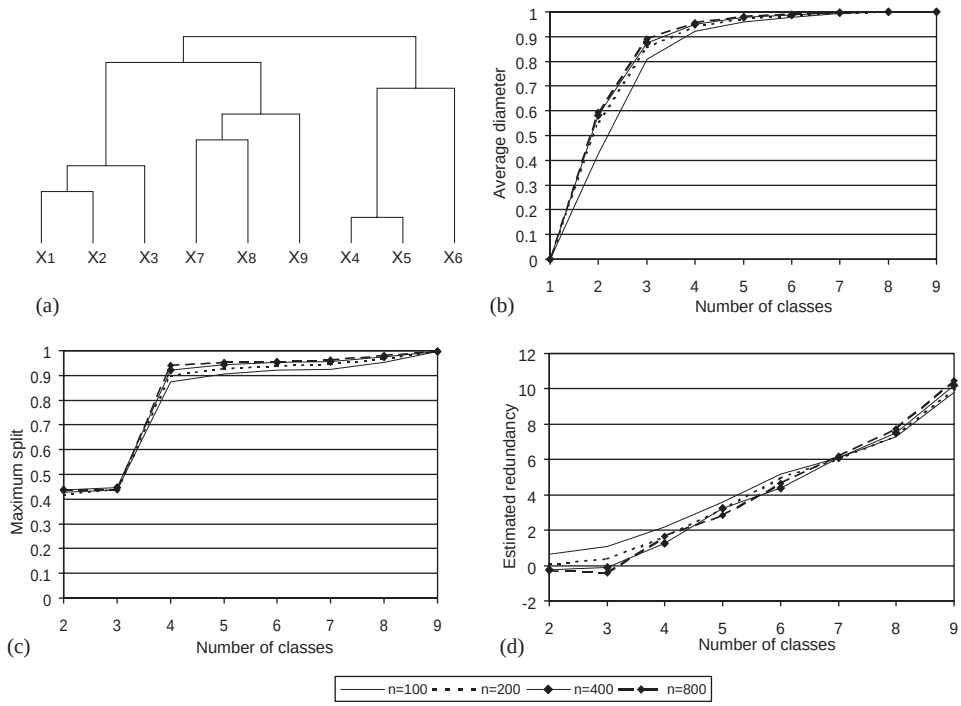

Fig. 3. Results of the hierarchical clustering of $\aleph_1$ from $n = 100$, 200, 400 and 800 realizations.

dependence. It therefore clearly appears that taking into account only linear dependencies among variables does not enable to solve this type of problem in an appropriate way.

The results presented so far have been obtained from $n = 1600$ observations. In order to empirically study the influence of the sample size on the results of the clustering, we have proceeded to similar simulations from $n = 100$, 200, 400 and 800 realizations. In any of these cases, the clustering algorithm has produced the same hierarchy of classes as previously, which is represented by the dendrogram of Fig. 3(a). In Figs. 3(b)–(d) the average diameter, the maximum split and the estimated redundancy are plotted against the number of classes of the compatible partitions. We can thus notice that, for $n = 100$, 200, 400 and 800, it seems also reasonable to designate as optimal compatible partition, in terms of homogeneity *and* separation, the 3-class partition given in (11). Note however that the estimated redundancy enables to designate Partition (11) as the finest most separated partition only when $n \geqslant 400$.

5.3.2. *A problem with a more complex "structure"*

The previous problem is characterized by the fact that the optimal compatible partition in terms of homogeneity and size coincide with the optimal compatible partition in terms of separation and size. This being rarely the case in real situations, we have

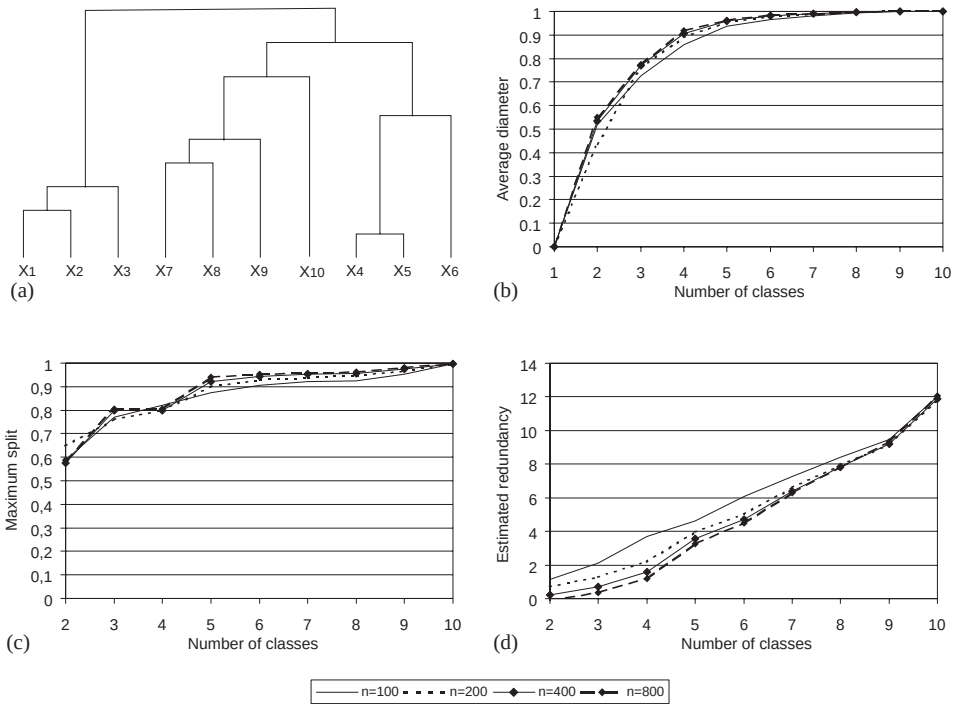

Fig. 4. Results of the hierarchical clustering of $\aleph_1'$ from $n = 100$, 200, 400 and 800 realizations.

decided to increase the difficulty by adding a tenth variable to $\aleph_1$, denoted X_{10}, such that its "link" with the three classes $\{X_1, X_2, X_3\}$, $\{X_4, X_5, X_6\}$ and $\{X_7, X_8, X_9\}$ of the optimal partition of $\aleph_1$ in terms of separation and homogeneity be non negligible and approximately identical. This new variable is defined by

$$X_{10} := X_1 + X_4 + X_7 + \varepsilon_{10},$$

where ε_{10} is a white noise of variance 0.01. The new set of variables thus obtained is denoted $\aleph_1' := \aleph_1 \cup \{X_{10}\}$. A natural way to describe the interdependencies among variables in $\aleph_1'$, i.e. its "structure", is to use the covering

$$\{\{X_1, X_2, X_3, X_{10}\}, \{X_4, X_5, X_6, X_{10}\}, \{X_7, X_8, X_9, X_{10}\}\}. \tag{12}$$

In order to study the clustering of $\aleph_1'$, we have run the algorithm for $n = 100$, 200, 400 and 800 realizations. In the four cases, the same hierarchy of classes, represented by the dendrogram of Fig. 4(a), has been obtained. As previously, the average diameter, the maximum split and the estimated redundancy of the partitions compatible with the hierarchy of Fig. 4(a) are plotted against the number of classes of these partitions in Figs. 4(b)–(d), respectively. From Fig. 4(b), we can notice that the homogeneity of the compatible partitions starts strongly decreasing as soon as the number of classes becomes strictly lower than 4. Thus, among the 10 partitions compatible with the

obtained hierarchy, the one that seems optimal in terms of homogeneity and size is the 4-class partition

$$\{\{X_1,X_2,X_3\},\{X_4,X_5,X_6\},\{X_7,X_8,X_9\},\{X_{10}\}\}.$$

The maximum split and the estimated redundancy seem to designate partition

$$\{\{X_1,X_2,X_3\},\{X_4,X_5,X_6,X_7,X_8,X_9,X_{10}\}\}$$

as the finest most separated compatible partition.

Thus, as expected, the partition designated as optimal in terms of separation and size does not coincide with the one designated as optimal in terms of homogeneity and size. Note however that the two designated partitions can be considered as "approximations" of the covering given in (12).

5.3.3. A problem enabling the study of $\hat{s}$ as aggregation criterion

In the considered context, the quantity $\hat{s}$ plays at the same time the role of a similarity measure and that of an aggregation criterion since it enables not only to measure the similarity between variables but also the similarity between disjoint classes of variables. In order to verify that $\hat{s}$, considered as an aggregation criterion, does not lead to the formation of excessively large classes (as its is the case for the *single linkage* for instance), it seems important to empirically study the influence of a larger number of functionally dependent variables on the results of the clustering algorithm. In the framework of this artificial problem, we thus study the set of random variables $\aleph_2 = \{X'_1,\ldots,X'_9\}$. The variable X'_1 is standard normal, the variable X'_8 is normal with expectation 0 and variance 100, the variable X'_9 has a uniform density on $[-1,1]$, these three variables being mutually stochastically independent. The variables $X'_2,\ldots,X'_7$ are obtained from X'_1 by means of the following transformations:

$$X'_2 := \sin(X'_1) + \varepsilon'_2,$$

$$X'_3 := (X'_1)^2 + \varepsilon'_3,$$

$$X'_4 := |X'_1| + \tanh(X'_1) + \varepsilon'_4,$$

$$X'_5 := \tanh(X'_1) + \varepsilon'_5,$$

$$X'_6 := \cos(X'_1) + \varepsilon'_6,$$

$$X'_7 := 2X'_1 - 8 + \varepsilon'_8.$$

As previously, the variables ε'_i in the expressions above are white noises of variance 0.01.

It seems reasonable to consider that the optimal partition of $\aleph_2$ in terms of separation *and* homogeneity is the well separated partition

$$\{\{X'_1,X'_2,X'_3,X'_4,X'_5,X'_6,X'_7\},\{X'_8\},\{X'_9\}\}. \tag{13}$$

As previously, we have run the clustering algorithm on $\aleph_2$ for $n = 100$, 200, 400 and 800 realizations. For the four sample sizes, the same hierarchy of classes, represented

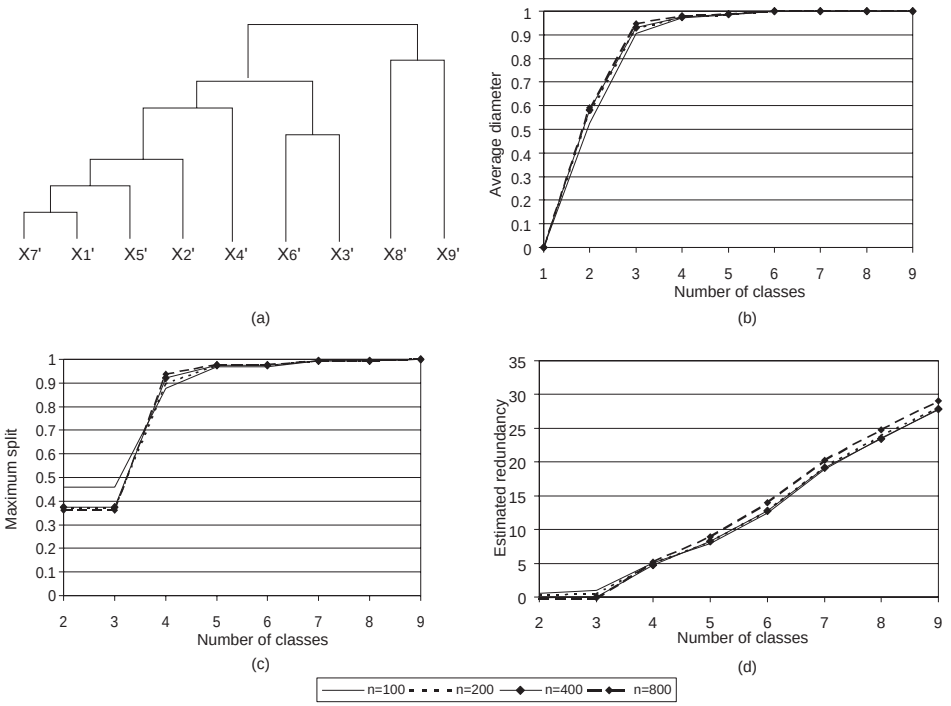

Fig. 5. Results of the hierarchical clustering of $\aleph_2$ from $n = 100$, 200, 400 and 800 realizations.

by the dendrogram of Fig. 5(a), has been obtained. The average diameter, the maximum split and the estimated redundancy of the partitions compatible with the hierarchy of Fig. 5(a) are plotted against the number of classes of these partitions in Figs. 5(b)–(d), respectively. From Fig. 5(b), we can notice that the homogeneity of the partitions compatible with the obtained hierarchy starts decreasing rapidly as soon as the number of classes becomes strictly inferior to 3. Similarly, from Figs. 5(c) and (d), we see that the indices of separation, i.e., the maximum split and the estimated redundancy, reach their optimal values as soon as the number of classes becomes lower than 3. Thus, it seems natural to designate as optimal compatible partition, in terms of separation *and* homogeneity, the 3-class partition given in (13). Therefore, the presence of a larger number of functionally dependent variables seems not to affect the results of the clustering algorithm.

5.3.4. A problem with a larger number of variables

Finally, it seems important to empirically study the behavior of the clustering algorithm in the case of a larger number of variables to be clustered. In order to so, we have chosen to run the algorithm first on the set $\aleph_1 \cup \aleph_2$ and then on the set $\aleph_1' \cup \aleph_2$.

The partition of $\aleph_1 \cup \aleph_2$ that can be considered optimal, in terms of separation *and* homogeneity, is clearly

$$\{\{X_1, X_2, X_3\}, \{X_4, X_5, X_6\}, \{X_7, X_8, X_9\}, \{X_1', X_2', X_3', X_4', X_5', X_6', X_7'\}, \{X_8'\}, \{X_9'\}\}. \tag{14}$$

This well separated partition can be considered as an appropriate description of the "structure" of $\aleph_1 \cup \aleph_2$.

In order to study the influence of the sample size on the results of the clustering, we have run the algorithm on $\aleph_1 \cup \aleph_2$ for $n = 100$, 200, 400, 800, 1600 and 3200 observations. The obtained results show that Partition (14) is only compatible with the hierarchies built from 1600 and 3200 observations.

We have performed a similar study of $\aleph_1' \cup \aleph_2$ which "structure" could be described by the following covering

$$\{\{X_1, X_2, X_3, X_{10}\}, \{X_4, X_5, X_6, X_{10}\}, \{X_7, X_8, X_9, X_{10}\}, \{X_1', X_2', X_3', X_4', X_5', X_6', X_7'\}, \{X_8'\}, \{X_9'\}\}.$$

The partition

$$\{\{X_1, X_2, X_3\}, \{X_4, X_5, X_6\}, \{X_7, X_8, X_9\}, \{X_{10}\}, \{X_1', X_2', X_3', X_4', X_5', X_6', X_7'\}, \{X_8'\}, \{X_9'\}\}, \tag{15}$$

which can be considered as the optimal partition of $\aleph_1' \cup \aleph_2$ in terms of homogeneity and size, is compatible only with the hierarchies built from 1600 and 3200 observations.

Thus, while for a small number m of variables, the clustering algorithm enables to identify (at least approximately) the "structure" of the set of variables even for a low number n of observations, for this last problem, it seems necessary that the number of available realizations be much higher. To our opinion, this behavior follows from the difficulty of the estimation of the similarity between two disjoint classes of high cardinal when the number of realizations is low, that is, from the difficulty of density estimation in high dimensional spaces. Indeed, for a fixed number n of data points, a consequence of the so-called *curse of dimensionality* (Bellman, 1961; Silverman, 1986) is that the similarity between two disjoint classes of low cardinal can generally be estimated with a better accuracy than the similarity between two disjoint classes of high cardinal. A way of solving this problem is to use, when n is low and m is large, an *aggregation criterion* to measure the similarity degree between disjoint classes.

5.4. Classical aggregation criteria

The notion of *aggregation criterion*, as seen in Section 2, is classically one of the two fundamental elements at the root of hierarchical clustering. The role of this concept is to determine how the similarity between two disjoint classes should be calculated, generally, from the similarities on pairs of variables.

A natural way of proceeding consists in using classical aggregation criteria such as the *single linkage*, the *complete linkage* or the *average linkage* (Chandon and Pinson, 1981). In the considered context, using the single linkage (resp. the complete linkage) is equivalent to defining the similarity between two disjoint classes $\mathbb{X}$ and $\mathbb{Y}$ of $\aleph$ by

$$\max_{\substack{X\in\mathbb{X}\\ Y\in\mathbb{Y}}} [\hat{s}(X,Y)] \left(\text{resp. } \min_{\substack{X\in\mathbb{X}\\ Y\in\mathbb{Y}}} [\hat{s}(X,Y)] \right).$$

The average linkage between two disjoint classes $\mathbb{X}$ and $\mathbb{Y}$ is defined by

$$\frac{1}{|\mathbb{X}||\mathbb{Y}|} \sum_{X\in\mathbb{X}} \sum_{Y\in\mathbb{Y}} \hat{s}(X,Y).$$

Among the three aggregation criteria presented above, the single linkage is probably the one which mathematical properties have been studied the most (Hansen and Jaumard, 1997). It is characterized by its tendency to "rapidly" link similar objects. On the contrary, the complete linkage tend to "delay" the aggregation of similar objects. The average link can be considered as a compromise between these two strategies. It is usually preferred in applications because of its "robustness" (Chandon and Pinson, 1981, Chapter 5).

5.5. Empirical study of the algorithm based on the average linkage

The choice of the aggregation criterion is fundamental since it determines the hierarchy of classes built by the clustering algorithm. Because of its higher "robustness", we have chosen to use the average linkage.

As a first empirical study, we have run the clustering algorithm based on the average linkage on $\aleph_1$, $\aleph_1'$ and $\aleph_2$ for $n=100$, 200, 400 and 800 observations. The same results as for $\hat{s}$ have been obtained with respect to optimal compatible partitions in terms of homogeneity and separation.

As a second study, it seemed important to us to verify that the use of an aggregation criterion enables to solve the problems encountered during the clustering of $\aleph_1\cup\aleph_2$ and $\aleph_1'\cup\aleph_2$ for small samples. In order to do so, we have first run the clustering algorithm based on the average linkage on $\aleph_1\cup\aleph_2$ for $n=100$, 200, 400 and 800 observations. In the four cases, the same hierarchy of classes has been obtained. The minimum diameter and the maximum split of the partitions compatible with the obtained hierarchy on $\aleph_1\cup\aleph_2$ are plotted against the number of classes of these partitions in Figs. 6(a) and (b), respectively. As we can notice, the compatible partition that seems optimal in terms of separation *and* homogeneity is composed of 6 classes. A closer look on the obtained hierarchy shows that this 6-class partition is the optimal partition of $\aleph_1\cup\aleph_2$, in terms of homogeneity *and* separation, that is the partition given in (14). We have then performed a similar study of $\aleph_1'\cup\aleph_2$. The results of the clustering based on the average linkage are given in Fig. 6(c) and (d). As we can see, the optimal compatible partition in terms of homogeneity and size is composed of seven classes. A closer look on the obtained hierarchy shows that this 7-class partition is the optimal partition of $\aleph_1'\cup\aleph_2$, in terms of homogeneity and size, that is the partition given in (15).

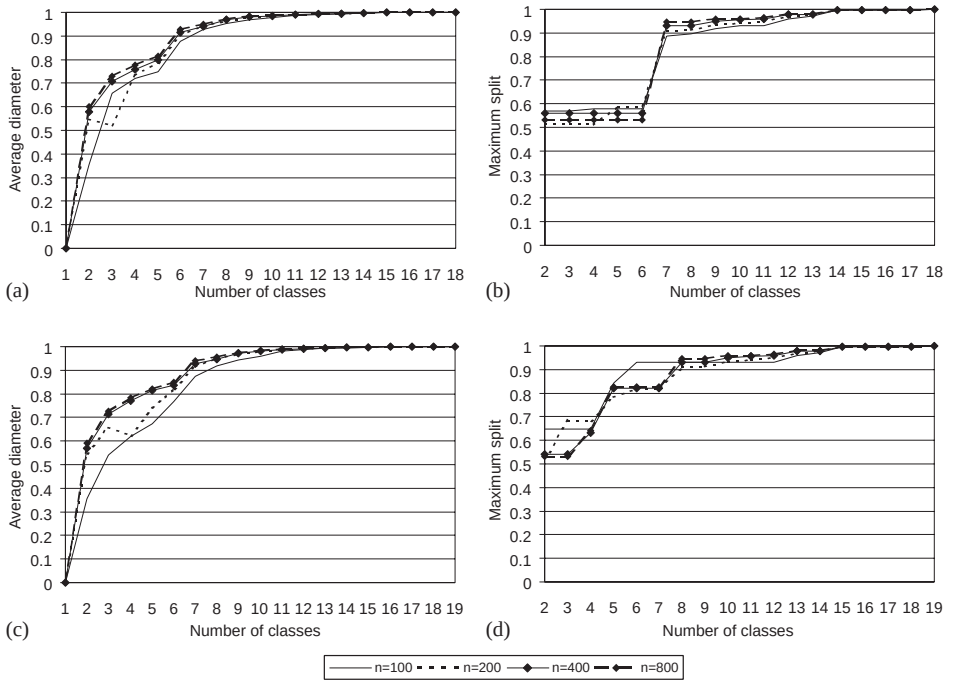

Fig. 6. Results of the hierarchical clustering (a), (b) of $\aleph_1 \cup \aleph_2$ and (c), (d) of $\aleph_1' \cup \aleph_2$.

Notice that simulations on even larger sets of variables based on the average linkage have been carried out and turned out satisfying in terms of "structure" identification.

It therefore seems that the use of an aggregation criterion enables to solve the problems encountered when the number of variables to be clustered is large and the sample size is low. We shall therefore distinguish between two situations: an "optimal" situation characterized by a sample which size is "exponentially larger" then the number of variables (Silverman, 1986) and a more realistic situation that consists of a large number of variables and little data. The simulations performed thus far give an idea, although vague, of the values of m and n corresponding to these two situations.

6. Conclusion

In this paper, as a mean to study interdependencies among continuous variables in the framework of a data analysis problem, we have proposed to perform an agglomerative hierarchical clustering of the variables under consideration. The clustering algorithm, based on the notion of mutual information, has been empirically studied on several artificial problems and its behavior can be considered as satisfying. It is important to note that the generated problems, because of the strongly nonlinear interdependencies involved, cannot be solved using more classical similarity measures based for example on Pearson's linear correlation or on Spearman's or Kendall's rank correlation.

We conclude this paper by mentioning the use of such a clustering algorithm as a preprocessing step in the framework of a *subset variable selection procedure* for regression or discrimination. The need for such a procedure arises when the number of variables that could be used to *explain* a variable of interest is very high. The aim of the procedure then consists in selecting a subset of *candidate explanatory variables* that can be used to explain the variable of interest in an "optimal" way. Clustering the set of candidate explanatory variables according to their interdependencies could then result in structural information that could be used to reduce the search space of the selection procedure. For instance, assume that one is looking for a reasonable subset of p candidate explanatory variables: the set of candidate explanatory variables could first be clustered and the result given in the form of a partition of size p; then, it would be natural to consider as candidate subsets, subsets of cardinal p which contain one and only one variable from each class of the obtained partition.

Acknowledgements

The author is very grateful to Jean Diatta for fruitful discussions and to the two anonymous referees whose suggestions greatly improved the quality of this paper.

Appendix A. Adaptive kernel density estimation

The most used and best studied probability density estimation technique is probably the *kernel method*. In this appendix, we restrict ourselves to an algorithmical presentation of this technique. A theoretical presentation can be found in (Silverman, 1986; Fukunaga, 1990; Scott, 1992).

Consider n realizations $x_1,\ldots,x_n$ of a r-dimensional random vector $\vec{\mathbb{X}}$ supposed absolutely continuous with respect to the Lebesgue measure. The probability density of $\vec{\mathbb{X}}$, estimated by the kernel method from the sample $x_1,\ldots,x_n$, is then given by

$$\hat{p}_{\vec{\mathbb{X}}}(x) := \frac{1}{nh^r}\sum_{i=1}^{n}\kappa\left(\frac{x-x_i}{h}\right), \quad x\in\mathbb{R}^r, \tag{A.1}$$

where $\kappa : \mathbb{R}^r \to \mathbb{R}$, is a symmetric, non negative function, called the *kernel function*, and satisfying

$$\int_{\mathbb{R}^r}\kappa(x)\,\mathrm{d}x = 1,$$

and where h is a real parameter, called the *kernel bandwidth*, which role is to determine the "region of influence" of the kernel function. This parameter is usually optimized by an automatic procedure which minimizes an approximation of the integrated mean squared error (Silverman, 1986, Chapters 3 and 4).

A frequent choice for the kernel function is the standard normal density, i.e.,

$$\kappa(x) := (2\pi)^{-r/2}\exp(-\tfrac{1}{2}xx^t), \quad x\in\mathbb{R}^r,$$

which is symmetric, of infinite support and has good regularity properties.

By considering Expression (A.1), we see that the idea behind the kernel method is simply to associate with each observation a "density element" of the shape of the kernel κ and in summing up all these elements. Histogram density estimation can thus be seen as a particular version of the kernel method in which the "density element" is a "small rectangle" in the class of the observation.

From a practical perspective, the shape of the kernel function has little influence on the estimated density (Scott, 1992, Chapter 6). On the contrary, the choice of the bandwidth h is crucial. If h is too small, the estimated density is in general too spiky and spurious structural details may appear. On the contrary, if h is too high, the estimated density is usually too smooth and many structural details may be hidden.

In the classical kernel method, the same bandwidth is used for all the available observations. In order to avoid too large differences of spread in the data, it has been suggested, before performing the density estimation, to linearly transform the data in order to obtain a sample with zero mean and identity variance-covariance matrix (Silverman, 1986; Hwang et al., 1994). The first step of this preprocessing phase consists in determining the eigenvalues and the eigenvectors of estimated variance-covariance matrix S of the random vector $\vec{\mathbb{X}}$. The matrix S can then be written as

$$S = UDU^t,$$

where U is an orthonormal matrix and D is the diagonal matrix of eigenvalues. The new standardized sample, denoted $z_1,\ldots,z_n$, is then obtained by performing, for all $i \in \{1,\ldots,n\}$, the following linear transformation:

$$z_i := S^{-1/2}(x_i - \bar{x}), \tag{A.2}$$

where $S^{-1/2} = UD^{-1/2}U^t$ and where $\bar{x}$ is the sample mean of $x_1,\ldots,x_n$. This classical transformation is known as *data sphering* (or *whitening*) or as the *Mahalanobis transformation*. The new sample $z_1,\ldots,z_n$ has thus zero mean and identity variance-covariance matrix. By integrating the transformation of the sample in the expression of the estimated density, we obtain

$$\hat{p}_{\vec{\mathbb{X}}}(x) = \frac{det(S)^{-1/2}}{nh^r} \sum_{i=1}^{n} \kappa\left(\frac{S^{-1/2}(x - x_i)}{h}\right), \quad x \in \mathbb{R}^r.$$

The last step before being able to perform the estimation of the density consists in determining a value for the bandwidth h. In the case of a normal sample with zero mean and identity variance–covariance matrix and when the normal kernel is used, an optimal choice for the bandwidth h is given by (Silverman, 1986, Chapter 4)

$$h = \left[\frac{4}{(2r+1)n}\right]^{1/(r+4)}. \tag{A.3}$$

However, in most real situations, the available observations are not gaussian, which is why the value of the bandwidth given by (A.3) generally leads to an oversmoothed density estimate. In order to obtain a density estimate which is closer to the real density (in terms of integrated mean squared error), the bandwidth h can be tuned locally to each observation. This improved version of the kernel method is known as *adaptive kernel density estimation* (Silverman, 1986; Scott, 1992; Hwang et al., 1994).

The tuning of the bandwidth h within the adaptive kernel method is performed according to the smoothness of a "pilot" density estimate. The different steps of the estimation are given below:

(1) *Data sphering*: The available observations $x_1,\ldots,x_n$ are linearly transformed using (A.2) in order to obtain a standardized sample $z_1,\ldots,z_n$.
(2) *Estimation of the pilot density*: The pilot density is estimated at each z_i by using as value for the bandwidth that given by (A.3). The obtained density values are denoted $\tilde{p}(z_i)$, for all $i\in\{1,\ldots,n\}$.
(3) *Local tuning of the bandwidth*: For all $i\in\{1,\ldots,n\}$, a tuning parameter local to the observation z_i, denoted λ_i, is computed by

$$\lambda_i := (\tilde{p}(z_i)/g)^{-\gamma},$$

where g is the geometric mean of the $\tilde{p}(z_i)$, i.e.,

$$\log g = \frac{1}{n}\sum_{i=1}^{n}\log\tilde{p}(z_i),$$

and where γ is a real parameter between 0 and 1 used to determine the influence of the pilot density on the final density estimate: the closer γ is to 1, the more sensitive is the density estimate with respect to the pilot density. A frequent choice for γ is $1/2$ (Silverman, 1986, Chapter 4).

The density estimate obtained using the adaptive kernel method is thus given by

$$\hat{p}_{\mathbb{X}}(x) = \frac{det(S)^{-1/2}}{n}\sum_{i=1}^{n}\frac{1}{h^r\lambda_i^r}\,\kappa\left(\frac{z-z_i}{h\lambda_i}\right) \quad \text{for all } x\in\mathbb{R}^r. \tag{A.4}$$

Empirical studies have shown that the adaptive kernel method can lead to a density estimate 45% more accurate in dimension 2 and 30% more accurate in dimension 3 in terms of integrated mean squared error then the classical kernel method (Scott, 1992, Chapter 6).

References

Beirlant, J., Dudewicz, E., Györfi, L., van der Meulen, E., 1997. Nonparametric entropy estimation: an overview. Int. J. Math. Stat. Sci. 6, 17–39.
Bellman, R., 1961. Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton.
Benzécri, J.-P., 1976. L'analyse de donnes: la taxonomie. Dunod, Paris.
Chandon, J., Pinson, S., 1981. Analyse Typologique: Théories et Applications. Masson, Paris.
Cover, T., Thomas, J., 1991. Elements of Information Theory. Wiley, New York.
DeGroot, M.H., 1962. Uncertainty, information and sequential experiments. Ann. Math. Statist. 33, 404–419.
Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, San Diego.
Gordon, A.D., 1999. Classification, 2nd Edition. Chapman & Hall, London.
Hall, P., Morton, S., 1993. On the estimation of entropy. Ann. Inst. Statist. Math. 45, 69–88.
Hansen, P., Jaumard, B., 1997. Cluster analysis and mathematical programming. Math. Program. 79, 191–215.
Hwang, J., Lay, S., Lippman, A., 1994. Nonparametric multivariate density estimation: a comparative study. IEEE Trans. Signal Process. 5 (10), 2795–2810.

Joe, H., 1989a. Estimation of entropy and other functionals of a multivariate density. Ann. Inst. Statist. Math. 41, 683–697.

Joe, H., 1989b. Relative entropy measures of multivariate dependence. J. Am. Statist. Assoc. 84, 157–164.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22, 79–86.

Kus, V., 1999. Divergences and generalized score functions in statistical inference. Ph.D. Thesis, Czech Technical University, Prague, Czech Republic.

Lerman, I., 1981. Classification et Analyse Ordinale de Données. Dunod, Paris.

Lerman, I., Peter, P., Leredde, H., 1993. Principes et calculs de la methode implante dans le programme chavl (classification hirarchique par analyse de la vraisemblance des liens). Modulad 33–101.

Nicolau, F., Bacelar-Nicolau, H., 1998. Some trends in the classification of variables. In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H., Baba, Y. (Eds.), Data science, Classification and Related Methods. Springer, Berlin, pp. 89–98.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The Art of Scientific Computing, 2nd Edition. Cambridge University Press, Urbana.

Saporta, G., 1990. Probabilités, Analyse de Données et Statistique. Editions Technip, Paris.

SAS, 1995. SAS User Manual, 3rd Edition. Version 6.

Scott, D., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley Intersciences, New York.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Systems Tech. J 27, 379,623.

Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, New York.

Ullah, A., 1996. Entropy, divergence and distance measures with econometric applications. J. Statist. Planning Inference 49, 137–162.

Wienholt, W., Sendhoff, B., 1996. How to determine the redundancy of noisy chaotic time series. Int. J. Bifurcation Chaos 6 (1), 101–117.