RAIRO. RECHERCHE OPÉRATIONNELLE

GILLES CELEUX

Étude exhaustive de l'algorithme de réallocation-recentrage dans un cas simple

RAIRO. Recherche opérationnelle, tome 20, n° 3 (1986), p. 229-243

<http://www.numdam.org/item?id=RO_1986__20_3_229_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 20, n° 3, août 1986, p. 229 à 243)

ÉTUDE EXHAUSTIVE DE L'ALGORITHME DE REALLOCATION-RECENTRAGE DANS UN CAS SIMPLE (*)

par Gilles CELEUX (1)

Résumé. — *Dans le but de préciser les limites d'application de l'algorithme de « reallocation-recentrage » ou des « centres mobiles », nous reprenons une étude récente de Lerman concernant la recherche des conditions de convergence de cet algorithme vers la solution optimale dans le cas où la structure à découvrir est constituée de deux intervalles disjoints de la droite réelle. Nous résolvons complètement ce problème et en tirons des conclusions utiles pour la pratique de cet algorithme.*

Mots-clés : Classification non hiérarchique; optimum global.

Abstract. — *In order to state precisely the application's bounds of the "k-means" algorithm, we return to a Lerman's recent study about convergence conditions of this algorithm in the simple situation of two no-overlapping intervals on the real line. We quite resolve this problem and give some useful comments for the practice of this algorithm.*

Keywords: Non-hierarchical classification; global optimum.

1. INTRODUCTION

Depuis quelques années, un nombre croissant de publications sur les méthodes d'analyse des données concernent la stabilité des résultats de techniques largement répandues.

Ces études témoignent d'une orientation nouvelle de la recherche en analyse des données. Après le développement rapide des techniques d'analyse des données et de reconnaissance des formes durant les années soixante et soixante-dix, de plus en plus de chercheurs ressentent le besoin de préciser les fondements théoriques et les domaines d'application des techniques proposées aux utilisateurs, de mieux cerner la stabilité et la validité des résultats afin de rendre l'utilisation des méthodes d'analyse des données plus efficace et même dans certains cas plus pertinente. Ce souci explique le grand succès des techniques de

(*) Reçu en février 1986.
(1) INRIA, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/86/03 229 15 $ 3.50,

reéchantillonnage (*cf.* [Efr82]) qui visent à mesurer la précision des résultats d'analyses statistiques.

Cette orientation de la recherche nous apparaît cruciale en classification, où les méthodes proposées abondent sans que leurs domaines d'application soient toujours clairement et précisément définis.

On trouvera une large bibliographie sur les problèmes de validité en classification automatique dans [Per83], nous voulons également mentionner l'article de Bock (Boc85] sur les tests en classification et les articles de Pollard ([Pol81]), [Pol82]) et Lemaire ([Lem83a], [Lem83b]) sur le comportement asymptotique des méthodes usuelles de partitionnement.

C'est dans ce cadre de validation des méthodes de classification que Lerman [Ler86] a étudié le comportement de l'algorithme de reallocation-recentrage, connu également sous le nom d'algorithme des centres mobiles et qui est la version la plus simple de l'algorithme des nuées dynamiques (*cf.* [Did80]), dans le cas d'une structure en classes particulièrement simple à savoir deux intervalles disjoints de la droite réelle.

Lerman exprime ainsi l'intérêt d'une telle étude :

« On ne peut admettre (comme il arrive parfois) de voir poser une axiomatique d'une classification, sans que cette parfaite partition en deux classes satisfasse une telle axiomatique. Il s'agit-là en effet d'une question de cohérence logique ».

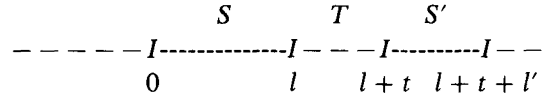
Nous ajouterons qu'une telle étude est particulièrement intéressante pour l'algorithme de reallocation-recentrage qui est de loin l'algorithme de partitionnement le plus utilisé, sans que les praticiens se soucient toujours de ses limites d'application.

Lerman exhibe des conditions suffisantes de convergence de cet algorithme vers la classification naturelle en deux intervalles disjoints. Mais ces conditions ne rendent compte que partiellement du comportement de l'algorithme de reallocation-recentrage dans ce cas simple (*cf.* paragraphe 3).

Dans cet article, nous résolvons complètement le problème posé par Lerman et nous faisons dans le dernier paragraphe des commentaires utiles pour le praticien des méthodes de classification.

2. LE CADRE DE L'ÉTUDE

Soient sur la droite réelle deux intervalles bornés S et S' disjoints et soit T l'intervalle entre S et S'.



On posera $S = [0, l]$ (l : longueur de S),

$T =]l, l + t[$ (t : longueur de T),

$S' = [l + t, l + t + l']$ (l' : longueur de S').

On supposera dans la suite, sans perte de généralité, que $l \geqslant l'$. Enfin, on notera $E = S \cup S'$ et $C = S \cup T \cup S'$.

Le but de l'étude est d'examiner la capacité de l'algorithme de reallocation-recentrage à retrouver cette classification en deux classes S et S'.

Pour ce faire, nous utilisons une généralisation naturelle de cet algorithme, bien connu dans le cas discret, au cas continu (*cf.* [Ler86], [Lec74]). Cette extension diffère de celle plus générale présentée dans [DMS77] et commentée dans [Ler86].

Pour ne pas alourdir inutilement l'exposé, nous présentons cette généralisation sur la droite réelle munie de la mesure de Lebesgue, la distance considérée étant la distance euclidienne usuelle.

L'algorithme de reallocation-recentrage continu recherche une partition $P = (P1, ..., Pk)$ en k classes (k spécifié) d'un ensemble E de $\mathbb{R}$ minimisant l'inertie intra-classe associée à la partition P par rapport à la mesure de Lebesgue.

$$W(P) = 1/\mu(E)\Sigma\left(\int_{Pj} (x - g_j)^2 \, dx, j = 1, k \right)$$

g_j étant le centre de gravité de la classe Pj :

$$g_j = \int_{Pj} x \, dx \Big/ \int_{Pj} dx \qquad \text{et} \qquad \mu(E) = \int_E dx.$$

Partant d'un système initial $g^0 = (g_1^0, ..., g_k^0)$ de centres, l'algorithme de reallocation-recentrage est un algorithme itératif utilisant alternativement deux fonctions :

– Une fonction d'affectation ou d'allocation G de $\mathbb{R}^k$ dans P_k (ensemble des partitions en k classes de E) définie par :

$$G(g_1, ..., g_k) = (P1, ..., Pk) \qquad \text{avec pour tout } j = 1, k$$
$$Pj = \{ x \in E / (x - g_j)^2 \leqslant (x - g_{j'})^2, \qquad \text{pour tout } j' \neq j \}$$

(avec j plus petit que j' en cas d'égalité).

– Une fonction de représentation ou de centrage F de P_k dans $\mathbb{R}^k$ définie par :

$$F(P1, ..., Pk) = (g_1, ..., g_k) \qquad \text{avec pour tout } j = 1, k$$
$$g_j = \int_{Pj} x \, dx \Big/ \int_{Pj} dx.$$

On montre (*cf.* [Lec74]) que :

— Cet algorithme fait décroître à chaque itération le critère d'inertie $W(P)$.

— La suite $g^n = (F \circ G)^n(g^0)$ converge dans R^k.

— La suite $u_n = W(P^n)$, P^n étant la partition associée au système de centres g^n, converge vers un minimum local du critère d'inertie.

Notons que, contrairement au cas discret, rien n'assure que la convergence se fait en un nombre fini d'itérations.

Le but de l'étude est le suivant :

Donner les conditions portant sur les quantités l, l', t et le système initial de centres (g, g'), le nombre de classes étant spécifié à deux, pour que l'algorithme de reallocation-recentrage défini ci-dessus converge vers la partition (S, S'), autrement dit que la suite des couples de centres générés par l'algorithme converge vers (s, s') avec $s = l/2$ centre de gravité de S et $s' = l + t + (l'/2)$ centre de gravité de S'.

3. LES PRINCIPAUX RÉSULTATS DE LERMAN [LER86]

Dans le cadre défini ci-dessus Lerman a démontré un théorème donnant des conditions suffisantes de convergence de l'algorithme de reallocation-recentrage vers la solution (S, S'). Il a par ailleurs démontré certaines propriétés adjacentes utiles pour la mise en évidence de son théorème.

Nous énonçons ci-dessous l'une de ces propriétés qui a un caractère fondamental puis le théorème assurant la convergence optimale de l'algorithme sous certaines conditions.

PROPRIÉTÉ. — Avec les notations et conventions du paragraphe 2, une condition nécessaire et suffisante pour que (s, s') soit un point fixe de l'algorithme de reallocation-recentrage est que t soit supérieur ou égal à $(l - l')/2$.

Cette propriété nous assure d'ores et déjà que, quel que soit le système initial des centres, l'algorithme de reallocation-recentrage ne peut conduire à la classification naturelle (S, S') si $t < (l - l')/2$.

Ce cas de figure se produit lorsque d'une part les longueurs l et l' des deux intervalles sont assez différentes et lorsque la distance t entre les deux intervalles est assez faible.

En fait, dans de tels cas, la propriété ci-dessus montre que le critère d'inertie intra-classe associée à la partition est inadéquat pour retrouver la partition

naturelle (S, S') pour la simple raison que cette partition ne minimise pas ce critère comme le montre l'exemple suivant :

$$l = 12, \qquad t = 4, \qquad l' = 2 \qquad (\text{on a } t < (l - l')/2).$$

Notons $S^* = (S, S')$:

$$W(S^*) = 1/14 \int_S (x - 6)^2 \, dx + 1/14 \int_{S'} (x - 17)^2 \, dx.$$

Ce qui donne $W(S^*) = 10.33$.

Soit maintenant la partition $P = (P1, P2)$ avec :

$$P1 = [0,10], \qquad P2 = [10,12] \cup [16,18]$$

$g_1 = 5$ est le centre de gravité de $P1$.

$g_2 = 14$ est le centre de gravité de $P2$.

On a :

$$W(P) = 1/14\left[\int_{P1} (x - 5)^2 \, dx + \int_{S-P1} (x - 14)^2 \, dx + \int_{S'} (x - 14)^2 \, dx\right]$$

d'où $W(P) = 8{,}62 < W(S^*)$.

THÉORÈME : *Avec les notations et conventions du paragraphe 2, si le système initial des centres est formé d'un couple de points appartenant respectivement aux intervalles S et S', si la longueur t de l'intervalle T séparant S et S' est telle que :* $(l - l')/2 \leqslant t < l'$ *et si* $l \leqslant 3l'$, *alors l'algorithme de reallocation-recentrage converge vers la solution optimale formée du couple de centres de gravité* (s, s'). *Cette solution est unique.*

Notons tout d'abord que ce théorème astreint le couple initial de centres à appartenir à $S \times S'$ ce qui ne recouvre pas tous les cas possibles de convergence optimale.

La condition $t \geqslant (l - l')/2$ découle directement de la propriété énoncée ci-dessus.

La condition $l \leqslant 3l'$ vient de ce que Lerman a imposé au système initial de centres d'appartenir à $S \times S'$.

La condition $t < l'$ est purement technique et limite la portée du théorème. Ainsi si $l = l'$, il découle de ce théorème que l'algorithme de reallocation-recentrage converge vers (s, s') si le système initial de centres appartient à $S \times S'$ et si $t < l$.

On verra en fait que dans ce cas l'algorithme converge vers la solution optimale pour tout t et quelle que soit la position initiale du système de centres dans $C \times C$.

4. CARACTÉRISATION DE L'ALGORITHME DE REALLOCATION-RECENTRAGE PAR UN OPÉRATEUR RÉEL H

On a vu au paragraphe 2, que l'algorithme de reallocation-recentrage était caractérisé par l'opérateur $F \circ G$ de $C \times C$ dans $C \times C$ qui à tout système de centres (g, g') associe un nouveau système de centres (h, h') par la formule : $(h, h') = F \circ G((g, g'))$.

PROPOSITION : (a) *Le couple* $(h, h') = F \circ G((g, g'))$ *ne dépend que de la position du point frontière* $m = (g + g')/2$ *entre les classes* $P1$ *et* $P2$ *définies par :*

$$P1 = \{ x \in E / (x - g)^2 \leqslant (x - g')^2 \} \qquad et \qquad P2 = E - P1$$

dont les centres de gravité sont respectivement h *et* h'.

(b) *De plus, le point frontière entre deux classes connexes* $P1$ *et* $P2$ *d'une partition* P *de* E *en deux classes détermine de manière unique la position des centres des deux classes* $P1$ *et* $P2$.

DÉMONSTRATION : D'après [Ler85] on a les formules (L) suivantes :

Si $0 \leqslant m \leqslant l$:

$$h = m/2$$
$$h' = [(l - m)(l + m) + l'(2l + 2t + l')]/2(l - m + l')$$

Si $l < m < l + t$:

$$h = s = l/2$$
$$h' = s' = (2l + 2t + l')/2.$$

Si $l + t \leqslant m \leqslant l + t + l'$:

$$h = [l^2 + (m - l - t)(m + l + t)]/2(m - t)$$
$$h' = (m + l + t + l')/2.$$

Le point (a) de la proposition se déduit immédiatement de ces formules.

Pour démontrer le point (b), il suffit de montrer que les formules (L) établissent une correspondance biunivoque entre tout couple (g, g') de $C \times C$ et tout point m de C.

D'après les formules (L), cette correspondance biunivoque existera, si les applications :

$$a : [0, l] \to C$$

définie par :

$$a(x) = [l^2 - x^2 + l'(2l + 2t + l')]/2(l + l' - x)$$

et :

$$b:[l+t,\ l+t+l'] \to C$$

définie par :

$$b(x) = [l^2 + x^2 - (l+t)^2]/2(x-t)$$

sont monotones.

Or un simple calcul montre que la dérivée $a'(x)$ de $a(x)$ est du signe du trinôme :

$$2x^2 - 4(l+l')x + 2(l+l')^2 + 4l't$$

et on vérifie facilement que ce trinôme est de signe constant. De même la dérivée $b'(x)$ de $b(x)$ est du signe du trinôme :

$$2x^2 - 4tx + 2(l+t)^2 + 4l^2$$

qui garde un signe constant.

Cette proposition nous montre que l'algorithme de reallocation-recentrage peut être caractérisé par l'opérateur H de C dans C qui a tout point frontière m de toute partition de E en deux classes connexes fait correspondre un nouveau point frontière $m' = H(m)$ entre les deux nouvelles classes connexes obtenues par l'algorithme.

La caractérisation de l'algorithme de reallocation-recentrage par cet opérateur réel va nous permettre de mener facilement l'étude projeté.

Par définition, pour tout m appartenant à C, $H(m) = m' = (h + h')/2$. Donc d'après les formules (L) ci-dessus, on a :

Si $0 \leqslant m \leqslant l$:

$$H(m) = [-2m^2 + (l+l')m + (l+l')^2 + 2tl']/[-4m + 4(l+l')].$$

Si $l < m < l + t$:

$$H(m) = (3l + 2t + l')/4.$$

Si $l + t \leqslant m \leqslant l + t + l'$:

$$H(m) = [2m^2 + (l+l')m - t(3l + 2t + l')]/[4m - 4t].$$

Cet opérateur H est continu en tout point de C du fait que :

$$H(l) = H(l+t) = (3l + 2t + l')/4.$$

Par ailleurs, l'étude de la dérivée de H sur chaque intervalle $[0, l]$ et $[l + t, l + t + l']$ montre facilement que H est une fonction croissante sur C.

5. ÉTUDE DES POINTS FIXES DE L'OPÉRATEUR H

Les points fixes de H vont évidemment nous permettre de caractériser les solutions possibles de l'algorithme de reallocation-recentrage dans les différents cas.

D'après la forme de H, nous sommes amenés à distinguer trois types de points fixes.

(*a*) Les points fixes compris entre 0 et l :

Ces éventuels points fixes sont solutions de l'équation :

$$[-2m^2 + (l + l')m + (l + l')^2 \pm 2tl']/[-4m + 4(l + l')] = m$$

équivalente à l'équation du second degré :

$$T(x) = 2m^2 - 3(l + l')m + (l + l')^2 + 2tl' = 0 \quad (T).$$

Le discriminant D de cette équation $D = (l + l')^2 - 16tl'$ est positif ou nul si :

$$t \leqslant (l + l')^2/16l'$$

les racines sont alors $m' = (3(l + l') + \sqrt{D})/4$ et $m'' = (3(l + l') - \sqrt{D})/4$ et sont positives.

Pour que ces racines soient des points fixes de H il faut de plus qu'elles soient majorées par l.

$m' \leqslant l \Leftrightarrow \sqrt{D} \leqslant l - 3l'$. Cette inégalité est toujours fausse si $l < 3l'$.

Si $l \geqslant 3l'$, $m' \leqslant l \Leftrightarrow D \leqslant (l - 3l')^2$. Ce qui après calcul est équivalent à $t \geqslant (l - l')/2$.

$m'' \leqslant l \Leftrightarrow (3l' - l) \leqslant \sqrt{D}$. Cette inégalité est toujours vraie si $l \geqslant 3l'$.

Si $l < 3l'$, $m'' \leqslant l \Leftrightarrow (3l' - l)^2 \leqslant D$ ce qui après calcul est équivalent à $t \leqslant (l - l')/2$.

En résumé :

H n'a pas de point fixe plus petit ou égal à l si $t > (l + l')^2/16l'$ ou si $(l - l')/2 \leqslant t \leqslant (l + l')^2/16l'$ et $l < 3l'$.

H a un seul point fixe plus petit ou égal à l si $t < (l - l')/2$. Ce point fixe est alors :

$$m'' = [3(l + l') - \sqrt{(l + l')^2 - 16tl'}]/4$$

H a deux points fixes plus petits ou égaux à l si :

$$(l - l')/2 \leqslant t \leqslant (l + l')^2/16l' \qquad \text{et} \qquad l \geqslant 3l'.$$

Ces points fixes sont alors :

$$m' = [3(l + l') + \sqrt{(l + l')^2 - 16tl'}]/4$$
$$m'' = [3(l + l') - \sqrt{(l + l')^2 - 16tl'}]/4.$$

(*b*) Les points fixes compris entre l et $l + t$:

Il y a au plus un point fixe compris entre l et $l + t$. Cet éventuel point fixe m^* de H est associé au couple optimal de centres (s, s') de la partition (S, S').

Pour que m^* existe, il faut et il suffit que $l \leqslant (3l + 2t + l')/4 \leqslant l + t$.

Il est facile de voir que ces deux inégalités sont vérifiées si et seulement si $t \geqslant (l - l')/2$.

On retrouve ainsi le fait que (s, s') est un point fixe de l'opérateur $F \circ G$ (*cf.* paragraphe 3) si et seulement si $t \geqslant (l - l')/2$ puisque (s, s') est définie de manière unique par le point fixe $m^* = (3l + 2t + l')/4$ de l'opérateur H.

(*c*) Les points fixes compris entre $l + t$ et $l + t + l'$:

Ces éventuels points fixes sont solutions de l'équation :

$$[2m^2 + (l + l')m - t(3l + 2t + l')]/[4m - 4t] = m$$

équivalente à l'équation du second degré :

$$T'(x) = -2m^2 + (l + l' + 4t)m - 3lt - 2t^2 - tl' = 0 \quad (T').$$

Le discriminant de cette équation s'écrit $D = (l + l')^2 - 16tl$.

A supposer que D soit positif ou nul les racines sont :

$$m1 = [(l + l' + 4t) + \sqrt{D}]/4$$

et :

$$m2 = [(l + l' + 4t) - \sqrt{D}]/4 \qquad \text{avec} \qquad m2 \leqslant m1.$$

Or :

$$m1 = [(l + l' + 4t) + \sqrt{(l + l')^2 - 16tl}]/4 < [2(l + l') + 4t]/4$$

comme $l' \leqslant l$, il vient $m1 < (4l + 4t)/4)/4 = l + t$.

Donc, l'opérateur H ne possède pas de point fixe plus grand ou égal à $l + t$.

Finalement :

– Si $t > (l + l')^2/16tl'$ ou si $(l - l')/2 \leqslant t \leqslant (l + l')^2/16l'$ et $l < 3l'$, H a pour seul point fixe m^*.

– Si $t < (l - l')/2$, H a pour seul point fixe m''.

– Si $l \geqslant 3l'$ et si $t = (l + l')^2/16l'$, H a deux points fixes distincts m' et m^* car alors $m'' = m'$.

— Si $l \geqslant 3l'$ et si $(l - l')/2 < t < (l + l')^2/16l'$, H a trois points fixes distincts m'' , m', m^*.

— Si $l \geqslant 3l'$ et si $t = (l - l')/2$, H a deux points fixes distincts m'' et m^* car alors $m' = m^* = l$.

6. CONVERGENCE DE L'ALGORITHME DE REALLOCATION-RECENTRAGE

Nous avons maintenant tous les éléments pour énoncer un théorème explicitant vers quelles solutions l'algorithme de reallocation-recentrage converge dans toutes les configurations, portant à la fois sur la position du système initial de centres et les grandeurs l, t et l', pour le cas simple étudié ici.

Pour simplifier l'énoncé de ce théorème, en plus des notations et conventions du paragraphe 2, on utilisera les notations suivantes :

$$m^* = (3l + 2t + l')/4$$
$$m' = [3(l + l') + \sqrt{(l + l')^2 - 16tl'}]/4$$
$$m'' = [3(l + l') - \sqrt{(l + l')^2 - 16tl'}]/4$$
$$f(m) = [(l - m)(l + m) + l'(2l + 2t + l')]/2(l - m + l').$$

THÉORÈME : *Avec les notations et conventions définies ci-dessus, les différentes valeurs possibles du couple de centres (g^*, g'^*) obtenu à la convergence de l'algorithme de reallocation-recentrage sont les suivantes :*

— *Si $t > (l + l')^2/16l'$ ou si $(l - l')/2 \leqslant t \leqslant (l + l')^2/16l'$ et $l < 3l'$, quelle que soit la position du couple initial de centres dans $C \times C$, $(g^*, g'^*) = (s, s')$.*

— *Si $t < (l - l')/2$, quelle que soit la position du couple initial de centres dans $C \times C$, $(g^*, g'^*) = (m''/2, f(m''))$.*

Dans tous les autres cas, (g^*, g'^*) dépend de la position du couple initial (g, g') ou plus précisément de la position du milieu m de g et de g'.

— Si $l \geqslant 3l'$ et si $t = (l + l')^2/16l'$ alors :

$$(g^*, g'^*) = (s, s') \qquad \text{si} \qquad m > m' = m''$$

et :

$$(g^*, g'^*) = (m'/2, f(m')) \qquad \text{si} \qquad m \leqslant m'.$$

— Si $l \geqslant 3l'$ et si $(l - l')/2 < t < (l + l')^2/16l'$,

$$(g^*, g'^*) = (s, s') \qquad \text{si} \qquad m > m',$$
$$(g^*, g'^*) = (m'/2, f(m')) \qquad \text{si} \qquad m = m',$$

et :

$$(g^*, g'^*) = (m''/2, f(m'')) \qquad \text{si} \qquad m < m'.$$

– Si $l \geqslant 3l'$ et si $t = (l - l')/2$, alors :

$$(g^*, g'^*) = (s, s') \qquad \text{si} \qquad m \geqslant m^* = m'$$

et :

$$(g^*, g'^*) = (m''/2, f(m'')) \qquad \text{si} \qquad m < m^*.$$

REMARQUE : Nous limitons le couple initial de centres à $C \times C$ et non à $\mathbb{R}^2$. En effet, dans tous les problèmes de classification les centres d'attraction ne sont jamais pour des raisons évidentes pris en dehors de l'enveloppe convexe du nuage de points à classer. De ce fait, considérer que l'un des centres initiaux n'appartient pas à C compliquerait inutilement l'énoncé du théorème, du fait de la possibilité d'apparition de solutions dégénérées (disparition d'une des deux classes dans le cas où le milieu des deux centres initiaux serait à l'extérieur de C) qui ne présentent pas d'intérêt.

DÉMONSTRATION : La proposition du paragraphe 4 montre que l'étude de la suite des couples de centres obtenue par l'algorithme de reallocation-recentrage peut être ramenée à l'étude de la suite des points-milieux de ces deux centres.

Nous allons donc, dans les différentes cas considérés, étudier la convergence de la suite récurrente (m_n) définie par :

$$m_0 = (g + g')/2 \qquad \text{et} \qquad m_{n+1} = H(m_n).$$

D'après les formules (L) du paragraphe 4, la convergence du couple de centres vers (s, s') est équivalente à la convergence de (m_n) vers m^* et la convergence du couple des centres vers $(m/2, f(m))$ est équivalente à la convergence de (m_n) vers m.

– cas 1 : $t > (l + l')^2/16l'$ ou $[(l - l')/2 \leqslant t \leqslant (l + l')^2/16l'$ et $l < 3l']$

Dans ce cas, la fonction H étant continue, la limite si elle existe de la suite récurrente (m_n) est nécessairement l'unique point fixe m^* de H (voir paragraphe 5). Il reste à montrer que la suite (m_n) converge quelle que soit la valeur de m_0.

Si $m_0 \leqslant m^*$ alors $H(m) \leqslant H(m^*) = m^*$ car H est croissante. Donc pour tout n, $m_n \leqslant m^*$. Il s'en suit que la suite (m_n) est croissante car $H(m) \geqslant m$ pour $m \leqslant m^*$. En effet :

Si $m \leqslant l$, il est facile de voir que $H(m) - m$ est du signe du trinôme du second degré $T(x)$ introduit dans l'équation (T) au paragraphe 5 qui dans le cas considéré n'a pas de racine ou a des racines plus grandes que l et est positif pour $m < l$.

Si $l < m \leqslant m^*$ on a $H(m) = m^*$ d'où $H(m) - m \geqslant H(m) - m^* = 0$.

Finalement la suite (m_n) croissante et majorée converge.

Maintenant si $m_0 > m^*$, on a $H(m) \geqslant H(m^*) = m^*$ et donc pour tout n, $m_n \geqslant m^*$. Il s'en suit que la suite (m_n) est décroissante car $H(m) \leqslant m$ pour $m \geqslant m^*$.

En effet si $m \geqslant l + t$, il est facile de voir que $H(m) - m$ est du signe du trinôme du second degré $T'(x)$ introduit dans l'équation (T') au paragraphe 5 et ce trinôme n'a pas de racine ou a des racines toujours plus petites que $l + t$ et est négatif pour $m > l + t$.

Si $m^* \leqslant m < l + t$, on a $H(m) = m^*$ d'où $H(m) - m \leqslant H(m) - m^* = 0$.

Finalement la suite (m_n) décroissante et minorée converge.

– Cas 2 : $t < (l - l')/2$.

Dans ce cas, H a pour unique point fixe m'' strictement plus petit que m^*. Pour montrer que la suite (m_n) converge vers m'' il suffit de recopier la démonstration pour le cas 1, en remplaçant partout m^* par m''.

– Cas 3 : $t = (l + l')^2/16l'$ et $l \geqslant 3l'$.

Dans ce cas, H a deux points fixes distincts $m'(= m'')$ et m^* avec $m' < m^*$. La suite (m_n) a donc deux limites possibles m' et m^*.

Si $m_0 \leqslant m'$, on voit simplement de manière analogue au cas 1 que la suite (m_n) est majorée par m' et qu'elle est croissante car dans ce cas le trinôme $T(x)$ est un carré parfait. Donc (m_n) converge vers m' si $m_0 \leqslant m'$.

Si $m' < m_0 \leqslant m^*$, on voit pareillement que la suite (m_n) est majorée par m^* et est croissante. Elle converge donc vers m^*.

Enfin si $m_0 > m^*$, il suffit de recopier la partie correspondante à la même hypothèse dans le cas 1 pour montrer que (m_n) converge alors vers m^*.

– Cas 4 : $(l - l')/2 < t < (l + l')^2/16l'$ et $l \geqslant 3l'$.

Dans ce cas H admet trois points fixes distincts m'', m' et m^* avec $m'' < m' < m^*$. Ces trois points fixes sont les seules limites possibles de la suite (m_n).

Or dans ce cas le trinôme $T(x)$ a pour racines m'' et m'. Il est positif pour $m < m''$ ou pour $m > m'$ et négatif entre m'' et m'. Il s'en suit aisément que :

Si $m_0 \leqslant m''$, (m_n) est une suite croissante majorée par m'' et donc converge vers m''.

Si $m'' < m_0 < m'$, (m_n) est une suite décroissante minorée par m'' et donc converge vers m''.

Si $m_0 = m'$, pour tout n, $m_n = m'$ et donc (m_n) converge vers m'.

Si $m_0 > m'$, l'étude se mène exactement de la même manière que pour le cas 1 et la suite (m_n) converge vers m^*.

Remarquons que le point m' n'est pas un point fixe attractif de l'opérateur H et qu'en conséquence le couple de centres $(m'/2, f(m'))$ est un optimum instable de l'algorithme de reallocation-recentrage.

– cas 5 : $t = (l - l')/2$ et $l \geqslant 3l'$.

Dans ce cas, H a deux points fixes distincts m'' et m^* avec $m'' < m^* = m'$. Ce cas se traite de la même façon que le précédent, la seule nuance résidant dans le fait que $m' = m^*$.

D'où si $m < m^*$, la suite (m_n) converge vers m''.

D'où si $m \geqslant m^*$, la suite (m_n) converge vers m^*.

7. COMMENTAIRES SUR LE THÉORÈME DE CONVERGENCE

Si la longueur t de l'intervalle T est suffisamment grande, l'algorithme de reallocation-recentrage permet de trouver la classification naturelle en deux classes S et S'. Ce résultat est intuitivement évident. Le suivant est moins trivial.

Plus l'écart entre les longueurs l et l' des intervalles S et S' est grand, plus t doit être grand pour assurer la reconnaissance de la classification naturelle par l'algorithme de reallocation-recentrage. Par contre, si $l = l'$ on déduit du théorème que l'algorithme de reallocation-recentrage donne toujours la classification naturelle aussi petit soit l'écart t entre les deux intervalles S et S'.

Ainsi si $t < (l - l')/2$, l'algorithme de reallocation-recentrage converge toujours vers la même unique solution qui est différente de la classification naturelle. Dans ce cas, ce n'est pas la procédure algorithmique qui est en cause mais le critère de minimisation de l'inertie intra-classe de la partition : L'algorithme joue parfaitement son rôle et trouve la partition d'inertie intra-classe minimum mais ce n'est pas la partition naturelle (S, S') (*cf.* le commentaire sur la propriété énoncée au paragraphe 3). On retrouve le fait que l'algorithme des nuées dynamiques avec le critère d'inertie intra-classe a tendance à donner des classes de mêmes volumes (*cf.* [ScSy71], [CeDi84]) et a de ce fait des difficultés à reconnaître des classes de volumes différents mais aussi de formes différentes (ce dernier point ne pouvant pas se déduire de la présente étude, mais plutôt d'une présentation dans le cadre de décomposition d'un mélange Gaussien du problème de la recherche d'une partition à inertie intra-classe minimum (*cf.* [ScSy71], [CeDi84]) pour peu que l'écart entre ces classes soit faible.

Ainsi dans le cas où l'on subodore une structure en classes de cette nature, il faut éviter de choisir ce critère d'inertie intra-classe, mais encore une fois la procédure algorithmique (de type nuées dynamiques) n'est pas en cause. Nous ne pouvons ici passer en revue les méthodes permettant de trouver de telles

structures, signalons toutefois que si des méthodes de partitionnement comme par exemple la méthode des distances adaptatives (*cf.* [Gov75], [Di80]) sont à même de reconnaître des classes de formes différentes, les méthodes hiérarchiques (*cf.* [Ler81]) sont plus à même de reconnaître des classes de volumes très différents.

D'un autre côté, le théorème du paragraphe 6 fait bien apparaître la faiblesse de la procédure algorithmique : Les solutions obtenues dépendent de l'initialisation pour peu que les classes soient relativement proches. Il existe différentes techniques pour tenter de s'affranchir de ce problème de l'initialisation.

La méthode des pôles d'attraction (*cf.* [LeLe77]) est certainement l'une des méthodes les plus pertinentes pour obtenir une bonne initialisation de l'algorithme de réallocation-recentrage.

On peut montrer (s'adresser à l'auteur pour une démonstration détaillée) que pour que la méthode des pôles d'attractions conduise à la partition naturelle si $l \geqslant 3l'$ et si $(l - l')/2 \leqslant t \leqslant (l + l')^2/16l'$ il faut de plus que $l > 7l'$ et que $t > \sup((l + l')/2, l - 3l')$.

Ces deux conditions sont assez restrictives. Dans le cas étudié ici, la méthode des pôles d'attraction ne résout que très partiellement le problème d'initialisation de l'algorithme de réallocation-recentrage.

BIBLIOGRAPHIE

[Boc85] H. H. Bock, *On some significance tests in Cluster Analysis,* Journal of Classification, 1985.

[CeDi84] G. Celeux, J. Diebolt, *Reconnaissance de mélanges et classification. Un algorithme d'apprentissage probabiliste : l'algorithme SEM,* Rapport de recherche INRIA n° 348.

[Did80] E. Diday et collaborateurs, *Optimisation en classification automatique,* Éditeur : INRIA, 1980.

[DMS77] J. Diebolt, W. L. Miranker, J. C. Simon, *The dynamic cluster algorithm with continuous data,* IBM Research Report, 1977.

[Efr82] B. Efron, *The Jackknife, the Bootstrap and others resampling Plans,* SIAM, 1982.

[Go75] G. Govaert, *Classification automatique et distances adaptatives,* Thèse de troisième cycle Université Paris 6, 1975.

[Lec74] Y. Lechevallier, *Optimisation de quelques critères en Classification Automatique,* Thèse de troisième cycle Université Paris 6, 1974.

[Lem83a] J. Lemaire, *Propriétés asymptotiques en classification. Convergence des solutions approchées,* Statistique et Analyse des Données, juin 1983.

[Lem83b] J. Lemaire, *Propriétés asymptotiques en classification. Convergence d'un schéma d'approximation stochastique,* Actes des troisièmes journées internationales d'Analyse des Données, North Holland, 1983.

[LeLe77] I. C. Lerman, H. Leredde, *La méthode des pôles d'attraction,* Actes des premières journées internationales d'Analyse des Données, North Holland, 1977.

[Ler81] I. C. Lerman, *Classification et analyse ordinale des données,* Dunod 1981.

[Ler86] I. C. Lerman, *Convergence optimale de l'algorithme de reallocation-recentrage dans le cas continu le plus simple,* RAIRO R.O., 1986 no 1.

[Per83] C. Perruchet, *Significance tests for clusters: overview and comments,* Numerical Taxonomy, 1983.

[Pol81] D. Pollard, *Strong consistency of k-means clustering,* Annals of Statistics, 1981.

[Pol82] D. Pollard, *A central limit theorem for k-means clustering,* Annals of Probability, 1982.

[ScSy71] A. J. Scott, M. J. Symons, *Clustering methods based on likelihood ratio criteria,* Biometrics, Vol. 27, 1971.