Chapter 2
Two Methods of Non-hierarchical Clustering

2.1 Preamble

As mentioned in the Preface, the development provided in this book is dominated by the potential of applying *ascendant agglomerative hierarchical clustering* to all types of data. Nonetheless, the specific methodology devoted to non-hierarchical clustering is also very important. In these conditions, we shall describe in this chapter two mutually very different methods of non-hierarchical clustering. The first one, called "Central Partition" method, is due to S. Régnier [35–37]. The second method called "méthode des Nuées Dynamiques" or "Dynamic cluster method" is due to E. Diday and collaborators [10, 11, 13, 16]. This approach corresponds to a vast generalization of the K-means method, initiated by Forgy [17] and Jancey [19].

A very important idea, common to both approaches, consists of summarizing a finite set E situated in a space endowed with a distance function, by a ***centre***. For the central partition method, E is a set of partitions on a finite set of objects and for the dynamic cluster method, E is a cluster of elements of a finite set of objects, or, more generally, a finite set of structures on an object set $\mathcal{O}$. In the case presented here (see Sect. 2.3) the set $\mathcal{O}$ to be clustered is represented by points in the geometrical space $\mathbb{R}^p$, where $\mathbb{R}$ denotes the reals and p a given positive integer.

The presentation of the central partition method may partially require the formalism established in Chap. 3. Some recent theoretical and methodological developments of this method will be indicated, particularly, in Sect. 2.2.4. Otherwise, the K-means algorithm is certainly the basic technique which has received the greatest development in non-hierarchical clustering [3, 18]. We shall try to comment on some of their elements.

I.C. Lerman, *Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-6793-8_2

2.2 Central Partition Method

2.2.1 Data Structure and Clustering Criterion

2.2.1.1 The Data

A set $\mathcal{A} = \{a^m | 1 \leq m \leq p\}$ of nominal categorical attributes (see Sect. 3.3.1 of Chap. 3) is assumed established in order to describe a set $\mathcal{O}$ of objects. Let $\mathcal{C}_m$ be the value set of the attribute a^m, $1 \leq m \leq p$. Designating by $\boldsymbol{a}$ the vector attribute $(a^1, a^2, \ldots, a^m, \ldots, a^p)$, the description can be expressed in terms of the mapping

$$\begin{aligned} \boldsymbol{a} &\longrightarrow \mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_m \times \cdots \times \mathcal{C}_p \\ x &\mapsto \boldsymbol{a}(x) = \left(a^1(x), a^2(x), ..., a^m(x), ..., a^p(x)\right) \end{aligned} \tag{2.2.1}$$

where $x \in \mathcal{O}$ and where $a^m(x)$ is the category of a^m possessed by x, $1 \leq m \leq p$. The sets $\mathcal{C}_m$ are finite, without any specific structure. Thereby, each of the a^m attributes induces a partition P^m on the object set $\mathcal{O}$, $1 \leq m \leq p$. If K_m is the class number of P^m, the latter can be written as

$$P^m = \left\{O_1^m, O_2^m, \ldots, O_k^m, \ldots, O_{K_m}^m\right\} \tag{2.2.2}$$

where O_k^m ($O_k^m \subset \mathcal{O}$) is composed of the set of objects having the kth value of the attribute a^m, $1 \leq k \leq K_m$, $1 \leq m \leq p$. Thus, the data are defined by the partition vector

$$(P^1, P^2, \ldots, P^m, \ldots, P^p) \tag{2.2.3}$$

on $\mathcal{O}$.

As indicated above, this data structure will be expressed again in Sect. 3.3.1 of Chap. 3. Equally, it will intervene in Sect. 7.2.3 of Chap. 7 devoted to the construction of a similarity index between objects described by nominal categorical attributes. According to (2.2.3), a partition P that we have to compare with the family of partitions P^m, $1 \leq m \leq p$, will be represented as a partition vector having p components equal to the same partition P:

$$(P, P, \ldots, P, \ldots, P) \tag{2.2.4}$$

2.2.1.2 The Notion of Central Partition

Let $\mathbb{P} = (P^m | 1 \leq m \leq p)$ be a sequence of partitions on a set $\mathcal{O}$ of objects. The notion of *central partition* of $\mathbb{P}$ assumes the definition of a distance δ on the set $\mathcal{P}$ of all partitions on $\mathcal{O}$. In these conditions, a central partition P of $\mathbb{P}$ is defined as

$$Arg\left\{\min_{P\in\mathcal{P}}\left[\frac{1}{p}\sum_{1\le m\le p}\delta(P^m,P)\right]\right\} \tag{2.2.5}$$

Clearly, in the expression of this criterion, the multiplicative factor $\frac{1}{p}$ can be omitted. However, according to [35, 36], we will keep it. In fact, it refers to the *mean* notion.

If P and Q are two partitions of $\mathcal{O}$, S. Régnier considers for $\delta(P, Q)$ the cardinal of the symmetrical difference of the graphs in $\mathcal{O}\times\mathcal{O}$ of the equivalence relations associated with P and Q. Without risk of ambiguity, the latter relations can also be designated as P and Q.

Notice that the graph of the equivalence relation associated with P^m defined above (see (2.2.2)) is given as

$$gr(P^m)=\left\{(x,y)|(x,y)\in\sum_{1\le m\le K_m}O_k^m\times O_k^m\right\} \tag{2.2.6}$$

Thereby, clearly, the representation of a partition of $\mathcal{O}$, by the graph of the associated equivalence relation, is situated in the Cartesian product $\mathcal{O}\times\mathcal{O}$.

In these conditions, the criterion to be minimized becomes

$$\delta[(P^1,P^2,\ldots,P^m,\ldots,P^p),(P,P,\ldots,P,\ldots,P)]=\frac{1}{p}\sum_{1\le m\le p}card\big(gr(P^m)\Delta gr(P)\big) \tag{2.2.7}$$

where Δ designates the symmetrical difference.

Without loss of generality, we may consider the representation of the graph of an equivalence relation at the level of the set $F=P_2(\mathcal{O})$ of unordered distinct object pairs of $\mathcal{O}$, that is, by a subset of F (see Sect. 1.5 of Chap. 1 and Sect. 4.3.2 of Chap. 4). This representation at the F level is more reduced ($card(F)=n(n-1)/2$) than that in $\mathcal{O}\times\mathcal{O}$. The diagonal of the latter set, that is, the subset of pairs of the form (x, x) ($x\in\mathcal{O}$) does not take part in the new representation. Clearly, we can deduce one of both representations from the other.

Now, by considering the representation at the F level, we can observe that the distance index

$$\delta(P,Q)=card\big(gr(P)\Delta gr(Q)\big) \tag{2.2.8}$$

where P and Q are two partitions of $\mathcal{O}$, can be reduced to the Rand similarity index expressed in Sect. 4.3.2 of Chap. 4. If $Rand(P, Q)$ designates this index, we can show easily that

$$\delta(P,Q)=n\big(1-Rand(P,Q)\big) \tag{2.2.9}$$

where $n=card(\mathcal{O})$.

2.2.1.3 Analysis of a Classification (Clustering) Criterion

As considered by S. Régnier [35, 36], let us return now to the representation in $\mathcal{O} \times \mathcal{O}$ of a partition P of an object set $\mathcal{O}$. To $gr(P)$ (where P denotes here the equivalence relation associated with a partition P) corresponds the indicator function ϖ defined as follows:

$$\begin{aligned}\big(\forall (x, y) \in \mathcal{O} \times \mathcal{O}\big)\ \varpi(x, y) = 1 \text{ if } & x \text{ and } y \text{ are in a same class of } P \\ \big(\forall (x, y) \in \mathcal{O} \times \mathcal{O}\big)\ \varpi(x, y) = 0 \text{ if } & x \text{ and } y \text{ are in different classes of } P\end{aligned} \tag{2.2.10}$$

By coding the set $\mathcal{O}$ with the set $\mathbb{I} = \{1, 2, \ldots, i, \ldots, n\}$ of the first n integers, the partition P can be represented by the point S of the cube $\{0, 1\}^{n \times n}$, whose coordinate sequence is

$$\big(\varpi(x, y) | (x, y) \in \mathcal{O} \times \mathcal{O}\big) \tag{2.2.11}$$

We have

$$\begin{aligned}&\big(\forall (x, y) \in \mathcal{O} \times \mathcal{O}\big),\ \varpi(x, y) = \varpi(y, x) \\ &\big(\forall (x, y) \in \mathcal{O} \times \mathcal{O}\big),\ \varpi(x, y) = 1 \text{ and } \varpi(y, z) = 1 \Rightarrow \varpi(x, z) = 1\end{aligned} \tag{2.2.12}$$

Conversely, each point of the cube $\{0, 1\}^{n \times n}$, for which (2.2.11) and (2.2.12) are satisfied, is the representation of a partition of $\mathcal{O}$.

We will designate by ϖ^m and ϖ the indicator functions associated with the partitions P^m and P in (2.2.3) and (2.2.4), $1 \leq m \leq p$.

Notice that the logic cube $\{0, 1\}^{n \times n}$ represents all the binary relations on $\mathcal{O}$. By immersing it in the geometrical space $\mathbb{R}^{n \times n}$, endowed with the usual Euclidean metric, the square distance between two binary relations R and R', represented by their respective indicator functions ρ and ρ', is

$$d^2(\rho, \rho') = \sum_{(x,y) \in \mathcal{O} \times \mathcal{O}} [\rho(x, y) - \rho'(x, y)]^2 \tag{2.2.13}$$

We have

$$card\big(gr(R) \Delta gr(R')\big) = d^2(\rho, \rho')\big) \tag{2.2.14}$$

In the case where R and R' are partitions of $\mathcal{O}$, as in Eq. (2.2.7) for P^m and P, this index gives the number of ordered pairs (x, y) of $\mathcal{O} \times \mathcal{O}$ which are joined in the same class for one of both partitions and separated into two classes for the other.

In these conditions, the criterion defined in (2.2.7) becomes

$$\delta\big((P^1, P^2, \ldots, P^m, \ldots, P^p), (P, P, \ldots, P, \ldots, P)\big) = \frac{1}{p} \sum_{1 \leq m \leq p} d^2(S^m, S) \tag{2.2.15}$$

where S^m (resp., S) is the point of $\mathbb{R}^{n \times n}$ representing the equivalence relation P^m (resp., P), $1 \leq m \leq p$. The coordinates of S^m (resp., S) are given by the indicator function ϖ^m (resp., ϖ). More precisely, $\big(s_{ij}^m | (i,j) \in \mathbb{I} \times \mathbb{I}\big)$ (resp., $\big(s_{ij} | (i,j) \in \mathbb{I} \times \mathbb{I}\big)$) will represent the coordinate vector of S^m (resp., S). s_{ij}^m (resp., $s(i,j)$) is defined by $\varpi^m(x,y)$ (resp., $\varpi(x,y)$), where (x,y) is coded by $(i,j) \in \mathbb{I} \times \mathbb{I}$. Then, we will denote below indifferently, by s_{ij}^m or s_{xy}^m (resp., s_{ij} or s_{xy}), $1 \leq m \leq p$.

Equation (2.2.15) can be written as

$$\frac{1}{p} \sum_{1 \leq m \leq p} (s_{xy}^m - s_{xy})^2 = \frac{1}{p} \sum_{1 \leq m \leq p} \|S^m - S\|^2 \tag{2.2.16}$$

where $\|\bullet\|$ designates the Euclidean norm in $\mathbb{R}^{n \times n}$.

The interest in immersing the cube of the binary relations on $\mathcal{O}$ in $\mathbb{R}^{n \times n}$ is due to the barycenter properties in the latter space. Let us designate by G the centre of gravity of the family of points $\{S^m | 1 \leq m \leq p\}$ of $\mathbb{R}^{n \times n}$, equally weighted as

$$G = \frac{1}{p} \sum_{1 \leq m \leq p} S^m \tag{2.2.17}$$

We have

$$\sum_{1 \leq m \leq p} \frac{1}{p} \|S^m - S\|^2 = \|S - G\|^2 + \sum_{1 \leq m \leq p} \frac{1}{p} \|S^m - G\|^2 \tag{2.2.18}$$

This is a classical formula—easily demonstrable—concerning a cloud of points in an Euclidean space (see Sect. 10.3.3) where the moment of order 2 with respect to a given vertex S is decomposed relatively to the cloud inertia (moment of order 2 with respect to the gravity centre of the cloud).

In these conditions, a ***central*** partition can be defined as realizing the minimum of $\|S - G\|^2$, with respect to S. Considering the general property (2.2.10) in which s_{xy} can be substituted for ϖ_{xy}, the function $\|S - G\|^2$ of S can be simplified. In fact all of the points of the cube $\{0,1\}^{n \times n}$ are at equal distance of the point H, whose coordinates are all equal to $\frac{1}{2}$. We have

$$\big(\forall (x,y) \text{ in } \mathcal{O} \times \mathcal{O}\big) \left(s_{xy} - \frac{1}{2}\right)^2 = \frac{1}{4}$$

and then,

$$\|S - H\|^2 = \sum_{(x,y) \in \mathcal{O} \times \mathcal{O}} \left(s_{xy} - \frac{1}{2}\right)^2 = \frac{n^2}{4} \tag{2.2.19}$$

We can say, in geometrical terms, that the cube of the binary relations on $\mathcal{O}$ is inscribed in the ball centred in H of radius $n^2/4$. It follows that

$$\|S-G\|^2 = \|(S-H)+(H-G)\|^2 = \|S-H\|^2 + \|H-G\|^2 + 2\cdot\langle S-H, H-G\rangle \tag{2.2.20}$$

where $\langle \bullet, \bullet \rangle$ denotes the symmetrical bilinear form associated with the Euclidean norm. In these conditions, (2.2.20) can be written as

$$\|S-G\|^2 = \frac{n^2}{4} + \|G-H\|^2 - 2\cdot\langle S-H, G-H\rangle \tag{2.2.21}$$

Therefore, minimizing $\|S-G\|^2$ is equivalent to maximizing the scalar product $\langle S-H, G-H\rangle$ which can be expanded as follows:

$$\sum_{(x,y)\in\mathcal{O}\times\mathcal{O}} \left(s_{xy}-\frac{1}{2}\right)\left(g_{xy}-\frac{1}{2}\right) \tag{2.2.22}$$

where the g_{xy}, $(x, y) \in \mathcal{O}\times\mathcal{O}$, are the coordinates of the point G, that is,

$$g_{xy} = \frac{1}{p}\sum_{1\le m\le p} s_{xy}^m \tag{2.2.23}$$

By putting $t_{xy} = g_{xy} - \frac{1}{2}$, the central partition is defined as maximizing the linear form

$$L(S) = \sum_{(x,y)\in\mathcal{O}\times\mathcal{O}} t_{xy}.s_{xy} \tag{2.2.24}$$

For a partition $P = \{O_k | 1 \le k \le K\}$ of $\mathcal{O}$, whose classes are the O_k, $1 \le k \le K$, (2.2.24) can be written as

$$L(S) = \sum_{1\le k\le K}\ \sum_{(x,y)\in O_k\times O_k} t_{xy} \tag{2.2.25}$$

where S is the point of $\{0, 1\}^{n\times n}$ representing the equivalence relation P, associated with the partition P. The kth term of the sum over $\{1, 2, \ldots, k, \ldots, K\}$ is a measure of cohesion or density of the class O_k. $L(S)$ is then the sum of densities of the different classes of P.

It is easy to see that $g_{xy} < 1/2$ for every (x, y) in $\mathcal{O}\times\mathcal{O}$, is a necessary and sufficient condition that the finest partition (each class is a singleton class) is optimal for $L(S)$. On the other hand, the coarsest partition into a single class can be optimal without having necessarily $g_{xy} > 1/2$ for all (x, y) in $\mathcal{O}\times\mathcal{O}$. Finally, we can notice that when g_{xy} is constant and equal to $1/2$, every partition of $\mathcal{O}$ is a central partition.

Example of Computing $L(S)$

Relative to the incidence data Table 4.1 given in Sect. 4.2.1.2 of Chap. 4, let us compare the partitions

$$P = \{C_1, C_2\} = \big\{\{o_a, o_d, o_f\}, \{o_b, o_c, o_e\}\big\}$$

and

$$Q = \{D_1, D_2\} = \big\{\{o_a, o_d, o_e, o_f\}, \{o_b, o_c\}\big\}$$

of the set $\mathcal{O} = \{o_a, o_b, o_c, o_d, o_e, o_f\}$, where we have $D_1 = C_1 + \{o_e\}$ and $D_2 = C_2 - \{o_e\}$.

This incidence data table illustrates the description of 6 objects by 18 Boolean attributes. In order to carry out the comparison mentioned, each Boolean attribute is interpreted here as a nominal categorical attribute with two values, 0 and 1. Table 2.1 gives the respective values of t_{xy} on $\mathcal{O} \times \mathcal{O}$. The rows of Table 2.1 are arranged in such a way that both partitions P and Q can clearly be distinguished.

For P, the criterion value is

$$\frac{41}{18} + \frac{37}{18} = \frac{13}{3} = \frac{39}{9}$$

and for Q,

$$\frac{46}{18} + \frac{24}{18} = \frac{35}{9}$$

Therefore, P is better than Q, according to the criterion concerned.

Admissible Metrics in $\mathbb{R}^{n\times n}$

The reason for which seeking for a central partition leads to maximizing a linear form (see (2.2.24)) is due to the fact that in $\mathbb{R}^{n\times n}$, all the points of the cube $\{0, 1\}^{n\times n}$

Table 2.1 Cohesion table

$\mathcal{O} \setminus \mathcal{O}$	o_a	o_d	o_f	o_e	o_b	o_c
o_a	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{9}$	$-\frac{1}{6}$	$-\frac{1}{9}$	$-\frac{1}{9}$
o_d	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{9}$	$\frac{1}{18}$	0	$-\frac{1}{9}$
o_f	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{2}$	0	$-\frac{1}{6}$	$-\frac{7}{18}$
o_e	$-\frac{1}{6}$	$\frac{1}{18}$	0	$\frac{1}{2}$	$\frac{1}{9}$	0
o_b	$-\frac{1}{9}$	0	$-\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{2}$	$\frac{1}{6}$
o_c	$-\frac{1}{9}$	$-\frac{1}{9}$	$-\frac{7}{18}$	0	$\frac{1}{6}$	$\frac{1}{2}$

are—for the ordinary Euclidean metric—equally distant from a specific point (called H above). We shall now characterize the symmetrical linear forms, giving metrics in $\mathbb{R}^{n\times n}$, which possess this property. In order to set forth the following proposition, let us designate by $\{E_I | 1 \leq I \leq n^2\}$ the canonical base of $\mathbb{R}^{n\times n}$. E_I is a vector with n^2 components for which the Ith component is equal to 1 and the others to 0. We have

Proposition 24 *Let $B(X, Y)$ $(X, Y \in \mathbb{R}^{n\times n})$ be a symmetrical bilinear form giving rise to a metric in $\mathbb{R}^{n\times n}$. A necessary and sufficient condition for which the metric associated with $B(X, Y)$ is such that there exists a point K $(K \in \mathbb{R}^{n\times n})$ equally distant from all points of the cube $\{0, 1\}^{n\times n}$, is that the canonical base $\{E_I | 1 \leq I \leq n^2\}$ is orthogonal.*

Clearly, the stated condition is sufficient. If the canonical base is orthogonal for $B(X, Y)$, we have

$$\big(\forall(I, J), 1 \leq I, J \leq n\big), B(E_I, E_J) = 1(\text{resp., } 0) \text{ if } I = J(\text{resp., } I \neq J)$$

If H is the point of $\mathbb{R}^{n\times n}$ whose coordinates are all equal to $\frac{1}{2}$, then, for any point X of $\{0, 1\}^{n\times n}$, we have

$$X - H = \sum_{1\leq I\leq n^2} \frac{1}{2} E_I \tag{2.2.26}$$

Therefore,

$$B(X - H, X - H) = \frac{1}{4} n^2 \tag{2.2.27}$$

Conversely, to begin with, let us verify that if all points of the cube $\mathbb{C} = \{0, 1\}^{n\times n}$ are at equal distance of a given point H' of $\mathbb{R}^{n\times n}$, then, necessarily, $H' = H$. In fact, if X is an arbitrary point of the cube $\mathbb{C}$, associate the point X' with X, defined by $X' = U - X$, where U is the point of $\mathbb{C}$, whose coordinates are all equal to 1. Due to these assumptions, we have

$$Q(X - H) = Q(X' - H) \text{ and } Q(X - H') = Q(X' - H')$$

where $Q(Y) = B(Y, Y)$ is the square of the norm of Y, corresponding to B.

Now, let us expand $Q(X - X')$ of two different ways, by referring to H and K, respectively,

$$\begin{aligned} Q(X - X') &= B[(X - H) - (X' - H), (X - H) - (X' - H)] \\ &= 2B[(X - H), (X - H)] - 2B[(X - H), (X' - H)] \\ &= 2B[(X - H), (X - X')] \end{aligned} \tag{2.2.28}$$

Similarly, we have

$$Q(X - X') = 2B[(X - H'), (X - X')] \tag{2.2.29}$$

From (2.2.28) and (2.2.29), we obtain

$$B(H' - H, X - X') = 0 \tag{2.2.30}$$

for all X and $X' = U - X$. Hence, $H' = H$.

H being the point of $\mathbb{R}^{n\times n}$ whose coordinates are all equal to $\frac{1}{2}$, if $Q(X - H)$ is a constant c^2, then, for every point Y whose coordinates are equal to 1 or -1, $Q(Y) = 4c^2$. In fact, there exists a single X of $\mathbb{C}$, for which $Y = 2(X - H)$. On the other hand, $Q(Y) = Q(-Y)$. It results that Q cannot include in its expression with respect to the canonical base, rectangle terms. If not, the value of $Q(Y)$ would change by substituting $Y' = -Y$ for Y. Therefore, $Q(X)$ is necessarily of the form

$$\sum_{(i,j)\in\mathbb{I}\times\mathbb{I}} q_{ij} x_{ij}^2$$

where the x_{ij}, $(i, j) \in \mathbb{I} \times \mathbb{I}$ are the coordinates of the vertex X in $\mathbb{C}$. □

Following the end of this proof, notice that if X and X' are two points of the cube $\mathbb{C}$, representing two binary relations R and R' on $\mathcal{O}$, $Q(X - X')$ can be interpreted as a measure—defined by $\{q_{ij}|(i, j) \in \mathbb{I} \times \mathbb{I}\}$—of the symmetrical difference of the graphs of R and R' in $\mathcal{O} \times \mathcal{O}$.

2.2.2 *Transfer Algorithm and Central Partition*

2.2.2.1 Transfer Algorithm

The ***transfer*** algorithm is independent of the nature of the criterion to optimize. However, its expression is more consistent in the case of a *linear* criterion with respect to the set of object pairs. In these conditions, we pay attention to a criterion of the form (2.2.25). In the latter, $t_{xy} = g_{xy} - 0.5$ plays the role of a similarity index between x and y, where g_{xy} is a similarity index comprised between 0 and 1, $(x, y) \in \mathcal{O} \times \mathcal{O}$. In fact, up to a coherent scale transformation, the latter property can be filled by a very general family of similarity indices. Nevertheless, referring to the value 0.5 cannot be justified for most of the similarity indices on $\mathcal{O}$.

For a partition P on $\mathcal{O}$, let $c(P)$ be the value of a criterion c on P, which has to be maximized. From P, we can build a very near partition P' by moving one object of its class and putting it in another one. The latter might be a new class. To fix ideas, we consider the set $\mathcal{P}_K$ of partitions on $\mathcal{O}$ into K classes at most ($\mathcal{P}_n$ is the set of all partitions). By labelling the partition classes, a partition P can be represented by a mapping f which makes correspondence between $\mathcal{O}$ and the class labels:

$$f : \mathcal{O} \longrightarrow \{1, 2, \ldots, k, \ldots, K\} \tag{2.2.31}$$

The *transfer* of an object x from its class labelled l to the class labelled k, is defined by substituting for the mapping f, that f' of $\mathcal{O}$ into $\{1, 2, \ldots, k, \ldots, K\}$, defined as follows:

$$\begin{aligned} f'(x) &= k \neq f(x) = l \\ f'(y) &= f(y) \text{ if } y \neq x \end{aligned} \tag{2.2.32}$$

This transfer is denoted in the following by $T_l^k(x)/f$.

This algorithm has to be initialized. A quick suggestion in the framework of the data structure presented in Sect. 2.2.1, consists of taking in (2.2.3) the partition P^m which maximizes $c(P)$.

By iterating the transformation (2.2.32) we can describe all the set $\mathcal{P}_K$ (see above). The algorithm consists of passing from one partition to the other (from P to P'), by transfer, choosing at each step the transfer $T_l^k(x)/f$ (see (2.2.32)), which maximizes the criterion c: $c(P') - c(P)$ maximum.

The development will be stopped when there does not remain any profitable transfer: $c(P') - c(P) < 0$ for all passages from P to P' (see (2.2.32)).

The algorithm definition above gives rise to a new distance between partitions. By denoting it τ, for a pair (P, Q) of partitions of $\mathcal{O}$, $\tau(P, Q)$ is the minimum number of transfers (2.2.32) needed to transform one of both partitions into the other. In [7, 8] some interesting formulas are given in certain conditions for the value of this distance.

2.2.2.2 Computing Central Partitions

As expressed above, the criterion to be maximized is linear with respect to the set $\mathcal{O} \times \mathcal{O}$ represented by the cube $\mathbb{C} = \{0, 1\}^{n \times n}$, see (2.2.25) which we again take up here:

$$L(S) = \sum_{1 \leq k \leq K} \left(\sum_{(x,y) \in O_k \times O_k} t_{xy} \right) \tag{2.2.33}$$

where S is the point of $\mathbb{C}$ representing the partition $P = \{O_k | 1 \leq k \leq K\}$. Besides, the notation P can replace S in $L(S)$. By considering the representation of the partition P by means of the mapping f (see (2.2.31)), $L(S)$ can be written as

$$L(S) = \sum_{1 \leq k \leq K} \left(\sum_{(x,y) \in \mathcal{O} \times \mathcal{O}, f(x) = f(y) = k} t_{xy} \right) \tag{2.2.34}$$

Let O_k and O_h be two classes of the partition P and let u be an element of O_k, the transfer of u from O_k to O_h modifies two elements of the first sum in (2.2.34), namely

$$\sum_{(x,y)\in O_k\times O_k} t_{xy} \text{ and } \sum_{(x,y)\in O_h\times O_h} t_{xy}$$

The sum of these two sums become

$$\sum_{(x,y)\in(O_k-\{u\})\times(O_k-\{u\})} t_{xy} + \sum_{(x,y)\in(O_h+\{u\})\times(O_h+\{u\})} t_{xy} \tag{2.2.35}$$

The variation of $L(S)$ is

$$\Delta\big(L(S)\big) = -\sum_{y\in O_k} t_{uy} - \sum_{x\in O_k} t_{xu} + \sum_{x\in O_h} t_{xu} + \sum_{y\in O_h} t_{uy} \tag{2.2.36}$$

By considering,

$$\big(\forall x, y, u \in \mathcal{O}\big), t_{ux} = t_{xu} \text{ and } t_{uy} = t_{yu}$$

and setting

$$A_u^l = \sum_{x|x\in O_l, x\neq u} t_{xu}\text{: attraction of the object } u \text{ by the class } l,\ 1 \le l \le K$$

we get

$$\Delta\big(L(S)\big) = 2(A_u^h - A_u^k) \tag{2.2.37}$$

$\Delta\big(L(S)\big)$ gives the gain in transferring the object u from the class O_k to the class O_h. As mentioned above, O_h might be a new class initially empty. In this case $A_u^h = t_{uu} = 0.5$.

Now, consider the table indexed by $\mathcal{O} \times \{1, 2, \ldots, l, \ldots, K\}$ with n rows and K columns

$$\mathbb{A} = \{A_u^l | u \in \mathcal{O}, 1 \le l \le K\} \tag{2.2.38}$$

In the transfer $T_k^h(u)$, only the columns k and h change according to

$$\begin{aligned} A_u^l &\longleftarrow A_u^k - t_{ux} \\ A_u^h &\longleftarrow A_u^h + t_{ux} \end{aligned} \tag{2.2.39}$$

Example

Consider the partition $Q = \big\{\{o_a, o_d, o_e, o_f\}, \{o_b, o_c\}\big\}$ of the previous example. We can observe that the transfer of the object o_e from the cluster $D_1 = \{o_a, o_d, o_e, o_f\}$ to $D_2 = \{o_b, o_c\}$, is profitable.

2.2.3 Objects with the Same Representation

The transfer algorithm has been defined at the level of a set $\mathcal{O}$ of objects. A unit element corresponds to a single object of $\mathcal{O}$. However, there may exist objects of $\mathcal{O}$ having the same descriptive representation in

$$\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_m \times \cdots \times \mathcal{C}_p \tag{2.2.40}$$

(See (2.2.1)). $\boldsymbol{a}$ designating the descriptive vector attribute $(a^1, a^2, \ldots, a^m, \ldots, a^p)$, two objects x and y have the same representation if and only if

$$\big(\forall m, 1 \leq m \leq p\big), a^m(x) = a^m(y) \tag{2.2.41}$$

If so, x and y are expected to be clustered together in the central partitions.

In fact, if we assume that x and y, for which $\boldsymbol{a}(x) = \boldsymbol{a}(y)$, are separated by a partition $P = \{O_1, O_2, \ldots, O_k, \ldots, O_K\}$, for example, $x \in O_j$ and $y \in O_h$, with $1 \leq j \neq h \leq K$, then transferring x in O_h or y in O_j increases strictly the linear form

$$L(P) = \sum_{1 \leq k \leq K} \left(\sum_{(u,v) \in O_k \times O_k} \right) t_{uv} \tag{2.2.42}$$

considered in (2.2.33).

Transferring x in O_h, entails the variation

$$\Delta_1\big(L(P)\big) = \frac{1}{2} + \sum_{v \in O_h} t_{xv} - \sum_{u \in O_j} t_{xu} \tag{2.2.43}$$

of $L(P)$. On the other hand, transferring y in O_j, gives the variation

$$\Delta_2\big(L(P)\big) = \frac{1}{2} + \sum_{u \in O_j} t_{yu} - \sum_{v \in O_h} t_{yv} \tag{2.2.44}$$

Now, the sum $\Delta_1\big(L(P)\big) + \Delta_2\big(L(P)\big)$ is equal to 1 and then, necessarily $\Delta_1\big(L(P)\big)$ or $\Delta_2\big(L(P)\big)$ is strictly positive.

In these conditions, it comes to work at the level of the image $\boldsymbol{a}(\mathcal{O})$ which defines a cloud of points in $\mathcal{C}$:

$$\boldsymbol{a}(\mathcal{O}) = \{(\xi, n_\xi) | \xi \in \mathcal{C}\} \tag{2.2.45}$$

where n_ξ is the number of objects in $\mathcal{O}$ having the same representation ξ. Clearly, the cardinal of the set of ordered pairs

$$\{(\xi, n_\xi) | \xi \in \mathcal{C}, n_\xi \neq 0\} \tag{2.2.46}$$

is lower than $n = card(\mathcal{O})$, thereby we may denote by q ($q \leq n$), the cardinality of $\boldsymbol{a}(\mathcal{O})$.

Therefore, it is advisable to work in the set $\boldsymbol{a}(\mathcal{O})$. In these conditions, the criterion $L(P)$ can be written as

$$L(P) = \sum_{(\xi,\eta)\in \boldsymbol{a}(\mathcal{O})\times \boldsymbol{a}(\mathcal{O})} n_\xi \cdot n_\eta (\varpi_{\xi\eta} - g_{\xi\eta})^2 \tag{2.2.47}$$

where $\varpi_{\xi\eta}$ is the common value of ϖ_{xy} for (x, y) in $\boldsymbol{a}^{-1}(\xi) \times \boldsymbol{a}^{-1}(\eta)$. Moreover,

$$g_{\xi\eta} = \frac{1}{p} \sum_{1\leq m\leq p} s^m(x, y) \tag{2.2.48}$$

where (x, y) belongs to $\boldsymbol{a}^{-1}(\xi) \times \boldsymbol{a}^{-1}(\eta)$. Recall that $s^m(x, y)$ stands for $\varpi^m(x, y)$, where ϖ^m is the indicator function of the partition of $\mathcal{O}$ associated with the mth attribute a^m, $1 \leq m \leq p$.

To summarize, a partition of $\mathcal{O}$ is represented as a partition of $\boldsymbol{a}(\mathcal{O})$ ($\boldsymbol{a}(\mathcal{O}) \subseteq \mathcal{C}$). The relational representation of a partition of $\mathcal{O}$ is realized at the level of the cross product $\boldsymbol{a}(\mathcal{O}) \times \boldsymbol{a}(\mathcal{O})$:

$$\big(\forall(\xi, \eta) \in \boldsymbol{a}(\mathcal{O}) \times \boldsymbol{a}(\mathcal{O})\big), \varpi_{\xi\eta} = \varpi_{xy} \text{ for } (x, y) \in \boldsymbol{a}^{-1}(\xi) \times \boldsymbol{a}^{-1}(\eta) \tag{2.2.49}$$

The linear form of the criterion (see (2.2.24)) becomes

$$L(P) = \sum_{(\xi,\eta)\in \boldsymbol{a}(\mathcal{O})\times \boldsymbol{a}(\mathcal{O})} t_{\xi\eta} s_{\xi\eta} \tag{2.2.50}$$

Due to the previous development and in particular to Eq. (2.2.47), we have

Proposition 25 *By adopting on* $\mathbb{R}^{n\times n}$ *the metric*

$$Q(R) = \sum_{(\xi,\eta)\in \boldsymbol{a}(\mathcal{O})\times \boldsymbol{a}(\mathcal{O})} n_\xi \cdot n_\eta \rho^2_{\xi\eta} \tag{2.2.51}$$

where $R = \{\rho_{\xi\eta} | (\xi, \eta) \in \boldsymbol{a}(\mathcal{O}) \times \boldsymbol{a}(\mathcal{O})\}$ *is a valuated binary relation on* $\boldsymbol{a}(\mathcal{O})$*, central partitions defined at the level of* $\boldsymbol{a}(\mathcal{O})$ *are identical to those defined at the level of* $\mathcal{O}$*.*

This new metric has the same algebraic properties as those of the metric defined initially, which are expressed in Sect. 3.3.1; particularly, the barycenter property and that stated in Proposition 24. However, previously, the square distance between two relations X and Y on $\mathcal{O}$ was the cardinal of the symmetrical difference of their respective graphs in $\mathcal{O} \times \mathcal{O}$. Now, X and Y have to be defined on the representation set $\mathcal{C}$ (see (2.2.40)). Moreover, $Q(X - Y)$ represents the sum of the products $n_\xi . n_\eta$ over all the pairs (ξ, η) in $\mathcal{C} \times \mathcal{C}$ for which exactly one of both relations is satisfied. Consequently, it is a measure—with integer values—of the symmetrical difference of the graphs of X and Y in $\mathcal{C} \times \mathcal{C}$.

2.2.4 Statistical Asymptotic Analysis

2.2.4.1 Preamble

As it will be expressed in Sect. 3.1 of Chap. 3, the set $\mathcal{O}$ of objects is generally a sample of a *universe* (we can also say *population*) $\mathcal{U}$, much more vast than $\mathcal{O}$ and even infinite. Mostly, $\mathcal{U}$ is impossible to observe in its totality. And even if exhaustive observation is possible, it is of importance—particularly for computational complexity reasons—to know the validity of the inference obtained from a sample of $\mathcal{U}$.

On the other hand, dually, a complete description of the $\mathcal{O}$ elements is impossible. More often than not, the attributes retained for the description of $\mathcal{O}$, constitute a sample of a much larger set of attributes.

In these conditions, it is interesting to study some statistical asymptotic problems. We present two convergence statistical problems concerning the central partition method analysed by S. Régnier (1966), reported in [37] and worked on again in [23], Chap. 4, or [25], Chap. 1. In the case concerned, the descriptive attributes are nominal categorical (see Sect. 3.3.1 of Chap. 3). Extension of this analysis to other types of descriptions might be envisaged. The presentation of these asymptotic convergence problems require basic notions of probability theory, which can be found in [28, 33]. In the following development, only results obtained will be stated.

2.2.4.2 The First Convergence Problem

The object set $\mathcal{O}$ denoted here by $\{o_i | 1 \leq i \leq n\}$ is viewed as the realization of a random sample $\mathcal{O}^\star = \{o_i^\star | 1 \leq i \leq n\}$ of independent random elements provided from a universe $\mathcal{U}$ of objects. $\mathcal{U}$ might be infinite, nonetheless, generally, $\mathcal{U}$ is finite, its cardinality N ($N = \mathcal{U}$) being "very" large. $\mathcal{U}$ is supposed to be endowed with a probability measure that we designate by Φ. A realization $\mathcal{O}$ of $\mathcal{O}^\star$ is obtained by a random drawing with replacement of a sequence of n objects. Due to the bigness of N with respect to n, the probability to get more than once the same object in $\mathcal{O}$, is extremely small.

The mapping $\boldsymbol{a} = (a^1, a^2, \ldots, a^m, \ldots, a^p)$ of the universe $\mathcal{U}$ into the Cartesian product

$$\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_m \cdots \times \mathcal{C}_p$$

defining the description of $\mathcal{U}$ (see (2.2.1)) is supposed fixed. $\mathcal{C}$ is finite and $\boldsymbol{a}$ is assumed measurable for Φ and for the uniform measure on $\mathcal{C}$. Thereby, every part of $\mathcal{C}$ is measurable and equally, the reverse image $\boldsymbol{a}^{-1}$ of the subset set of $\mathcal{C}$ into the subset set of $\mathcal{U}$ determines a set of $\mathcal{U}$ subsets, measurable for Φ (see [28, 33] for the definition of the measurability of a mapping).

A probability measure Ψ on $\mathcal{C}$ is induced from Φ as follows:

$$\Psi(\xi) = \Phi\left(\boldsymbol{a}^1(\xi)\right) \tag{2.2.52}$$

The mapping $\boldsymbol{a}$ defines the representation of the random set $\mathcal{O}^\star$ of objects in $\mathcal{C}$. The sequence of images

$$\{\Omega_i = \boldsymbol{a}(o_i^\star) | 1 \leq i \leq n\} \tag{2.2.53}$$

is a sequence of n independent random elements of $\mathcal{C}$, endowed with the probability measure Ψ

The realization of the sequence (2.2.53) in the form

$$\{\omega_i = \boldsymbol{a}(o_i) | 1 \leq i \leq n\} \tag{2.2.54}$$

determines a discrete probability measure on $\mathcal{C}$, defined by

$$(\forall \omega \in \mathcal{C}),\ \Psi_n(\omega) = \frac{n_\omega}{n} \tag{2.2.55}$$

where n_ω is the number of objects of $\mathcal{O}$, whose representation is ω: $n_\omega = card\left(\boldsymbol{a}^{-1}(\omega) \cap \mathcal{O}\right)$.

According to the previous proposition, a central partition of $\mathcal{O}$ (resp., $\mathcal{O}^\star$) is a reciprocal image by $\boldsymbol{a}$ of a central partition of $\mathcal{C}$, when $\mathcal{C}$ is endowed with the probability measure Ψ_n (resp., Ψ). In these conditions, we consider the central partition search in $\mathcal{C}$.

Π designating the partition set of $\mathcal{C}$, let π be an arbitrary element of Π ($\pi \in \Pi$). According to (2.2.47), define

$$D_n(\pi) = \sum_{(\xi,\eta) in \mathcal{C} \times \mathcal{C}} (\varpi_{\xi\eta} - g_{\xi\eta})^2 \Psi_n(\xi) \cdot \Psi_n(\eta) \tag{2.2.56}$$

where $\varpi_{\xi\eta}$ is the indicator function of the equivalence relation on $\mathcal{C}$ associated with π ($\varpi_{\xi\eta} = 1$ if ξ and η are in the same class of π and 0 if not). On the other hand,

$$g_{\xi\eta} = \frac{1}{p} \sum_{1 \leq m \leq p} s^m_{\xi\eta}$$

where $s^m_{\xi\eta}$ is the value for (ξ, η) of the indicator function on $\mathcal{C}$, associated with the image of the partition of $\mathcal{O}$, defined by the attribute a^m, $1 \leq m \leq p$.

The set of central partitions of $\mathcal{C}$, provided with the probability measure Ψ_n, is defined as the following set C_n of partitions π where $D_n(\pi)$ (see Eq. (2.2.56)) reaches its lower bound:

$$C_n = \{\chi | D_n(\chi) \leq D_n(\pi) \text{ for all } \pi \in \Pi\} \tag{2.2.57}$$

Similarly, putting

$$D(\pi) = \sum_{(\xi,\eta) in \mathcal{C} \times \mathcal{C}} (\varpi_{\xi\eta} - g_{\xi\eta})^2 \Psi(\xi).\Psi(\eta) \tag{2.2.58}$$

we have

$$C = \{\chi | D(\chi) \leq D(\pi) \text{ for all } \pi \in \Pi\} \tag{2.2.59}$$

Relative to the latter equations, recall that $D_n(\pi)$ (resp., $D(\pi)$) is associated with $\mathcal{O} = \{o_i | 1 \leq i \leq n\}$ (resp., with $\mathcal{O}^\star = \{o_i^\star | 1 \leq i \leq n\}$). In the framework of $\mathcal{C}^n$—realization of $\mathcal{C}$ on $\mathcal{O}$—endowed with its subset set and the probability measure induced by Ψ, the statistic $D_n(\pi)$, for a given π, is a realization of a random variable, also denoted—without ambiguity—by $D_n(\pi)$. To the random vector $\big(D_n(\pi) | \pi \in \Pi\big)$ corresponds the centre C_n (see (2.2.57)) which is a random subset of Π. The asymptotic behaviour of C_n $(n \to \infty)$ is stated as follows:

Proposition 26 *Almost surely*

$$\bigcap_n \bigcup_{l \geq n} C_l \subset C \tag{2.2.60}$$

that is to say,

$$Prob\left\{ \bigcap_n \bigcup_{l \geq n} C_l \subset C \right\} = 1$$

On the other hand,

$$lim_{n \to \infty} Prob\{C_n \subset C\} = 1 \tag{2.2.61}$$

The left member of (2.2.60) corresponds to $limsup_{n \to \infty} C_n$

2.2.4.3 The Second Convergence Problem

We assume here that the set $\mathcal{O}$ of objects is fixed. Each descriptive attribute a^m ($a^m : \mathcal{O} \to \mathcal{C}_m$) induces a partition P^m of $\mathcal{O}$, $1 \leq m \leq p$. The central partitions are those which minimize the function, already considered in (2.2.7):

$$\delta_p(P) = \frac{1}{p} \sum_{1 \leq m \leq p} \delta(P, P^m) \tag{2.2.62}$$

where P is an element of the set $\mathcal{P}(\mathcal{O})$ of all partitions and where $\delta(P, P^m)$ designates the cardinal of the symmetrical difference between the graphs in $\mathcal{O} \times \mathcal{O}$ of the equivalence relations associated with P and P^m, respectively, $1 \leq m \leq p$.

In this problem of asymptotic behaviour, the sequence of attributes $(a^m | 1 \leq m \leq p)$ is regarded as a sample of a large set of attributes. More precisely, we suppose that the sequence of partitions $\{P^m | 1 \leq m \leq p\}$ is a sample of p independent random elements, provided from $\mathcal{P}(\mathcal{O})$, where the latter partition set is endowed with the algebra of the set of its subsets and with a probability measure Λ.

$\{P^m | 1 \leq m \leq p\}$ determines an empirical probability measure $\Lambda_p(X)$, for X belonging to $\mathcal{P}(\mathcal{O})$, which is the relative frequency (proportion) of partitions identical to X. Equation (2.2.62) becomes

$$\delta_p(P) = \sum_{X \in \mathcal{P}(\mathcal{O})} \delta(P, X) \Lambda_p(X) \tag{2.2.63}$$

In this empirical case, the set denoted by C_p of central partitions of $\mathcal{O}$ is defined with respect to the family $\{P^m | 1 \leq m \leq p\}$ of $\mathcal{P}(\mathcal{O})$ elements, as follows:

$$C_p = \{P | P \in \mathcal{P}(\mathcal{O}) \text{ and } \delta_p(P) \leq \delta_p(Y) \text{ for all } Y \in \mathcal{P}(\mathcal{O})\} \tag{2.2.64}$$

When p tends to infinity, the empirical distribution Λ_p tends to a distribution Λ. Putting

$$\delta(P) = \sum_{X \in \mathcal{P}(\mathcal{O})} \delta(P, X) \Lambda(X) \tag{2.2.65}$$

the associated centre C, where $\delta(P)$ reaches its minimum, is written as

$$C = \{P | P \in \mathcal{P}(\mathcal{O}) \text{ and } \delta(P) \leq \delta(Y) \text{ for all } Y \in \mathcal{P}(\mathcal{O})\} \tag{2.2.66}$$

For P fixed, by associating a family of random partitions

$$\{P^{1\star}, P^{2\star}, \ldots, P^{m\star}, \ldots, P^{p\star}\}$$

with the observed one, $\delta_p(P)$ becomes a random variable defined on $\mathcal{P}(\mathcal{O})^p$ endowed with the algebra of its subset set and the probability measure induced by Λ. In other words, the limit distribution

$$\{(X, \Lambda(X))|X \in \mathcal{P}(\mathcal{O})\}$$

is substituted for the empirical distribution

$$\{(X, \Lambda_p(X))|X \in \mathcal{P}(\mathcal{O})\}$$

In this situation

$$\left(\delta_p(P)|P \in \mathcal{P}(\mathcal{O})\right)$$

becomes a random vector and the associated centre C_p becomes a random subset of $\mathcal{P}(\mathcal{O})$.

Λ is assumed to be obtained as a limit of Λ_p, when p tends to infinity. C_p (resp., C) indicating the centre for Λ_p (resp., Λ), we have

Proposition 27 *Almost surely*

$$limsup_{p\to\infty} = \bigcap_p \bigcup_{l\geq p} C_l \subset C \tag{2.2.67}$$

that is to say,

$$Prob\left\{\bigcap_p \bigcup_{l\geq p} C_l \subset C\right\} = 1$$

On the other hand,

$$lim_{p\to\infty} Prob\{C_p \subset C\} = 1 \tag{2.2.68}$$

The proofs of the latter two propositions can be found in the references given in the introductory section (see Sect. 2.2.4.1).

2.2.5 *Remarks on the Application of the Central Partition Method and Developments*

In Sect. 2.2.1.3, relative to a pair of objects (x, y) in $\mathcal{O} \times \mathcal{O}$, we have seen introduced in a natural way the similarity index

$$g_{xy} = \frac{1}{p} \sum_{1\leq m\leq p} s_{xy}^m$$

(See (2.2.23)) which represents the proportion of categorical attributes having the same value in both objects x and y. g_{xy} or more exactly $t_{xy} = g_{xy} - 0.5$ (see (2.2.24)) is the similarity index sustaining the *Central Partition Method*.

The weakest point of this technique is due to the fact that, in principle, all the descriptive categorical attributes are considered as equally discriminant. A consequence of this is the importance given to the reference value $\frac{1}{2}$ for g_{xy}. Thus, the value 0 of t_{xy} must correspond to the case for which x and y are moderately similar (middle resemblance).

Nonetheless, in the general case where the object set $\mathcal{O}$ is provided with a numerical similarity index $Sim(x, y)$, deduced from a description by attributes of any sort (see Chaps. 3 and 7), it is always possible to reduce the value scale of $Sim(x, y)$ to the interval [0, 1], in such a way that the value 0.5 corresponds to a middle resemblance.

This reduction is in fact considered by S. Régnier in order to adapt the transfer algorithm to any similarity index *Sim*. The similarity function *Sim* is then represented by a point in the continuous cube $[0, 1]^{n\times n}$, included in $\mathbb{R}^{n\times n}$. On the other hand, the central partition P to be sought is represented in the latter cube by a point S whose coordinates ϖ_{xy} $((x, y) \in \mathcal{O} \times \mathcal{O})$ are 0 or 1 (0 (resp., 1) if x and y are separated (resp., joined) by P).

With this formalism, P is all the more preferred that the Euclidean distance $\|S - Sim\|$ is small ($\|\bullet\|$ is the Euclidean norm in $\mathbb{R}^{n\times n}$). Therefore, the criterion established in Sect. 2.2.1.3 applies. It becomes

$$L(S) = \sum_{(x,y)\in\mathcal{O}\times\mathcal{O}} \left(Sim(x, y) - \frac{1}{2}\right) \varpi_{xy} \tag{2.2.69}$$

However, this criterion becomes independent of the central partition notion.

This linear form of an association coefficient between a similarity *Sim* and a partition P of $\mathcal{O}$ appears clearly in our statistical analysis of a criterion established to validate a clustering of $\mathcal{O}$. This criterion is written in the form

$$C(Sim, P) = \sum_{\{x,y\}\in F} \varpi_{xy}.c(x, y) \tag{2.2.70}$$

(See Eq. 9.3.71 of Sect. 9.3.5.2 of Chap. 9) where $c(x, y)$ is a normalized version of $Sim(x, y)$. This normalization of statistical nature is carried out empirically with respect to the set F of unordered object pairs of $\mathcal{O}$. It plays a fundamental role in the *LLA* clustering approach. Therefore, $C(Sim, P)$ is not a linear function of the similarity table

$$\{Sim(x, y)|(x, y) \in \mathcal{O} \times \mathcal{O}\} \tag{2.2.71}$$

In the method built by F. Marcotorchino and P. Michaud [31] the criterion to maximize takes the linear form

$$L(Sim, P) = \sum_{(x,y)\in\mathcal{O}\times\mathcal{O}} \varpi_{xy} \cdot Sim(x, y) = \sum_{1\le k\le K} \sum_{(i,j)\in\mathbb{I}_k\times\mathbb{I}_k} \varpi_{i,j} \cdot Sim_{ij} \qquad (2.2.72)$$

where $\mathbb{I}_k$ is the index set coding the kth $\mathcal{O}_k$, $1 \le k \le K$.

The originality of the technique mentioned resides in using *linear programming under constraints* as follows:

Maximize $L(Sim, P)$ under the conditions

$$\begin{aligned} &\varpi_{i,j} \in \{0, 1\} \\ &\varpi_{i,i} = 1 \\ &\varpi_{i,j} = \varpi_{j,i} \\ &(\forall(i, j, l) \in \mathbb{I} \times \mathbb{I} \times \mathbb{I})\ \varpi_{i,j} + \varpi_{j,l} - \varpi_{i,l} \le 1 \end{aligned} \qquad (2.2.73)$$

The solution provided corresponds to a *local* maximum.

In [30] an interesting and rich formalization is proposed for similarity indices on $\mathcal{O}$ in the case of a description by *Boolean* attributes. A large ensemble of similarity functions is mutually compared according to algebraic properties. We cannot here discuss this subject any further.

An important research direction indicated by A.K. Jain in an overview article [18] concerns the development of *ensemble clustering*. The aim consists of deriving a clustering on the set concerned $\mathcal{O}$ from a sequence of partitions of $\mathcal{O}$, obtained by applying sequentially a parametrized clustering algorithm on $\mathcal{O}$. In this process, the parametrization changes from one application to the next one. Iteration of the K-means algorithm is considered in [18]. In these conditions, each of the partitions obtained infers a nominal categorical attribute on $\mathcal{O}$. Consequently, the methods presented above apply and enable us to obtain a consensus partition of the different partitions obtained, in particular by the K-means algorithm. This algorithm is involved in the next section.

By an analogous way, we can also propose to summarize different partitions of $\mathcal{O}$, obtained from repeatable applications of a parametrized non-hierarchical clustering algorithm, by means of an *ascendant agglomerative hierarchical clustering*. For this purpose, the general method expressed in Sect. 7.2.1 of Chap. 7 can be followed. More particularly, we put for $S^j(x, y)$ of (7.2.1) the index corresponding to $S^j(o_i, o_{i'})$ of (7.2.15), $1 \le j \le p$, where p is the number of partitions obtained by repeating the non-hierarchical clustering algorithm (e.g. the K-means algorithm).

2.3 Dynamic and Adaptative Clustering Method

2.3.1 Data Structure and Clustering Criterion

As mentioned in the Preamble (see Sect. 2.1), we shall describe the initial version of the *dynamic cluster* method proposed by Diday [10, 11]. In this version the object set

$\mathcal{O}$ to be clustered is represented by a set of points in the geometrical space $\mathbb{R}^p$, where $\mathbb{R}$ is the reals and p, a positive integer, corresponding to the number of numerical attributes describing $\mathcal{O}$. To fix ideas, we may suppose $\mathbb{R}^p$ endowed with the ordinary Euclidean metric. On the other hand, if $P = \{O_1, O_2, \ldots, O_k, \ldots, O_K\}$ is a partition by proximity of $\mathcal{O}$, each of its classes is represented by a subset of $\mathcal{O}$, comprising a few elements, $1 \leq k \leq K$. This cluster representation, adequately determined, is expected to be more accurate than that defined by the centre of gravity (centroid) of the cluster concerned, as in the K-means algorithm (see Sect. 2.3.2). For simplicity, historical reasons and theoretical properties, the K-means version is mostly employed [18]. An interesting historical overview of this general approach is given in [3].

To begin let us specify our notations. $\mathcal{V} = \{v^j | 1 \leq j \leq p\}$ designates the set of numerical descriptive attributes and $\mathcal{O} = \{o_i | 1 \leq i \leq n\}$, the set of the objects described (see Sect. 3.2.2 of Chap. 3). The set $\mathbb{I} = \{1, 2, \ldots, i, \ldots, n\}$ codes $\mathcal{O}$ and can be decomposed, according to the partition P (see above), as $\{\mathbb{I}_k | 1 \leq k \leq K\}$, where $\mathbb{I}_k$ is the subset of $\mathbb{I}$, coding the class O_k, $1 \leq k \leq K$. n_k denoting the cardinality of O_k, or equivalently $\mathbb{I}_k$, we have

$$\sum_{1 \leq k \leq K} n_k = n = card(\mathcal{O}) \tag{2.3.1}$$

The point of $\mathbb{R}^p$, representing an object x of $\mathcal{O}$, can be written as

$$v(x) = \left(v^1(x), v^2(x), \ldots, v^j(x), \ldots, v^p(x)\right) \tag{2.3.2}$$

In the K-means algorithm, the definition of the criterion evaluating a partition P of the object set $\mathcal{O}$ depends *uniquely* on P. $P = \{O_1, O_2, \ldots, O_k, \ldots, O_K\}$, designating the partition of $\mathcal{O}$ into K non-empty classes, this criterion is a measure evaluating globally and additively the respective cohesions of the different classes O_k, $1 \leq k \leq K$. As mentioned above, each cluster O_k is represented by its centre of gravity, which we denote by g_k, $1 \leq k \leq K$. More precisely,

$$g_k = \frac{1}{n} \sum_{i \in \mathbb{I}_k} o_i \tag{2.3.3}$$

where—without ambiguity—o_i stands for the point of $\mathbb{R}^p$, representing the object o_i, $i \in \mathbb{I}$. Equation (2.3.3) can be written as

$$g_k - O = \frac{1}{n_k} \sum_{i \in \mathbb{I}_k} (o_i - O)$$

it reflects the vector writing

$$\vec{Og_k} = \frac{1}{n_k} \sum_{i \in \mathbb{I}_k} \vec{Oo_i}$$

where O designates here the origin point of $\mathbb{R}^p$.

The cohesion of a given cluster O_k is defined by the sum of squared distances of the different points of $\mathbb{R}^p$, representing O_k, to the gravity centre of these points, namely

$$W_k = \sum_{i \in \mathbb{I}_k} d^2(o_i, g_k) = \sum_{i \in \mathbb{I}_k} \|\overrightarrow{o_i g_k}\|^2 \tag{2.3.4}$$

Thereby, the cohesion of O_k is all the more strong as W_k is small. In these conditions, the global cohesion of the partition P is evaluated by the smallness of

$$W(P) = \sum_{1 \le k \le K} W_k \tag{2.3.5}$$

Consider now the cloud of points in $\mathbb{R}^p$ and designate it by $\mathcal{N}(\mathbb{I})$ (see Sect. 10.3.3 of Chap. 10). The inertia moment of $\mathcal{N}(\mathbb{I})$ can be written as

$$\mathcal{M}\big(\mathcal{N}(\mathbb{I})\big) = \sum_{i \in \mathbb{I}} d^2(o_i, G) \tag{2.3.6}$$

where G is the centre of gravity of $\mathcal{N}(\mathbb{I})$. We have

$$G = \frac{1}{n} \sum_{i \in \mathbb{I}} o_i = \frac{1}{n} \sum_{1 \le k \le K} n_k g_k \tag{2.3.7}$$

where n_k, specified above, is the cardinality of O_k, $1 \le k \le K$.

The general inertia decomposition formula with respect to the partition P is

$$\mathcal{M}\big(\mathcal{N}(\mathbb{I})\big) = \sum_{1 \le k \le K} (W_k + B_k) \tag{2.3.8}$$

where

$$B_k = n_k d^2(g_k, G) \tag{2.3.9}$$

In the previous formulas (2.3.3)–(2.3.9) all vertices in the cloud of $\mathbb{R}^p$, associated with $\mathcal{O}$, are equally weighted. It is easy to generalize the methods studied below, in the possible case of a representation by a cloud of points in $\mathbb{R}^p$, non-equally weighted. In the latter case, the inertia decomposition formula becomes (10.3.10) of Sect. 10.3.3.1 of Chap. 10.

The K-means algorithm described in the next section comprises successive steps of the same nature. In each of them, the criterion (2.3.5) is minimized at best. This criterion is interpreted as the "lost" (or "residual") inertia for the representation of $\mathcal{O}$ by the cloud

$$\mathcal{N}(\mathcal{G}) = \{(g_k, n_k) | 1 \le k \le K\} \tag{2.3.10}$$

Other criteria having the same general analytical form as (2.3.5) can be substituted for $W(P)$, that is,

$$\sum_{1\leq k\leq K}\sum_{i\in\mathbb{I}_k}\delta(o_i, e_k) \tag{2.3.11}$$

where δ is a dissimilarity index on $\mathcal{O}$ and e_k, a prototype of the cluster O_k, $1 \leq k \leq K$.

Equation (2.3.11) can be denoted by $\Delta(L, P)$, where

$$L = (e_1, e_2, \ldots, e_k, \ldots, e_K) \tag{2.3.12}$$

is the prototype sequence, e_k, representing the cluster O_k, $1 \leq k \leq K$.

Equation (2.3.11) is the form of the criterion employed in dynamic cluster methods [3, 10, 11]. In order to make explicit this criterion, let us suppose the partition P with labelled classes. Hence, P is defined as the ordered sequence

$$P = (O_1, O_2, \ldots, O_k, \ldots, O_K) \tag{2.3.13}$$

$t(P) = (n_1, n_2, \ldots, n_k, \ldots, n_K)$, where $n_k = card(O_k)$, $1 \leq k \leq K$, is the partition *type* of P.

In this technique, the setting of the criterion depends on two arguments: the partition P (see (2.3.13)) and a sequence of integers

$$(m_1, m_2, \ldots, m_k, \ldots, m_K) \tag{2.3.14}$$

such that

$$(\forall k, 1 \leq k \leq K), m_k < n_k$$

where m_k specifies the cardinal of the subset e_k of the *entire* set $\mathcal{O}$, which represents the class O_k, $1 \leq k \leq K$. e_k will be called the *kernel* of O_k.

$\Delta(L, P)$ (see (2.3.11)) can be normalized in order to refer to a suitable scale taking into account the respective cardinalities of the e_k, $1 \leq k \leq K$. In practice, mostly, the different integers m_k are taken mutually equal in the form $m_k = [\alpha \times (n/K)]$, where α is a proportion chosen (e.g. $\alpha = 0.1$) and where $[\bullet]$ designates the integer part of $\bullet$. To fix ideas, without loss of generality, the latter option is considered in the following; that is, $m_k = m$ for all k, $1 \leq k \leq K$.

In these conditions, given P, the kernel e_k representing the kth class O_k, is defined as

$$Arg\left[\min_{e\in\mathcal{P}(m,\mathcal{O})}\left(\sum_{(x,y)\in O_k\times e}\delta(x, y)\right)\right] \tag{2.3.15}$$

where $\mathcal{P}(m, \mathcal{O})$ is the set of all subsets of cardinal m. There are in all $\binom{n}{m}$ such subsets.

In fact, e is obtained by carrying out a sort of the m $\mathcal{O}$ elements which are the nearest O_k cluster, according to the dissimilarity index

$$\delta(y, O_k) = \sum_{x \in O_k} \delta(x, y) \tag{2.3.16}$$

Generally, only one solution exists for e, answering (2.3.15). If more than a single solution can be provided, the first encountered is selected.

The global fitting criterion between L (see (2.3.12)) and P (see (2.3.13)), corresponding to (2.3.11), can be written as

$$\Delta(L, P) = \sum_{1 \le k \le K} \sum_{(x,y) \in O_k \times e_k} \delta(x, y) \tag{2.3.17}$$

Several types of δ dissimilarity indices can be considered [3]. The most classical one is associated with the Euclidean representation of $\mathcal{O}$ in $\mathbb{R}^p$. This index can be written as

$$d^2(x, y) = \sum_{1 \le j \le p} (\xi^j - \eta^j)^2 \tag{2.3.18}$$

where $(\xi^j | 1 \le j \le p)$ and $(\eta^j | 1 \le j \le p)$ are the respective coordinates of the points denoted by x and y, representing the objects x and y in $\mathbb{R}^p$. Notice that (2.3.18) is directly derived from the distance index used in the K-means method.

Let us indicate here that the *LLA* approach enables us to get with a unique principle, dissimilarities called *informational* dissimilarities, for a very large scope of data structures. Different types of objects can be compared. The general form of the dissimilarity table concerned is given in (7.2.7) of Sect. 7.2.1 of Chap. 7.

2.3.2 The K-Means Algorithm

We consider the geometrical representation of the object set $\mathcal{O}$ by a cloud of points designated by $\mathcal{N}(\mathbb{I})$ ($\mathcal{N}(\mathbb{I}) = \{o_i | i \in \mathbb{I}\}$), equally weighted, in the space $\mathbb{R}^p$, endowed with the usual Euclidean metric (see (2.3.18)).

The initial state of the algorithm might be either a partition

$$P^0 = \left\{O_1^0, O_2^0, \ldots, O_k^0, \ldots, O_K^0\right\}$$

of $\mathcal{O}$, or a set

$$G^0 = \left\{g_1^0, g_2^0, \ldots, g_k^0, \ldots, g_K^0\right\}$$

of K distinct points. These define a system of K attraction centres. They are determined at random or from expert knowledge. Without loss of generality, we assume to start with

$$G^0 = \left\{g_1^0, g_2^0, \ldots, g_k^0, \ldots, g_K^0\right\} \tag{2.3.19}$$

In these conditions, the first step of the K-means algorithm consists of attaching each vertex of the cloud $\mathcal{N}(\mathbb{I})$ to the nearest attraction centre belonging to G^0. In the case where more than a single element satisfies the proximity requirement, the first encountered is selected. $\mathcal{O}$ is then divided into K classes. Let us denote by

$$P^0 = \left\{O_1^0, O_2^0, \ldots, O_k^0, \ldots, O_K^0\right\} \tag{2.3.20}$$

the partition obtained. O_k^0 is composed of the set of objects of $\mathcal{O}$ assigned to g_k^0, $1 \leq k \leq K$. We have

$$O_k^0 = \{x | x \in \mathcal{O}, d(x, g_k^0) \leq d(x, g_{k'}^0) \text{ for } k' \neq k, 1 \leq k, k' \leq K\} \tag{2.3.21}$$

As mentioned above, for a given x in $\mathcal{O}$, k is the lowest label for which the condition

$$d(x, g_k^0) \leq d(x, g_{k'}^0) \text{ for } k' \neq k$$

holds.

Notice that some of the classes of P^0 might be empty. In fact, some attraction centres in G^0—which do not correspond to elements of $\mathcal{N}(\mathbb{I})$—can remain isolated in the assignation process. In these conditions, let

$$P^1 = \left\{O_1^1, O_2^1, \ldots, O_k^1, \ldots, O_{K(1)}^1\right\} \tag{2.3.22}$$

be the partition of $\mathcal{O}$ into the non-empty classes of P^0, $K(1) \leq K$.

From P^1 a new system

$$G^1 = \left\{g_1^1, g_2^1, \ldots, g_k^1, \ldots, g_{K(1)}^1\right\} \tag{2.3.23}$$

of attraction centres is determined, where

$$g_k^1 = \frac{1}{n_k^1} \sum_{i \in \mathbb{I}_k^1} o_i \tag{2.3.24}$$

where $\mathbb{I}_k^1$ is the subset of $\mathbb{I}$ coding the class O_k^1, whose cardinality is denoted by n_k^1, $1 \leq k \leq K(1)$.

It is exceptional that two different gravity centres of G^1 coincide. However, mathematically, this case cannot be excluded. Hence, we may introduce

$$G^2 = \left\{g_1^2, g_2^2, \ldots, g_k^2, \ldots, g_{K(2)}^2\right\} \tag{2.3.25}$$

which is composed of mutually distinct elements of G^1, $K(2) \leq K(1)$.

It is easy to observe that each step of this algorithmic process decreases the criterion $W(P)$ (see (2.3.5)). Consequently, this process converges. In the case where this convergence is reached too slowly, the process is stopped when the relative decreasing of $W(P)$ becomes negligible.

The asymptotic convergence of this algorithm under different probabilistic or continuous models is studied in [2, 4, 26, 34]. This facet will be described more in Sect. 2.3.4.2.

To end this section let us recall the version of the K-means algorithm as it was named and defined by McQueen [29]. In this, the reactualization of the attraction centre (centre of gravity) of a given cluster is carried out as the class formation goes along. More precisely, suppose that at a given step, C is a cluster already constituted and denote by g the gravity centre of C. If at the next step an object x of $\mathcal{O}$ is assigned to C—because g is the x nearest gravity centre, among the different gravity centres of the clusters formed—then g is replaced by the centre of gravity of $C + \{x\}$. For more details we may refer to [1] pages 162–163. In the following, we will not be concerned in generalizing the McQueen version.

2.3.3 *Dynamic Cluster Algorithm*

Assume that K is the upper bound of the number of classes of the partition P to be sought. Moreover, suppose that each of the P classes is represented by a subset of the object set $\mathcal{O}$, whose cardinality is m. As expressed in the preceding section, m is generally a small number with respect to $n = card(\mathcal{O})$. As said, m might correspond to the integer part of $\alpha \times (n/K)$, where α is a small proportion, for example $\alpha = 0.1$.

Now, let $\mathcal{P}_K$ be the set of all labelled partitions of $\mathcal{O}$ into K non-empty classes at the most. We have

$$card(\mathcal{P}_K) = K^n \tag{2.3.26}$$

On the other hand let $\mathcal{L}_K$ designate the set of sequences of K kernels at the most, of equal size m. Thus, each kernel is defined by a subset of $\mathcal{O}$ whose cardinality is m. Therefore, there are $\binom{n}{m}$ choices for a given kernel e. The latter corresponds to a given non-empty component of a given L in $\mathcal{L}_K$. L will be denoted by $(e_1, e_2, \ldots, e_k, \ldots, e_K)$ and called a system of kernels.

2.3.3.1 Two Mappings ν and π: Allocating and Centring

ν makes correspondence from $\mathcal{P}_K$ to $\mathcal{L}_K$. To each element P of $\mathcal{L}_P$, ν will associate an element L of $\mathcal{L}_K$ which maximizes the fitting of L to P. More formally, ν is a mapping of $\mathcal{P}_K$ to $\mathcal{L}_K$:

$$\begin{aligned} \nu : \mathcal{P}_K &\longrightarrow \mathcal{L}_K \\ P &\mapsto \nu(P) = L \end{aligned} \tag{2.3.27}$$

where L is a kernel system minimizing the criterion

$$\Delta(L, P) = \sum_{1 \le k \le K} \sum_{i \in \mathbb{I}_k} \delta(o_i, e_k) \tag{2.3.28}$$

considered in (2.3.11).

To fix ideas, assume in (2.3.27) that P is composed of K non-empty classes and follow the development of the previous section: Eqs. (2.3.11)–(2.3.17). The kernel e_k is obtained according to (2.3.15), as associated with the cluster O_k, $1 \le k \le K$. However, two distinct clusters might give rise to the same kernel. In these conditions, $\nu(P)$ is obtained by reducing the initial sequence $(e_1, e_2, \ldots, e_k, \ldots, e_K)$, preserving all mutually distinct kernels.

π is a mapping of $\mathcal{L}_K$ into $\mathcal{P}_K$:

$$\begin{aligned} \pi : \mathcal{L}_K &\longrightarrow \mathcal{P}_K \\ L &\mapsto \pi(L) = P \end{aligned} \tag{2.3.29}$$

where the partition P is obtained by assigning each of the $\mathcal{O}$ objects to the nearest kernel of L, according to the criterion

$$\delta(e, x) = \sum_{y \in e} \delta(y, x) \tag{2.3.30}$$

In this way, each of the obtained P classes surrounds at the best one kernel of L. More precisely, the kth class of P, $1 \le k \le K$, is defined as

$$O_k = \{x | \delta(e_k, x) \le \delta(e_{k'}, x) \text{ for } k' \ne k, 1 \le k' \le K\} \tag{2.3.31}$$

Now, if we consider the geometrical framework in which this algorithm lies, the dissimilarity $\delta(e, x)$ (see (2.3.30)) has to be defined by the sum of squared distances between x and the different elements of e, that is,

$$D(e, x) = \sum_{y \in e} d^2(y, x) \tag{2.3.32}$$

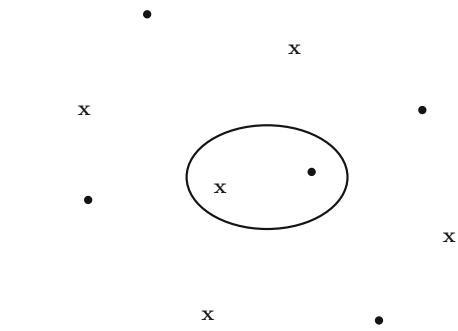

Fig. 2.1 A single kernel for two clusters

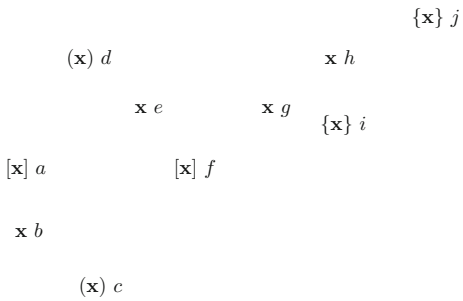

Fig. 2.2 A single cluster for two kernels

For this distance, as for possible others, a single P class might be obtained from two distinct kernels of L. In (2.3.32) we refer to (2.3.18) for $d^2(y, x)$.

A Geometrical Illustration

The distance between a set of points and a single point in a geometrical space is considered here defined as in (2.3.32). Two cases have to be illustrated.

1. A single kernel might correspond to two distinct classes;
2. Two distinct kernels might give rise to a single class.

For item 1, two clusters of points are considered in Fig. 2.1, where 10 points are involved. The points of one of both clusters are marked by "**x**" and those of the other one, by "•". Both clusters have the same kernel sized 2 which is surrounded.

For item 2, consider the set of 10 points in Fig. 2.2 and the kernel system $(\{a, f\}, \{c, d\}, \{i, j\})$. The elements of a given kernel are shown with brackets of the same type: (), [] or {}. Every point of the cluster $\{a, b, c, d, e, f\}$ is assigned to the kernel $\{a, f\}$; c and d are themselves assigned to $\{a, f\}$, according to the distance $D(e, x)$ (see (2.3.32))

2.3.3.2 Convergence Condition

Although the objective is to discover a "good" partition of $\mathcal{O}$, with clusters having strong cohesion, the dynamic cluster algorithm has to be viewed as searching a

"good" kernel system. In fact, a single step of this algorithm consists of substituting for a given kernel an improved one. This is obtained from the partition associated from the last partition obtained by means of the criterion $\delta(e, x)$ (see (2.3.30) and (2.3.31)).

Let L^0 designate the initial kernel system. The latter might be chosen at random. We have for the following kernel systems

$$\begin{aligned} L^1 &= \nu \circ \pi(L^0) \\ L^2 &= \nu \circ \pi(L^1) \\ &- - - - - - - - \\ L^i &= \nu \circ \pi(L^{i-1}) \\ &- - - - - - - - \end{aligned} \tag{2.3.33}$$

where ν and π were defined in (2.3.27) and (2.3.29). L^i is the kernel system obtained after the ith step of the algorithm. The latter stops after the tth step if

$$\nu \circ \pi(L^t) = L^t \tag{2.3.34}$$

L^t is a kernel system which cannot be enhanced. It defines a fixed point of the mapping $\nu \circ \pi$. We can write for the tth iteration

$$L^t = \nu \circ \pi(L^0) \tag{2.3.35}$$

where t is the lowest integer for which

$$(\nu \circ \pi)^{t+1}(L^0) = (\nu \circ \pi)^t(L^0) \tag{2.3.36}$$

Clearly, this algorithm is justified only if, for every i, the adequacy of L^i to the associated partition $\pi(L^i)$ is at least as good as that of L^{i-1} to $\pi(L^{i-1})$, that is, by referring to (2.3.28),

$$\Delta(L^i, \pi(L^i)) \leq \Delta(L^{i-1}, \pi(L^{i-1})) \tag{2.3.37}$$

Proposition 28 *The sequence* $\{\Delta\big(L^i, \pi(L^i)\big) | i \geq 0\}$ *decreases if for any pair* (L, M) *of kernel systems, the following condition is satisfied:*

$$\Delta\big(L, \pi(M)\big) \leq \Delta\big(M, \pi(M)\big) \Rightarrow \Delta\big(L, \pi(L)\big) \leq \Delta\big(M, \pi(M)\big) \tag{2.3.38}$$

In fact by construction (see (2.3.15)), we have

$$\Delta(L^i, \pi(L^{i-1})) \leq \Delta(L^{i-1}, \pi(L^{i-1})) \tag{2.3.39}$$

The condition (2.3.38) implies

$$\Delta(L^i, \pi(L^i)) \leq \Delta(L^i, \pi(L^{i-1})) \tag{2.3.40}$$

From (2.3.39) and (2.3.40) we obtain, by transitivity,

$$\Delta(L^i, \pi(L^i)) \leq \Delta(L^{i-1}, \pi(L^{i-1})) \tag{2.3.41}$$

Intuitively, in geometrical context, condition (2.3.38) appears as very natural. It expresses that if the kernel system L does better than M, with respect to the partition $\pi(M)$, then L does better for $\pi(L)$, than M for $\pi(M)$.

Now, by considering the dissimilarity index $D(e, x)$ (see (2.3.32)), we shall show that the condition (2.3.38) is effectively satisfied for a set of objects $\mathcal{O}$, which is represented by a cloud of points in the geometrical space $\mathbb{R}^p$.

Let

$$\big(M, \pi(M)\big) = \big((c_1, c_2, \ldots, c_k, \ldots, c_K), (C_1, C_2, \ldots, C_k, \ldots, C_K)\big) \tag{2.3.42}$$

be an ordered pair of a kernel system M and its associated partition $\pi(M)$. Recall that $card(c_k) = m$, for all k $(1 \leq k \leq K)$ and suppose, to fix ideas, that all of the classes C_k are non-empty. In these conditions, consider a kernel system

$$(e_1, e_2, \ldots, e_k, \ldots, e_K)$$

and suppose satisfied the left side of the implication (2.3.38), that is,

$$\sum_{1 \leq k \leq K} \sum_{y \in C_k} D(y, e_k) \leq \sum_{1 \leq k \leq K} \sum_{y \in C_k} D(y, c_k) \tag{2.3.43}$$

The partition

$$\pi(L) = (O_1, O_2, \ldots, O_k, \ldots, O_K)$$

where O_k is attracted by e_k, is necessary such that

$$\sum_{1 \leq k \leq K} \sum_{y \in O_k} D(y, e_k) \leq \sum_{1 \leq k \leq K} \sum_{y \in C_k} D(y, e_k) \tag{2.3.44}$$

In fact, for a given k and $y \in C_k$, if $D(y, e_h) < D(y, e_k)$, for $k \neq h$, y will be assigned to O_h in $\pi(L)$ and the associated criterion diminution is $D(y, e_k) - D(y, e_h)$. Effectively, O_h is constituted of all elements y, such that $D(y, e_h)$ is minimum.

2.3.4 Following the Definition of the Algorithm

2.3.4.1 Clustering Structures Associated with the Repetition of the Algorithm Process

To fix ideas and without real loss of generality, we suppose in this section that the set of objects $\mathcal{O}$, as represented by a cloud of n points in the geometrical space $\mathbb{R}^p$, endowed with the ordinary Euclidean metric.

The most vulnerable point in the K-means or dynamic cluster algorithms consists of initial choice of attraction centres in the K-means, or kernels, in the dynamic cluster cases, respectively. Generally, these choices are carried out at random and then, repeating the algorithm process from different systems of attraction centres (for the K-means version) or kernels (for the dynamic cluster version), is needed. This repetition enables different algorithmic results, associated with different initializations, to be compared and synthesized. In the following, we shall focus on these aspects in the case of dynamic cluster algorithm.

Strong and Weak Patterns

Let

$$(L^1, L^2, \ldots, L^j, \ldots, L^J) \tag{2.3.45}$$

be a sequence of kernel systems. We may assume that the L^j, $1 \leq j \leq J$, are obtained randomly and independently. L^j is supposed to be written in the form

$$L^j = \left(e_1^j, e_2^j, \ldots, e_k^j, \ldots, e_K^j\right) \tag{2.3.46}$$

where $card(e_k^j) = m$, $1 \leq k \leq K$, $1 \leq j \leq J$.

Now, let

$$(P^1, P^2, \ldots, P^j, \ldots, P^J) \tag{2.3.47}$$

be the sequence of partitions obtained (after convergence) by the dynamic cluster algorithm from the kernel system (2.3.45). P^j is obtained from L^j, $1 \leq j \leq J$.

In relation to the partition sequence (2.3.47), Diday introduces [10], intuitively, the respective notions of "strong pattern" ("forme forte") and "weak pattern" ("forme faible"). In fact these notions can be formally expressed with respect to the lattice order endowing the partition set of $\mathcal{O}$ (see Sect. 1.1 of Chap. 1). For this, let us begin by defining the two following partitions:

$$P^\star = \bigwedge_{1 \leq j \leq J} P^j \tag{2.3.48}$$

$$Q^{\star} = \bigvee_{1 \leq j \leq J} P^{j} \tag{2.3.49}$$

where $\bigwedge$ and $\bigvee$ denote, respectively, the *greatest lower bound* (infimum) and the *least upper bound* (supremum) operations in the partition lattice.

We shall consider hereafter the partition $P^{\star}$. If we denote the partition P^{j} as

$$P^{j} = \left\{ O_{k}^{j} | 1 \leq k \leq K_{j} \right\} \tag{2.3.50}$$

where $K_j \leq K$, a given class of $P^{\star}$ can be written as follows:

$$\bigcap_{1 \leq j \leq J} O_{k_j}^{j} \tag{2.3.51}$$

where k_j is a subscript comprised between 1 and K_j, $1 \leq j \leq J$.

Thereby, we have

$$P^{\star} = \left\{ \bigcap_{1 \leq j \leq J} O_{k_j}^{j} | 1 \leq k_j \leq K_j \right\} \tag{2.3.52}$$

The theoretical number of classes of the partition $P^{\star}$ is

$$H = \prod_{1 \leq j \leq J} K_j \tag{2.3.53}$$

However, in practice, many of the $P^{\star}$ classes, as expressed in (2.3.52), are empty. This, all the more, as $\mathcal{O}$ is classifiable with respect to the distance, which is provided (see Sect. 8.7 of Chap. 8). The classes of $P^{\star}$ are called "strong patterns" ("formes fortes").

Note here that $Q^{\star}$ is the partition of $\mathcal{O}$ associated with the equivalence relation defined by the transitive closure of the binary relation on $\mathcal{O}$

$$Q = P^{1} \vee P^{2} \vee \cdots P^{j} \vee \cdots \vee P^{J} \tag{2.3.54}$$

where P^{j} indicates here the equivalence relation on $\mathcal{O}$, defined by the partition P^{j} of $\mathcal{O}$, $1 \leq j \leq J$

The classes of the partition Q are called "weak patterns" (formes faibles).

Facing to repetition of the dynamic cluster algorithm process, from a sequence of kernel systems (see (2.3.45)), strong and weak patterns enable the interpretation of the results (see (2.3.47)) to be ordered.

Statistical Significance of Strong Patterns

The strong patterns (formes fortes) are the more interesting to be handled and interpreted. There are two ways to observe their respective validities. The first one is static and can be expressed after J iterations of the algorithmic process (see (2.3.45)–(2.3.47)). The second one is recursive and is defined after each iteration of the algorithmic process.

For the first approach, it comes to measure the internal agreement of the different partitions of (2.3.47). The most natural is to refer mutual independence statistical hypothesis between the partitions observed at the end of the algorithmic process. In this context, the most classical measure is provided by the chi-square statistic [20] (see (6.2.43) of Sect. 6.2.3.5 of Chap. 6). In order to express this statistic here, let us introduce, relative to the partition P^j (see (2.3.50)), the parameters

$$(\forall k),\ 1 \leq k \leq K,\ 1 \leq j \leq J,\ n_k^j = card(O_k^j) \text{ and } p_k^j = \frac{n_k^j}{n} \tag{2.3.55}$$

On the other hand, relative to the partition $P^\star$ (composed of strong patterns), introduce its class cardinals, that is,

$$n_{k_1 k_2 \ldots k_j \ldots k_J} = card\left(\bigcap_{1 \leq j \leq J} O_{k_j}\right) \tag{2.3.56}$$

In these conditions, the chi-square statistic χ^2 can be written as follows:

$$\chi^2 = \sum_{(k_1, k_2, \ldots, k_j, \ldots, k_J)} \frac{\left(n_{k_1 k_2 \ldots k_j \ldots k_J} - n p_{k_1} \times p_{k_2} \times \cdots \times p_{k_j} \times \cdots \times p_{k_J}\right)^2}{n p_{k_1} \times p_{k_2} \times \cdots \times p_{k_j} \times \cdots \times p_{k_J}} \tag{2.3.57}$$

where p_{k_j} is what we have denoted above by p_k^j (see (2.3.55)).

The independence hypothesis against to which the chi-square statistic (2.3.57) is established is far from the real nature of the problem in question. In fact, whatever is the relative arrangement of the respective points representing $\mathcal{O}$ in $\mathbb{R}^p$, the different partitions P^j, $1 \leq j \leq J$, are necessarily, near statistically, because they are obtained by the same algorithm, processed on the same set, endowed with a fixed distance.

In fact, the matter consists of studying the distribution of (2.3.57) in the case of a random class of n points in $\mathbb{R}^p$. The randomness might be defined by a multidimensional distribution (e.g. Gaussian distribution) adequately associated with the empirical distribution of the numerical attributes describing $\mathcal{O}$. For simplicity reasons, the latter distribution can be defined as the product of p independent Gaussian probability laws, where the jth one is associated with the empirical distribution of the jth descriptive attribute, $1 \leq j \leq p$. In these conditions, the mean and variance of the jth probability law might be chosen as identical to those of the jth empirical distribution.

The analytical complexity for establishing the probability law of (2.3.57) under the probabilistic model we have just described leads us to propose to proceed by simulation. R independent random realizations of the cloud of n points associated with the observed one are carried out. For each realization, J independent repetitions of the dynamic cluster algorithm, from J kernel systems as (2.3.46) above, are performed and then, the chi-square statistic (2.3.57) is calculated. The sequence of the R values of χ^2, so obtained, enables the distribution of (2.3.57) to be observed under an adequate random model of no relation between the data units. Consequently, the observed value of (2.3.57) in a real case can be validated.

Notice that the coefficients described in Sect. 6.2.3.5 of Chap. 6 show that other indices than the χ^2 might be considered. For this, a certain research turn out to be needed.

As mentioned above, a second method, which permits the validity or prehension of the strong patterns to be established, has a sequential nature. It takes place after each iteration of the algorithm. More precisely, imagine the partition $P^\star$ (see (2.3.48) and (2.3.50)) get after J iterations and let

$$P = \{O_k | 1 \leq k \leq K\} \tag{2.3.58}$$

be the partition obtained after a new iteration (the $(J+1)$th one). If C is a class of $P^\star$, the method evaluates how C is questioned by P.

For this evaluation—as suggested by Diday—the Shannon entropy can be used. It is written as follows:

$$\mathcal{H}(P/C) = -\sum_{1 \leq k \leq K} p(k/h) \log_K (p(k/h)) \tag{2.3.59}$$

where k and h stand for O_k and C and where

$$p(k/h) = \frac{card(O_k \cap C)}{card(C)} \tag{2.3.60}$$

$1 \leq k \leq K$.

A value 1 for one of the conditional probabilities (proportions) (2.3.60) means that the class C is included in one of the classes O_k. In this case, the class C is not questioned by the partition P and the value of $\mathcal{H}(P/C)$ is null. The opposite case is that for which C is equally distributed on the different classes O_k, $1 \leq k \leq K$. In this situation, each of the rations (2.3.60) is equal to $1/K$ and $\mathcal{H}(P/C) = 1$.

Consequently, the measure (2.3.59) permits us to evaluate the concentration of the class C with respect to the different classes of the partition P.

Now, to integrate the respective contributions of all the $P^\star$ classes, the following coefficient has to be adopted according to (2.3.59):

$$\mathcal{H}(P/P^\star) = -\sum_{1 \leq h \leq H} \sum_{1 \leq k \leq K} p(k/h) \log_K (p(k/h)) \tag{2.3.61}$$

Other coefficients than (2.3.59) can be used for the same purpose. The most known is the Gini index. It can be written as follows:

$$\phi(P/C) = 1 - \sum_{1 \le k \le K} p(k/h)^2 \tag{2.3.62}$$

Its value is comprised between 0 and $(1 - (1/K))$: 0 in the case where C is contained in one of the classes O_k and $(1 - (1/K))$ in the case where C is uniformly distributed among the different classes O_k, $1 \le k \le K$.

Consensus by a Partition or a Partition Chain

In terms of methodology, a few things have to be added to what was expressed at the end of Sect. 2.2.5. Nevertheless, the context here is very different. Let

$$(P^1, P^2, \ldots, P^j, \ldots, P^J) \tag{2.3.63}$$

be the sequence of partitions obtained by the dynamic cluster algorithm from the kernel system (2.3.45). In (2.3.63), P^j is the partition obtained from L^j at the jth iteration of the algorithmic process, $1 \le j \le J$.

Every partition P^j can be assimilated to a partition induced by a nominal categorical attribute on $\mathcal{O}$, with K_j values, $1 \le j \le J$. The specificity of this interpretation is that these attributes are statistically very close. In fact, the different partitions P^j of $\mathcal{O}$, $1 \le j \le J$, are obtained by applying the *same algorithm*, but with *different initial states*. These partitions are all the more near that the set $\mathcal{O}$ endowed with its distance function is classifiable (see Sect. 8.7 of Chap. 8).

In these conditions, every clustering method of a set of objects described by nominal categorical attributes applies, in particular, the partition central method, which is conceived for this type of description (see Sect. 2.2). Also, we have point out at the end of Sect. 2.2.5 the usage of ascendant agglomerative hierarchical clustering, with the similarity index defined in Sect. 10.2.1 of Chap. 10.

In fact, following the definition of a similarity or dissimilarity index between objects described by categorical nominal attributes, every clustering algorithm which can handle the similarity or dissimilarity index concerned is able to built a consensus between the different partitions P^j, $1 \le j \le J$. Clearly, the consensus result is strongly dependent on the method used.

Notice that in our particular situation (see (2.3.63)), whatever the similarity (resp., dissimilarity) index chosen, the similarity (resp., dissimilarity) function is maximal (resp., minimal) for any pair of objects belonging to the same strong pattern. Thus, the strong patterns belonging to $P^\star$ (see (2.3.52)) will appear as subclasses of larger classes. It is interesting to observe how the strong patterns are organized for a given clustering consensus method.

2.3.4.2 Some Important Developments of the *K*-Means or Dynamic Cluster Methods

Due to the geometrical nature of the notions handled in these adaptative reallocating-recentring algorithms, the clearest expression of these concerns clustering a cloud of points in an Euclidean space. However, their flexibility, combined with the additive form of the adequation criterion (see (2.3.30)), make that they apply to a large extent of data structures and cluster representations [12]. In [14] each cluster is characterized by a factorial plan corresponding to a local component analysis [21, 25], Chap. 6.

In another statistical situation [15] each cluster is defined by a random model represented by a multidimensional distribution. The latter, denoted by F_k, is supposed characterized by a vector parameter θ_k, $1 \leq k \leq K$. The whole cloud in $\mathbb{R}^p$ associated with the representation of the set of objects $\mathcal{O}$ is seen as the realization of a probabilistic mixing of K distributions. It can be set in the form

$$F = \sum_{1 \leq k \leq K} p_k F_k \tag{2.3.64}$$

where p_k is the probability of the F_k occurrence, $1 \leq k \leq K$

$$\sum_{1 \leq k \leq K} p_k = 1$$

In this approach, the kernel system becomes

$$(\theta_1, \theta_2, \ldots, \theta_k, \ldots, \theta_K) \tag{2.3.65}$$

The criterion maximized is the likelihood logarithm of the observed sample $\{x_i | 1 \leq i \leq n\}$, where x_i represents the ith object in $\mathbb{R}^p$. This criterion comes down to the additive form (2.3.30) [16, 38].

This formalization of the clustering problem is the same as that of *Expectation-Maximisation* (*EM*) algorithm [5, 9, 32, 39]. Relative to the general principle, *E* and *M*, respectively, correspond to the allocation and centring steps of the general algorithm type considered in this section.

The development of the *EM* algorithm is more global than that resulting from the adaptation of dynamic cluster algorithm [16]. The same general form of the distributions F_k, $1 \leq k \leq K$, is assumed in the *EM* algorithm. Mostly, this form is Gaussian. In [6] *EM* algorithm is proved to be equivalent to the K-means algorithm—with the inertia criterion—under a spherical Gaussian mixture.

Let us return now to the initial representation of a class O_k of a partition $P = \{O_k | 1 \leq k \leq K\}$ of an object set $\mathcal{O}$ by a subset e of $\mathcal{O}$ (see Sect. 2.3.3.1). As already indicated, the most vulnerable feature of the dynamic cluster algorithm concerns the initial choice of the kernel system L^0 (see Sect. 2.3.3.2). For simplicity reasons and also because it is mostly practiced, we have supposed above a common size, denoted by m, for the different kernels. Generally, the order of m is a few units

($m = 1, 2$ or 3). If we accept the representation of each of the P classes by one of its elements, the "Attraction poles" method [22, 24, 27] and [25], Chap. 8, provides an objective statistical technique in order to determine a system of attraction poles, playing efficiently the role of an initial kernel system. This determination method is based on a variance analysis of the mutual distances between the elements of the object set $\mathcal{O}$. More precisely, suppose that we want a system of K attraction poles. If $\{e_1, e_2, \ldots, e_{k-1}\}$ designates the set of the $k-1$ first poles extracted, $k \leq K$, the kth pole e_k is determined from a statistical coefficient combining two conditions:

1. The variance of the distances (or proximities) of e_k to the other elements of $\mathcal{O}$ is as large as possible;
2. e_k is as far as possible from $\{e_1, e_2, \ldots, e_{k-1}\}$.

The Pole attraction method was first conceived as a seriation method [24]. Afterwards, it was extended and gave rise to a large family of clustering methods [22, 27] and Chap. 8 of [25]. Let us precise briefly the general principle of these methods. An increasing sequence of the number K of classes is considered: $K = 2, 3, \ldots, l, \ldots, \mathbb{L}$. For a given value l of K, we determine a Pole attraction system $\{e_1, e_2, \ldots, e_l\}$, comprising l elements. Then, each of the objects of $\mathcal{O}$ is assigned to the (or one of the) nearest attraction pole among $e_1, e_2, \ldots, e_l$, creating thus a partition of $\mathcal{O}$ into exactly l clusters. In this way, e_i is naturally included in the ith cluster, $1 \leq i \leq l$.

Otherwise and importantly, this partition is evaluated according to the adequacy criterion studied in Sects. 9.3.5 and 9.3.6 of Chap. 9. Thereby, only some values of l are retained as corresponding to "natural" numbers of classes. Therefore, this method does not require to fix beforehand a number of classes. A top-down but *non-hierarchical* sequence of partitions is collected, each being accompanied by a high value of the adequacy criterion.

Determining an initial kernel system as the pole attraction system effectively contributes to any of the K-means or dynamic cluster algorithms. Experimentally, we observe a very quick convergence of these algorithms when the starting point is a system of attraction poles. On the other hand, theoretically, the attraction poles showed all interest in a continuous model of perfect classifiability [26].

This model is defined by a sequence of two consecutive intervals of an horizontal axis, I_1 and I_2, uniformely weighted with a density equal to 1, separated by an empty interval I. The respective lengths of I_1, I_2 and I are the parameters of the analysis of the K-means convergence algorithm studied in [26]. This analysis was discussed with great interest in [4].

The nature of the convergence problem considered in [34] is very different from that we have just mentioned. In [34] it is not a question of the convergence of the K-means algorithm towards an optimal solution. The matter is rather the convergence of an optimal solution of the K-means algorithm calculated for a random sample of infinite population towards the optimal solution of this algorithm, applied to the entire population. Therefore, the latter problem has a classical nature in terms of *Mathematical Statistics*. More precisely, consider the application of the K-means algorithm to the geometrical space $\mathbb{R}^p$, endowed with a probability measure P',

such that the algorithm converges to a unique optimal solution defined by a system $L = (g_1, g_2, \ldots, g_k, \ldots, g_K)$ of centres of gravity. Now, consider a random sample $\mathcal{O}^{(n)}$ of size n provided from the probabilized space $\mathbb{R}^p$ and let us designate by $L^{(n)} = (g_1^{(n)}, g_2^{(n)}, \ldots, g_k^{(n)}, \ldots, g_K^{(n)})$ the unique optimal solution of the K-means algorithm applied on $\mathcal{O}^{(n)}$, the Pollard theorem specifies the conditions under which $L^{(n)}$ converges almost surely to L. This convergence is point by point, and it implies associated labellings between the elements of $L^{(n)}$ and those of L.

References

1. Anderberg, M.R.: Cluster Analysis for Applications. Academic Press, New York (1973)
2. Bock, H.-H.: On some significance tests in cluster analysis. J. Classif. **2**, 77–108 (1985)
3. Bock, H.-H.: Clustering methods: a history of k-means algorithms. In: Cucumel, G., Brito, P., Bertrand, P., de Carvalho, F. (eds.) Selected Contributions in Data Analysis and Classification, pp. 161–172. Springer, Berlin (2007)
4. Celeux, G.: Étude exhaustive de l'algorithme de réallocation-recentrage dans un cas simple. R.A.I.R.O. Recherche opérationnelle/Operations Research **20**(3), 229–243 (1986)
5. Celeux, G.: Reconnaissance de mélanges de densités de probabilité et applications à la validation des résultats en classification, thèse de doctorat ès sciences. Ph.D. thesis, Université de Paris IX Dauphine, September 1987
6. Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. **14**, 315–332 (1992)
7. Charon, I., Denoeud, L., Guénoche, A., Hudry, O.: Maximum transfer distance between partitions. J. Classif. **23**, 103–121 (2006)
8. Charon, I., Denoeud, L., Hudry, O.: Maximum de la distance de transfert à une partition donnée. Revue Mathématiques et Sciences **179**, 45–83 (2007)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). JRSS **B. 39**, 1–38 (1977)
10. Diday, E.: Une nouvelle méthode de classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. Revue de Statistique Appliquée **XIX**(2), 19–33 (1971)
11. Diday, E.: The dynamic clusters method in nonhierarchical clustering. J. Comput. Inf. Sci. **2**(1), 61–88 (1973)
12. Diday, E. and Collaborators: Optimisation en Classification Automatique, vol. I, II. Institut National de Recherche en Informatique et en Automatique (INRIA), Le Chesnay (1979)
13. Diday, E., Govaert, G.: Classification automatique avec distances adaptatives. R.A.I.R.O. Inf. Comput. Sci. **11**(4), 329–349 (1977)
14. Diday, E., Schroeder, A.: The dynamic clusters method in pattern recognition. In: Rosenfeld, J.L. (ed.) Information Processing 74, pp. 691–697. North Holland, Amsterdam (1974)
15. Diday, E., Schroeder, A.: A new approach in mixed distribution detection. Research Report 52, INRIA, January 1974
16. Diday, E., Schroeder, A.: A new approach in mixed distribution detection. R.A.I.R.O. Recherche Opérationnelle **10**(6), 75–1060 (1976)
17. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. In: Biometrics, Biometric Society Meeting, vol. 21, p. 768 (1965)
18. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognit. Lett. **31**, 651–666 (2010)
19. Jancey, R.C.: Multidimensional group analysis. Aust. J. Bot. **14**, 127–130 (1966)
20. Lancaster, H.O.: The Chi-Squared Distribution. Wiley, New York (1969)
21. Lebart, L., Fenelon, J.-P.: Statistique et Informatique appliquées. Dunod, Paris (1973)

22. Leredde. H.: La méthode des pôles d'attraction; la méthode des pôles d'aggrégation: deux nouvelles familles d'algorithmes en classification automatique et sériation. Ph.D. thesis, Université de Paris, 6 Oct 1979
23. Lerman, I.C.: Les bases de la classification automatique. Gauthier-Villars, Paris (1970)
24. Lerman, I.C.: Analyse du phénomène de la sériation. Revue Mathématiques et Sciences Humaines **38** (1972)
25. Lerman, I.C.: Classification et analyse ordinale des données. Dunod, Paris. http://www.brclasssoc.org.uk/books/index.html (1981)
26. Lerman, I.C.: Convergence optimale de l'algorithme de "réallocation-recentrage" dans le cas continu le plus simple. R.A.I.R.O. Recherche opérationnelle/Operations Research **20**(1), 19–50 (1986)
27. Lerman, I.C., Leredde, H.: La méthode des pôles d'attraction. In: IRIA (ed.): Analyse des Données et Informatique. IRIA (1977)
28. Loève, M.: Probability Theory. D. Van Nostrand Company, New Jersey (1963)
29. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
30. Marcotorchino, F.: Essai de Typologie Structurelle des Indices de Similarité Vectorielles par Unification Relationnelle. In: Benani, Y., Viennet, E. (eds.) RNTI A3, Revue des Nouvelles Technologies de l'Information, pp. 203–319. Cepaduès, Toulouse (2009)
31. Marcotorchino, F., Michaud, P.: Agrégation de similarités en classification automatique. Revue de Statistique Appliquée **2**, 21–44 (1982)
32. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
33. Neveu, J.: Bases Mathématiques du Calcul des Probabilités. Masson, Paris (1964)
34. Pollard, D.: Strong consistency of *k*-means algorithm. Ann. Stat. **9**(1), 135–140 (1981)
35. Régnier, S.: Sur quelques aspects mathématiques des problèmes de la classification automatique. I.C.C. Bull. **4**, 175–191 (1965)
36. Régnier, S.: Sur quelques aspects mathématiques des problèmes de la classification automatique. Revue Mathématiques et Sciences Humaines **22**, 13–29 (1983)
37. Régnier, S.: Sur quelques aspects mathématiques des problèmes de la classification automatique. Revue Mathématiques et Sciences Humaines **22**, 31–44 (1983)
38. Sclove, S.L.: Population mixture models and clustering algorithms. In: Commun. Stat. Theory Methods **A6**, 417–434 (1977)
39. Xu, R., Wunsh, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)

http://www.springer.com/978-1-4471-6791-4

Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering

Lerman, I.C.

2016, XXIV, 647 p. 54 illus., Hardcover
ISBN: 978-1-4471-6791-4