LES CAHIERS DE L'ANALYSE DES DONNÉES

R. ROUSSEAU

Classification dichotomique optima et convexité des classes pour le critère de la variance

Les cahiers de l'analyse des données, tome 21, nº 3 (1996), p. 261-270

<http://www.numdam.org/item?id=CAD_1996__21_3_261_0>

Les Cahiers de l'Analyse des Données
Vol. XXI - 1996 - n°3 - pp.261-270.

CLASSIFICATION DICHOTOMIQUE OPTIMA ET CONVEXITÉ DES CLASSES POUR LE CRITÈRE DE LA VARIANCE

[CLASS. DICH. OPTIMA]

*R. ROUSSEAU**

Bien qu'offrant des représentations indispensables à l'analyse des données multidimensionnelles, la classification nous paraît être en butte à de nombreuses difficultés que l'on n'a pas résolues jusqu'ici; et dont certaines sont peut-être insolubles, tout au moins avec les moyens de calcul d'aujourd'hui.

1 Séparation linéaire des classes construites par CAH

Dans l'article [CONV. CAH], (in *CAD*, Vol.X, n°3; 1985), K. Ben SALEM et J.-P. BENZÉCRI, posent la question suivante:

Soit I un ensemble fini, muni de masses et distances, réalisé dans un espace euclidien (i.e. muni d'une métrique quadratique définie positive): appliquons à ces données l'algorithme de Classification Ascendante Hiérarchique (CAH) avec le critère d'agrégation suivant la variance (: i.e., en agrégeant, à chaque pas de la construction ascendante, une paire de sommets s et s' choisie telle que soit aussi faible que possible la variance interclasse de la partition que {s, s'} offre de $s \cup s'$). Soit deux classes A(q) et B(q) (généralement appelées Aîné et Benjamin) en lesquelles se partage une classe q de la hiérarchie: les polyèdres convexes engendrés par A(q) et B(q) sont-ils nécessairement d'intersection vide? [question qui, comme le rappelle le titre de l'article cité, équivaut à celle-ci: existe-t-il, dans l'espace ambiant à I, un hyperplan partageant l'espace en deux demi-espaces dont l'un contient A(q) et l'autre B(q)].

Le réponse est OUI si l'espace ambiant est une droite (cas de la dimension 1); mais peut être NON, dès la dimension 2.

De façon précise, l'étude citée, se fonde sur la proposition suivante:

Soit qa, qb, qs trois classes, qu'il est commode de désigner chacune par le même sigle que son centre, et dont on peut supposer qu'elles sont des sommets

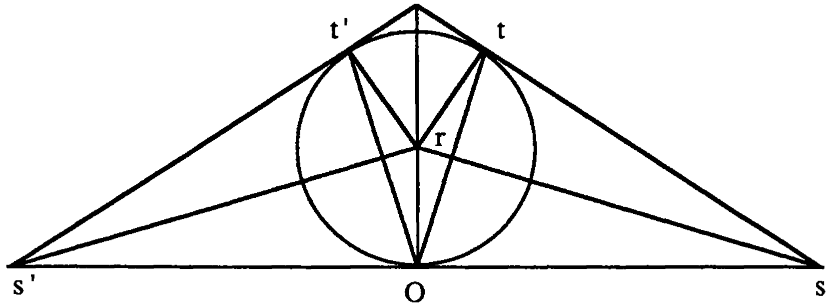
(*) Institut de Mathématiques Appliquées: Université Catholique de l'Ouest.

à une certaine étape de l'application de l'algorithme de CAH; alors si les deux segments (qs, qa) et (qs, qb) forment, en qs, un angle obtus (pour la structure euclidienne de l'espace ambiant), qa et qb ne peuvent former un couple de plus proches voisins réciproques. Ce qu'on peut exprimer en faisant image, sinon explicitement, en disant que la présence de qs, entre qa et qb, fait obstacle à l'agrégation de qa avec qb.

En dimension 1, il résulte de la proposition que, dans la CAH, l'aîné A(n) et le benjamin B(n), d'un nœud n, s'identifient chacun avec la partie de I comprise dans un segment, les deux segments afférents à A(n) et B(n) (qui ne sont autres que les supports convexes de ces classes) étant d'intersection vide. Ce résultat se démontre par récurrence en suivant le déroulement de l'algorithme de CAH: au départ, on a la partition triviale formée des classes réduites à un point de I; et, à un niveau de la construction, l'agrégation de deux sommets ne peut se faire que pour deux classes, qa et qb, dont chacune est la restriction de I à un segment Ia ou Ib, ces segments étant d'intersection vide (de par l'hypothèse de récurrence) et sans points de I entre les deux (selon la proposition rappelée ci-dessus).

Mais dès la dimension 2, la même proposition suggère un contre-exemple: on a un ensemble I, de 5 points munis de masses et distances; au sommet de la CAH, I se partage en deux classes; dont l'une A(I) comprend 4 points, sommets d'un quadrilatère convexe à l'intérieur duquel se trouve la classe B(I), réduite au seul point restant de I.

Dans [CONV. CAH], le contre-exemple est étudié minutieusement. Ici nous nous bornerons à en rappeler le schéma.



$r = \{0; 1\}$; $s = \{(1/\varepsilon); 0\}$; $s' = \{(-1/\varepsilon); 0\}$;

$m(r) = m(t) = m(t') = 1$; $m(s) = m(s') = \varepsilon^6$;

t et t' sont les points de contact des tangentes, au cercle de centre r et de rayon 1, autres que l'axe horizontal, et issues respectivement de s et de s'.

Pour une valeur suffisamment faible de ε, la CAH agrège, au plus bas niveau, chacun des deux couples $\{s, t\}$ et $\{s', t'\}$; et, du fait de la grande différences des masses, les centres de gravité des classes ainsi constituées diffèrent arbitrairement peu de t et t'. La CAH agrège ensuite $\{s, t\}$ et $\{s', t'\}$, dont les centres sont séparés par une distance équivalente à $4.\varepsilon$. Enfin la classe $\{s', t', t, s\}$, dont les quatre points sont les sommets d'un trapèze, s'agrège au point r, intérieur à ce trapèze; d'où la hiérarchie:

$$\{\{\{s, t\}, \{s', t'\}\}, r\}.$$

On notera que l'emploi de masses très inégales n'ôte rien à la généralité de l'exemple: car, au lieu des masses $\{1, 1, 1, \varepsilon^6, \varepsilon^6\}$, on peut prendre, par exemple, $\{10^6, 10^6, 10^6, 1, 1\}$; or, placer une masse de 10^6, équivaut à placer 10^6 individus de masse 1, au même point.

2 Classification hiérarchique optima selon le critère de la variance

Le contre-exemple qu'on a cité, laisse place à un nouveau problème: puisque, par la voie ascendante, on n'est pas assuré de construire une hiérarchie de classes dont les dichotomies soient linéairement séparables, peut-on, indépendamment de tout algorithme de classification efficace, définir une classification qui soit optima pour le critère de la variance ? Et, cette classification étant supposée définie, les dichotomies en sont-elles linéairement séparables.

[N.B. Le terme d'optima ne doit pas ici faire illusion: il s'agit d'un optimum quant au critère formel; non quant à la valeur taxinomique de la hiérarchie. En effet, même si le contre-exemple du §1 présente une imbrication des classes de la CAH qui n'est satisfaisante d'aucun point de vue, il se peut qu'en d'autres cas, existent, dans l'espace, des parties denses de I, couvrant des domaines non convexes imbriqués entre eux, et qu'on ne souhaite pas voir morcelés du fait de la séparation linéaire.]

De façon précise, posons la définition suivante:

Une classification hiérarchique binaire, d'un ensemble euclidien I, muni de masses et distances, sera dite Dichotomique Optima pour le critère de la variance si: pour tout nœud n de la hiérarchie, $\{A(n), B(n)\}$ constitue une partition de n en deux classes réalisant le maximum possible de la variance interclasse (pour une dichotomie de n): en d'autres termes, telle que les classes A(n) et B(n) soient chacune aussi concentrées que possible (en ce sens qu'est aussi faible que possible la somme de leurs variances intraclasses).

En particulier, au sommet de la hiérarchie, la dichotomie $\{A(I), B(I)\}$ doit réaliser le maximum de la variance interclasse pour les dichotomies de I.

Il va sans dire que, comme c'est presque toujours le cas en Classification, il n'y a pas unicité: la classification optima définie ci-dessus ne peut-être unique, en toute généralité.

On va démontrer qu'une Classification Dichotomique Optima (CDO) satisfait à deux propriétés souhaitées *a priori*. D'une part, dans la hiérarchie, il n'y a pas d'inversion des niveaux: en ce sens que, si deux classes A(c) et B(c) s'agrègent au niveau nu(c), pour constituer c; alors les deux classes A(A(c)) et B(A(c)) s'agrègent à un niveau nu(A(c)) $\leq$ nu(c). D'autre part, les classes de la CDO sont linéairement séparables.

Notons : $A(A(c)) = qa$; $B(A(c)) = qb$; $c - (qa \cup qb) = B(c) = rc$;

$$Crit(qa, qb) = nu(A(c)) \ ; \ Crit((qa \cup qb), rc) = nu(c) \ .$$

On va appliquer au système de trois parties, ou triangle, {qa, qb, rc}, ce qu'on appelle, en théorie de la classification, l'axiome de la médiane (auquel satisfait le critère de la variance). Dans l'axiome, comme dans le calcul du critère, seuls sont en jeu les centres et masses des parties {qa, qb, rc}; l'axiome affirme que la médiane (comprise comme valeur du critère pour l'agrégation d'une des trois parties, qualifiée de sommet, avec la réunion des deux autres, jouant le rôle du milieu d'un côté) opposée au plus petit côté est nécessairement supérieure à ce côté (entendu comme valeur du critère pour le couple de sommets, entre lesquels ce critère est minimum). Ceci posé, si (qa, qb) est le plus petit côté, avec nu(A(c)) pour critère, alors nu(A(c)) est $\leq$ nu(c) qui correspond à la médiane opposée à ce plus petit côté. Reste à montrer qu'il est absurde que (qa, qb) ne soit pas plus petit côté. En effet, supposons, par exemple, que la valeur du critère soit strictement moindre pour (qa, rc) que pour (qa, qb); alors l'on a:

$$Itot = Crit((qa \cup qb), rc) + Crit(qa, qb) = Crit((qa \cup rc), qb) + Crit(qa, rc) \ ;$$

formule où Itot représente l'inertie interclasse du système des trois classes {qa, qb, rc}; ou inertie du nuage de trois points, chaque classe étant assimilée à son centre. De cette égalité résulte l'implication, entre inégalités strictes:

$$[Crit(qa, rc) < Crit(qa, qb)] \quad \Rightarrow$$
$$[Crit((qa \cup qb), rc) < Crit((qa \cup rc), qb)] \quad ;$$

d'où il résulte que la dichotomie de c en (qa $\cup$ qb) et rc ne serait pas optima, contrairement à l'hypothèse qu'on a une une CDO.

Reste à montrer que les classes d'une CDO sont linéairement séparables. On retrouve ici des considérations souvent faites par E. DIDAY, puis L. LEBART, à propos de l'algorithme d'agrégation autour de centres mobiles [dit encore: algorithme des nuées dynamiques].

Notons A(s) = qa ; B(s) = qb; et proposons-nous de montrer que la dichotomie de s en {qa, qb} n'est pas optima si qa et qb ne sont pas linéairement séparables.

On peut, en fait, dire plus précisément: les classes qa et qb sont de part et d'autres de l'hyperplan H médiateur du segment joignant leurs deux centres gravités (notés, également qa et qb); et, de plus, aucune masse de s n'est portée par l'hyperplan lui-même; ce que nous exprimerons, en bref, en disant que qa et qb sont strictement séparables par H.

En effet, supposons qu'un point ia de qa soit dans le demi-espace fermé Hb, limité par H et contenant qb. Posons qa' = qa - {ia} et qb' = qb $\cup$ {ia}: on a une nouvelle dichotomie de s en {qa', qb'} pour laquelle on montre que le critère est plus fort qu'il n'est pour {qa, qb}. Comme de règle, on tient compte de la complémentarité entre variance interclasse et intraclasse. En faisant passer ia de qa dans qb, pour créer qa' et qb', on a affecté le point ia à un centre duquel il est plus proche que du précédent (à moins que la distance ne soit la même, si ia est sur l'hyperplan H); ceci suffit à montrer que la somme des inerties de qa', relativement au centre de qa, et de qb', relativement au centre de qb, ne peut être plus grande que la somme des inerties internes des classes initiales qa et qb; mais, quant à l'inertie interne de qa' et qb', elle est strictement moindre, car on doit la calculer, non relativement aux centres des classes précédentes, mais aux centres mêmes des nouvelles classes; ce qui ne peut que diminuer strictement l'inertie, dont le minimum pour une distribution de masse finie à support borné, est réalisé précisément au centre de gravité.

3 Classification hiérarchique optima et partition optima

Reprenons la donnée des §§ précédents: I, ensemble fini, muni de masses et distances, réalisé dans un espace euclidien.

Une partition de I en k classes sera dite Optima si elle réalise le minimum de l'inertie intraclasse (ou, corrélativement, le maximum de l'inertie interclasse) parmi toutes les partitions de I comptant ce nombre de classes. En reprenant les considérations du §2, on démontre que les classes d'une partition Optima sont, deux à deux, strictement séparables par un hyperplan. Donc les enveloppes convexes des classes sont deux à deux d'intersection vide.

On peut dire que, dans une Classification Dichotomique Optima (CDO), au sens du §2, pour tout nœud n, {A(n), B(n)} est une partition optima de n en deux classes.

Il vaut la peine de chercher à quel prix pourrait être construite une CDO, ou une partition optima. Certes, l'ensemble des partitions dont est susceptible un ensemble fini I, ou même l'ensemble des classifications dichotomiques de I, sont des ensembles finis: mais le cardinal de ces ensembles croît si vite avec cardI, qu'il est pratiquement impossible de chercher l'optimum en examinant tous les cas.

L'algorithme de E. DIDAY est bien connu: à partir d'une partition quelconque, P, de I en k classes, on construit une partition $\pi(P)$, comprenant au plus k classes non vides, chacune de celles-ci, associée à une classe p de P, étant formée de l'ensemble des éléments i de I dont la distance au centre de gravité de p réalise le minimum de la distance de i à un centre de classe de P; s'il y a une classe vide, on peut scinder une des classes obtenues comptant plus de 2 éléments.

De façon précise, il importe ici de considérer avec une attention particulière le cas où des distances sont égales. Soit un point i, pour lequel le minimum de la distance à un centre de classe est réalisé simultanément pour les deux centres, cp et cp', des classes p et p'. Le point i peut, *a priori*, être affecté aussi bien à p qu'à p'. Si l'affectation est faite arbitrairement pour plusieurs points, il se peut que les centres de gravité des classes ne soient même pas modifiés . C'est le cas, en particulier, si deux points, de même masse, i et i' étant superposés, la seule modification apportée à la partition consiste à faire passer *i* de p à p'; et *i'* de p' à p; cas où il est clair que les variances des classes ne changent pas. Pour être assuré de faire strictement décroître la variance intraclasse, même si les modifications possibles pour l'ensemble de la partition ne concernent que des points situés à égale distance de deux centres, il faut, en ce cas, (comme à la fin du §2), ne changer de classe qu'un seul point. Ainsi, on peut reprendre les considérations de la fin du §2, qui assurent que, les centres de gravités des deux classes concernées étant déplacés, l'inertie intraclasse totale diminue strictement.

[Afin d'accélérer la convergence, et, plus encore, en vue de la généralisation à des distributions continues, cf. *infra* §4, on peut procéder à des réaffectations multiples. Soit {p, p'} deux classes avec, sur l'hyperplan médiateur, H, du segment joignant leurs centres {cp, cp'}, des masses susceptibles d'être affectées, indifféremment, à cp ou à cp': toutes ces masses seront ensemble affectées au même centre; le choix entre cp et cp' étant indifférent, à ceci près qu'il faut que la partition soit effectivement modifiée].

Cette réserve étant faite, il est clair que, partant d'une partition initiale P0 arbitraire, on construit une suite de partitions P1 = π(P0); P2 = π(P1), etc., de telle sorte que la variance intraclasse décroît strictement à chaque pas; jusqu'à ce que π(Px) = Px: le processus s'arrête quand chaque point se trouve nécessairement affecté à la classe où il est déjà. Ce terme est atteint en un nombre fini de pas: car la décroissance stricte de l'inertie implique qu'il s'agit d'une suite de partitions distinctes. On dit alors qu'on est parvenu à une partition stationnaire.

Il résulte de la définition même du processus π de réaffectation, que les classes d'une partition stationnaire sont strictement séparables. Mais il s'en faut de beaucoup que toute partition stationnaire soit optima. C'est pourquoi, désirant approcher l'optimum, L. LEBART commence par faire une CAH de I; et, ensuite, applique le processus de réaffectation π, à partir de la partition P0, de I, définie par les k-1 nœuds les plus hauts de la CAH.

En particulier, si, cf. §1, P0 comporte des classes non séparables par hyperplan, le processus π aboutit à une partition Px où cette condition est satisfaite. Cependant, la structure peut être modifiée par π, en ce sens qu'une CAH faite sur les centres des classes de Px, considérés avec leurs masses, ne donne pas nécessairement la structure hiérarchique de P0. [Et, de plus, ainsi qu'on l'a noté dès le début du §2, il n'est pas souhaitable, d'imposer, après CAH, la séparabilité linéaire, si celle-ci n'est pas conforme à la structure géométrique des données.]

L'exemple d'une distribution linéaire de densité uniforme montre qu'une partition optima n'est pas toujours comprise dans une classification dichotomique optima; et qu'une partition extraite d'une CD0 (et non seulement celle obtenue par CAH usuelle) peut être modifiée par l'algorithme de réaffectation, bien que ses classes soient strictement séparables car l'hyperplan de séparation n'est pas nécessairement l'hyperplan médiateur du segment joignant les CdG.

Partons du cas limite d'une distribution de densité unité sur un segment de la droite réelle, munie de sa métrique usuelle. L'inertie d'un segment de longueur x relativement à son milieu est de la forme $\mu.x^3$ (où $\mu = 1/12$). Pour un segment donné, par exemple (0, 1), la dichotomie optima est obtenue par partage en deux segments égaux. D'où il résulte que la CDO d'un segment s'obtient en divisant tout nœud n en deux segments égaux A(n) et B(n).

Cherchons maintenant une partition optima d'un segment en trois: $k = 3$. Il y a trois segments consécutifs, dont les longueurs, peuvent être notées: a, b, c. Si $a \neq b$, on peut diminuer l'inertie intraclasse de la partition en substituant à $\{a, b\}$ deux segments égaux $\{a', b'\}$ avec $a \cup b = a' \cup b'$. Donc, dans la partition optima, a et b ont même longueur; et, semblablement b et c. La partition optima, formée de trois segments consécutifs égaux, ne peut donc se trouver dans la CDO, qui se fait en divisant indéfiniment en deux moitiés égales le segment donné et ses subdivisions.

Il est facile de reprendre la démonstration ci-dessus dans le cas où I est une suite de points équidistants ayant tous la même masse. Les calculs et notations sont seulement plus complexes: car il faut, en particulier, tenir compte de ce qu'une suite de points ne peut, dans certains cas, être partagée en deux - ou trois - parties égales qu'à une unité près.

4 Partition optima pour une distribution de masses

Cette excursion dans le domaine des distributions de masse non discrètes appelle notre attention sur une généralisation des propriétés géométriques considérées ici.

En effet, tandis que l'algorithme de CAH requiert un ensemble discret, les notions de partition optima et de CDO, peuvent se définir pour toute distribution de masse positive finie à support compact, μ; et les propriétés de convexité et de séparabilité qui ont été démontrées s'étendent à ce cas. Ou, même, plus généralement, au cas où μ, étant de masse finie sans être à support compact, a une inertie totale finie; comme c'est le cas, notamment, pour une distribution normale.

En bref, au lieu de parties finies a, b, ... de I, on considère des distributions de masse positive, μa, μb; et, au lieu de changer de classe, comme on l'a fait au §2, un point ia qui est dans le demi-espace Hb contenant b, on incorpore à μb la partie de la mesure μa portée par le demi-espace Hb, limité à l'hyperplan médiateur H du segment joignant les centres de gravité de μa et μb, Hb étant le demi-espace contenant ce dernier centre.

De façon précise, soit μ une distribution finie de masses positives sur un espace euclidien, E, de dimension finie. Afin de généraliser la recherche d'une partition optima en r classes, on considère les décompositions de μ en une somme d'un système, M, de r mesures positives, {μ1, μ2, ..., μs, ..., μr} : μ = Σ{μs | s = 1, ..., r}. Pour la décomposition M, est définie l'analogue, II(M) d'une inertie intraclasse: II(M) étant la somme des inerties de chacune des distribution de masse μs relativement à son centre de gravité g(μs). Et II(M) a un minimum, IImin, quand M parcourt l'ensemble de toutes les décompositions possibles. On va montrer que ce minimum est atteint; et cela pour un M, dont les μs ont des supports séparables strictement; i.e. par des portions, ne portant pas de massenon nulle, d'hyperplan médiateur.

Relativement à un système C = {c1, ..., cs, ..., cr} de r points de E, on définit une inertie d'affectation, I(μ, C) comme l'intégrale, prise pour la mesure μ, de la fonction D(C, x) qui à tout point x de E associe le minimum des distances au carré de x à tous les points du système C. Au système C, on peut associer une partition de E en parties mesurables, Es, par la condition que tout point x est compris dans la partie Es, afférente au centre cs dont il est le plus proche; ou, s'il y a égalité, à celles des parties concurrentes dont l'indice *s* est le plus faible. Ainsi, μ est *ipso facto* décomposé en une somme d'un système, M(C), de termes positifs, μs(C), défini chacun comme la restriction de μ à Es.

Il est clair que II(M(C)), calculé relativement aux centres de gravité, g(μs(C)), des mesures respectives, μs(C), est ≤ I(μ, C), calculé relativement aux cs; l'égalité ne pouvant être réalisée que si chacune des μs a pour centre de gravité cs = g(μs(C)). Donc, si ICmin désigne le minimum de I(μ, C) sur l'ensemble de tous les C, on a IImin ≤ ICmin.

D'autre part, si, partant d'une décomposition M = {μs}, on prend C = {g(μs)}, système des centres de gravités des μs: on a I(μ, C) = II(M); d'où, ICmin ≤ IImin. Donc, compte tenu de l'inégalité précédente, IImin = ICmin.

Supposons que pour un système C, soit réalisé le minimum strict de la fonction I(μ, C). Dans ce cas, II(M(C)) réalise le minimum IImin; et, de plus, les μs sont strictement séparables (au sens précisé ci-dessus); car autrement, on pourrait, (cf. *supra*), en transférant entre deux termes {μs, μs'} des masses portées par l'hyperplan médiateur de (cs, cs'), modifier effectivement les cs et diminuer l'inertie intraclasse.

Pour démontrer que le minimum IImin peut être réalisé par une décomposition convenable de μ, de forme M = {μs | s = 1, ..., r}, il suffit donc de démontrer que le minimum ICmin peut être réalisé par un système de centres C = {cs}. Ceci résultera de la continuité de I(μ, C) en fonction de C.

En effet, notons G le centre de gravité de la mesure μ considérée; et supposons, pour simplifier les notations, que l'inertie de μ relativement à G vaut 1. Considérons deux systèmes de centres C={cs} et C'={c's}. Notons eps le maximum pour s allant de 1 à r, de la différence entre les carrés des distances à G des points cs et c's. Entre l'inertie de μs(C) relativement à cs et son inertie relativement à c's, la différence est certainement moindre qu'entre l'inertie de μ, toute entière, relativement à ces deux centres; différence qui, avec les notations posées, ne peut dépasser eps. Donc l'inertie du système des μs(C) relativement aux c's est majorée par I(μ, C)+r.eps; *a fortiori*, I(μ, C') est majorée par I(μ, C)+r.eps. On a donc les deux inégalités analogues entre elles:

$$I(\mu, C') \leq I(\mu, C) + r.eps \ ; \ I(\mu, C) \leq I(\mu, C') + r.eps \ ;$$

dont résulte la continuité de I(μ, C) en fonction de C.

Dans le cas où la distribution de masse, μ, est à support compact, e.g. incluse dans la sphère S(G, R) de centre G et de rayon R, il est clair que, si un système C de centres comporte un centre, cs, situé à une distance de G supérieure à R, on peut diminuer I(μ, C) en rapprochant ce centre de G. Le minimum de I(μ, C) est donc à chercher dans l'ensemble des systèmes de centres renfermés dans la sphère S(G, R); ensemble compact où la fonction continue I(μ, C) atteint effectivement (pour un système C au moins) son minimum, ICmin.

Dans le cas où μ n'a pas de support compact, on peut considérer une suite {C(i) | i=1, 2...} de systèmes de centres, telle que I(μ, C(i)) tende vers ICmin quand i tend vers l'infini. On peut restreindre {C(i)} à une sous-suite {C(j)} pour laquelle chacun des cs(j) converge vers un point à distance finie, cs, ou tende vers l'infini. On va montrer que la limite ICmin est aussi bien atteinte si l'on substitue à la suite des C(j) une suite de C'(j) où les centres cs(j) qui tendent vers l'infini sont remplacés par G. En effet pour tout eps, on peut trouver un rayon R tel que soit < eps l'inertie relativement à G de la partie de μ située en dehors de la sphère S(G, R). Soit maintenant jeps un rang au-delà duquel, d'une part, IC(μ, C(j)) < ICmin + eps; et, d'autre part, les cs(j) qui tendent vers l'infini sont tous à une distance de G > 2.R. En substituant aux affectations à l'un de ces centres des affectations à G, on ne peut augmenter l'inertie IC que s'il s'agit de masses dont la distance à G dépasse R; mais pour ces masses, l'inertie totale relativement à G est déjà majorée par eps: ce qui prouve que, si j > jeps, IC(μ, C'(j)) $\leq$ IC(μ, C(j)) + eps $\leq$ ICmin + 2.eps. D'où il résulte que la limite des C'(j) réalise, à distance finie, le minimum ICmin.

N.B.: on montre que le cas envisagé ci-dessus (où l'un des centres tend vers l'infini) ne peut être l'optimum du fait que, d'une part, l'inertie intraclasse d'une partition optima en k classes est une fonction strictement décroissante de k, et, d'autre part: la limite de I(μ, C) quand l'un des centres de C tend vers l'infini, le système C' des (k-1) autres étant fixes, n'est autre que I(μ, C').

Nous terminerons en proposant une autre généralisation. On a considéré exclusivement le critère de la variance, intimement lié à la notion de convexité. Mais la notion de Classification Dichotomique Optima peut être étendue à tout critère; et, si l'axiome de la médiane est vérifié avec la condition complémentaire:

$$\forall\ a, b, c: \ \mathrm{Crit}(a \cup b, c) + \mathrm{Crit}(a, b) = \mathrm{Crit}(a \cup c, b) + \mathrm{Crit}(a, c)\ ,$$

on peut, comme au §2, démontrer qu'une CDO ne présente pas d'inversion dans la suite des niveaux des nœuds.

Référence bibliographique

K. Ben SALEM et J.-P. BENZÉCRI : “Sur la séparabilité linéaire des enveloppes convexes des classes d'une CAH”; in *CAD*, Vol.X, n°3; (1985).