LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI
K. BENSALEM

Sur la séparabilité linéaire des enveloppes convexes des classes d'une CAH

Les cahiers de l'analyse des données, tome 10, n° 3 (1985), p. 272-278

<http://www.numdam.org/item?id=CAD_1985__10_3_272_0>

Les Cahiers de l'Analyse des Données
Vol. X - 1985 - n° 3 - pp 272-278

SUR LA SÉPARABILITÉ LINÉAIRE DES ENVELOPPES CONVEXES DES CLASSES D'UNE CAH

[CONV. CAH]

*par J.P. Benzécri *, K. Bensalem ***

Il serait satisfaisant que les classes constituées en CAH soient linéairement séparables au sens suivant : que les enveloppes convexes conv(a(n)) et conv(b(n)) des deux descendants d'un noeud n soient d'intersection vide ; (ou que plus généralement, à chaque étape de la construction ascendante, les sommets soient des classes dont les enveloppes convexes ne se recoupent pas). La présente note (rédigée comme un problème suivi de sa solution) montre que la séparabilité parfaite existe en dimension 1 (i.e. quand l'espace ambiant est une droite ; cf. § 2.1 Remarque) ; un résultat partiel vaut dans le plan pour des classes réduites à deux points (§ 2.2) ; mais un contre exemple simple montre que la séparabilité n'existe pas en général (§ 2.4).

On sait que l'algorithme de classification hiérarchique, procède à partir d'un ensemble I d'individus, en créant des noeuds par agrégations successives de paires d'individus ou de noeuds déjà créés ; les éléments non encore agrégés sont appelés sommets ; deux sommets s et s' peuvent être agrégés s'ils sont plus proches voisins réciproques au sein de l'ensemble S des sommets ; c'est-à-dire si chacun réalise dans S le minimum de la distance à l'autre. En fait, le terme de distance est impropre : il s'agit plus exactement d'une quantité critère appelée "niveau d'agrégation"; pour le calcul de celle-ci, plusieurs formules peuvent être utilisées, la plus avantageuse étant, selon nous, celle du critère d'agrégation suivant l'inertie, définie par :

$$niv(s,s') = (m.m'/(m + m'))\,|ss'|^2 ,$$

où m et m' sont les masses respectives des classes s et s' ; et |ss'| la distance euclidienne entre leurs centres de gravité ; (centres qu'on désignera généralement dans la suite par la même lettre que les classes correspondantes)

1 *Enoncé du problème*

1.1 L'objet de la première partie est de répondre à la question suivante :

Supposons qu'il existe trois sommets notés t, s, s', dont les centres constituent un triangle ayant en t un angle obtus : est-il possible que s et s' soient agrégés en présence de t .

On adoptera les notations suivantes :

Les classes t, s, s' ont pour masses respectives n, m, m' ; les vecteurs $\vec{ts}$, $\vec{ts'}$ sont désignés par x et x'.

*(*) Professeur de statistique. Université Pierre et Marie Curie.*

(**) Assistant à la Faculté des Sciences de Tunis.
Docteur Ingénieur en Analyse des Données.

1.1.1 Montrer que l'inégalité :

$$\text{niv}(s,s') \leq \text{niv}(s,t)$$

peut se mettre sous la forme :

$$(1) \quad A|x - x'|^2 \leq (m + m')n\,|x|^2$$

où A est une fonction de m, m' et n qu'on précisera.

1.1.2 De l'inégalité ci-dessus et de l'inégalité analogue obtenue en échangeant les rôles de s et s', déduire que, pour que s et s' soient plus proches voisins réciproques, il faut que soit satisfaite une inégalité de la forme :

$$B|x - x'|^2 \leq (m + m')n\,(|x|^2 + |x'|^2),$$

où B est une fonction de m, m' et n qu'on précisera.

1.1.3 Compte tenu de l'inégalité du § 1.1.2, est-il possible que s et s' soient plus proches voisins réciproques si est négatif le produit scalaire $\langle x,x' \rangle$.

1.2 . L'objet de la deuxième partie est de préciser la région de l'espace où peut se trouver t si s et s' sont plus proches voisins réciproques. Pour simplifier le langage, on supposera que l'espace est restreint à un plan passant par s et s'.

1.2.1 En groupant dans le membre de droite les termes en n des deux membres de l'inégalité (1) du § 1.1.1., on obtient une inégalité de la forme :

$$(2) \quad U \leq V\,n,$$

où U et V sont des expressions que l'on précisera. Compte tenu du signe de n, U et V, montrer que l'inégalité (2) fournit pour $|st|^2$ une borne inférieure T, qu'on exprimera en fonction de m , m' et $|ss'|^2$.

1.2.2 On désigne par σ le centre de gravité de l'ensemble des deux points (s,s') munis de leurs masses (m,m').

Calculer $\overrightarrow{s\sigma}$ en fonction de $\overrightarrow{ss'}$, m et m'.

On considère le triangle rectangle shs' ayant pour hypothénuse ss', et tel que σ soit le pied de la hauteur issue de h.

Exprimer $|sh|^2$ en fonction de $|ss'|^2$, m et m' ; et comparer l'expression de $|sh|^2$ à la borne T demandée au § 1.2.1.

1.2.3 En cherchant pour $|s't|^2$ une borne inférieure, comme on l'a fait pour $|st|^2$, préciser dans quelle région du plan doit se trouver le point t pour que s et s' soient p.p. voisins réciproques. Pour une position donnée de t , préciser les valeurs possibles de n. (On accompagnera la solution d'une figure).

1.3 L'objet de la troisième partie est de répondre à la question suivante. Supposons qu'il existe dans le plan quatre points t, t', s, s' de masses respectives n, n', m, m' ; que les deux segments (t,t')

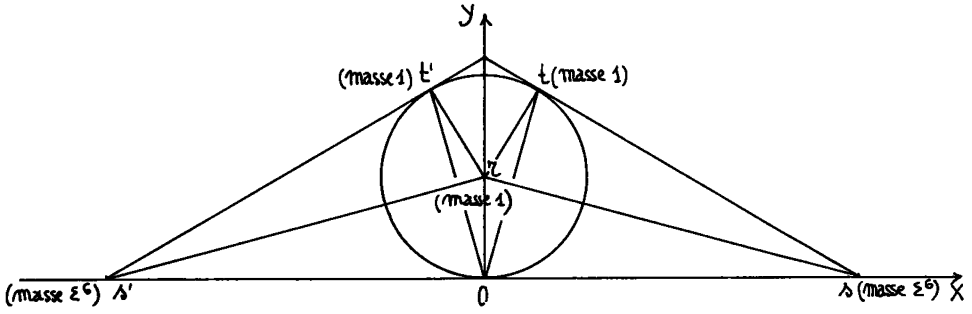
et (s,s') aient un point commun o. et que s et s' soient plus proches voisins réciproques (en présence de t et t'). Est-il possible qu'après agrégation de s et s' en leur centre de gravité σ, t et t' soient plus proches voisins réciproques (en présence de σ).

1.3.1 On suppose que le point o se trouve sur le segment (s,s') entre s et σ ; i.e. : $o \in (s,\sigma)$. Désignons par f et f' les points où la droite passant par t et t' coupe le cercle de centre s et de rayon $|sh|$ (cf. § 1.2.2) : Quel est le signe du produit scalaire $\langle \sigma f , \sigma f' \rangle$.

1.3.2 Compte tenu des résultats des §§ 1.2.3 et 1.3.1, préciser le signe du produit scalaire $\langle \sigma t , \sigma t' \rangle$.

1.3.3 Compte tenu des résultats des §§ 1.3.2 et 1.1.3, répondre à la question posée en tête du § 1.3.

1.4 L'objet de la quatrième partie est d'étudier sur un exemple ce que peut être la position relative des enveloppes convexes de deux parties c et c' qui à une étape donnée de l'agrégation, constituent des sommets. A cet effet, on considère dans le plan (rapporté à ces deux axes orthonormés OX , OY) un nuage de cinq points $I = (r, t, t', s, s')$ affectés des masses respectives $(1, 1, 1, \varepsilon^6, \varepsilon^6)$, où ε est un nombre petit qu'on peut prendre égal à 0,1 dans les calculs ; la disposition des points se voit sur la figure, et est précisée par les valeurs de leurs coordonnées.



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

$$t = ((2\varepsilon/(1+\varepsilon^2)) \; ; \; (2/(1+\varepsilon^2))) \; ; t' = (-(2\varepsilon/(1+\varepsilon^2)) \; ; \; (2/(1+\varepsilon^2))) \; .$$

En termes géométriques, la figure admet l'axe OY pour axe des symétries ; et le cercle de centre r et de rayon 1 admet pour tangentes respectives aux points 0, t, t', les droites ss' (axe 0X), st, et s't'. (On pourra dans la solution faire usage de ces propriétés sans les démontrer d'après les valeurs des coordonnées : on notera en particulier que $|rt| = 1$ et que $|st| = |s0| = \varepsilon^{-1}$).

1.4.1 Calculer les niveaux d'agrégations suivants :

$$niv(r,t) \; ; \; niv(s,t) \; ; \; niv(s,s') \; ; \; niv(t,t') \; ; \; niv(r,s) \; ;$$

donner pour chaque point i de I son ou ses plus proches voisins (s'il y a plusieurs plus proches voisins, on écrira : vois(i) = {i' , i"}) ; et signaler s'il existe des paires de plus proches voisins réciproques.

1.4.2 On désigne par σ le centre de gravité du système des deux points t et s, affectés de leur masses respectives 1 et ε^6 ; on définit de même, symétriquement σ'. Calculer l'abscisse $X(\sigma)$ du point σ ; calculer $niv(\sigma, \sigma')$, (les points σ et σ' étant chacun affecté de la masse $1 + \varepsilon^6$) : on pourra se borner à une expression approchée de le forme $\alpha\varepsilon^n$.

1.4.3 Décrire le résultat de la C.A.H. sur I. Préciser si à une étape, on a à agréger deux classes c et c' dont les enveloppes convexes se coupent.

2 *Solution du problème*

2.1.1 En appliquant la formule de définition de niv, l'inégalité proposée s'écrit :

$$(m\,m')/(m+m')) \; |x - x'|^2 \leq (m\,n/(m+n)) \; |x|^2 \quad ;$$

d'où après multiplication des deux membres par $((n+m)(m+m')/m)$, l'inégalité équivalente :

$$m'(n+m)\,|x - x'|^2 \leq (m+m')n|x|^2,$$

soit $A = m'(n+m)$.

2.1.2 Les sommets s et s' sont plus proches voisins réciproques en présence de t si et seulement si niv(s,s') est inférieur (ou égal) à niv(s,t) et à niv(s',t). Ce qui s'exprime par les deux inégalités ci-dessous (cf. § 2.1.1) :

$$m'(m+n) \; |x - x'|^2 \leq (m+m')n \; |x|^2$$

$$m(m'+n) \; |x - x'|^2 \leq (m+m')n \; |x'|^2$$

de ces deux inégalités on déduit par addition une troisième qui est seulement condition *nécessaire* (pour une condition n. et suffisante cf. § 2.2) pour que s et s' soient p.p.v.r. ; soit :

$$((m+m')\,n + 2\,m\,m')|x - x'|^2 \leq (m+m')n \; (|x'|^2 + |x|^2) \; ;$$

on a donc $B = (m + m')n + 2\,m\,m'$.

2.1.3 On peut récrire l'inégalité finale du § 2.1.2 en faisant apparaître le produit scalaire $\langle x , x' \rangle$; il vient :

$$((m+m')\,n + 2\,m\,m')\,(|x|^2 + |x'|^2 - 2\langle x , x' \rangle) \leq (m+m')n\,(|x|^2 + |x'|^2) \; ;$$

cette inégalité ne peut être satisfaite si $\langle x , x' \rangle \leq 0$; car alors elle prend la forme : $Bu \leq B'u'$; où B est strictement inférieur à B' et $u \leq u'$. En termes géométriques on répondra donc négativement à la question posée en 1.1.

S'il existe trois sommets t, s, s' dont les centres constituent un triangle ayant en t un angle obtus (ou droit), s et s' ne peuvent être agrégés en présence de t. (Nous spécifions "en présence de t" , car après agrégation éventuelle de t avec un autre sommet, s et s' pourraient se trouver p.p. voisins réciproques).

Remarque : il résulte de cette première partie, que si l'espace ambiant est une droite, à toute étape de la CAH, les sommets sont

des classes portées par des intervalles deux à deux non empiétants ; en effet la propriété est vraie au départ, quand les classes sont réduites à des points ; et elle est conservée à chaque agrégation, qui ne peut se faire qu'entre deux classes non séparées par une troisième.

2.2.1 En groupant les termes en n de l'inégalité (1) il vient :

$$(2) \quad m m' |s s'|^2 \leq n ((m+m')|st|^2 - m'|s s'|^2) ;$$

sont, avec les notations de l'énoncé :

$$U = m m' |s s'|^2 \quad ; \quad V = ((m+m')|st|^2 - m'|s s'|^2) ;$$

U étant positif (ainsi que la masse n) , l'inégalité (2) ne peut être vérifiée que si :

$$T = (m'/(m+m'))\ |s s'|^2 \leq |st|^2.$$

L'inégalité (2) fournit, de plus, pour chaque valeur de $|st|^2$ (supérieure à T), une borne inférieure de n , d'autant plus grande que $|st|^2$ est plus proche de T.

2.2.2 On a classiquement :

$$\overrightarrow{s\sigma} = (m'/(m+m'))\ \overrightarrow{s s'} \quad ;$$

$$|sh|^2 = |s\sigma|\ |s s'| = (m'/(m+m'))\ |s s'|^2 = T.$$

2.2.3 On a les deux inégalités analogues :

$$|sh|^2 \leq |st|^2 \quad ; \quad |s'h|^2 \leq |s't|^2 \quad ;$$

ces inégalités imposent à t de se trouver dans la région du plan extérieure à la réunion de deux disques, l'un de centre s et rayon |sh| , l'autre de centre s' et rayon s'h . Pour une position de t dans la région ainsi délimitée, ou pour n les inégalités (cf. § 2.2.1) :

$$m m'|s s'|^2/((m+m')(|s t|^2 - |s h|^2)) \leq n$$

$$m m'|s s'|^2/((m+m')(|s't|^2 - |s'h|^2)) \leq n$$

de ces deux inégalités, la plus contraignante est la première si t se projette orthogonalement sur la droite (s,s') du même côté que s par rapport à σ ; sinon c'est la deuxième.

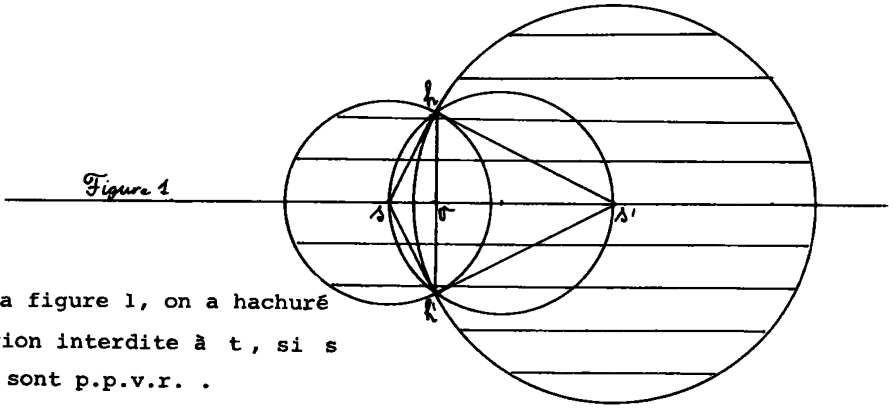
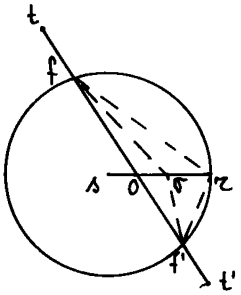


Figure 1

Sur la figure 1, on a hachuré la région interdite à t , si s et s' sont p.p.v.r. .

2.3.1 Le point σ est intérieur au cercle de centre s et de rayon |sh| ; et, par hypothèse le point 0 appartient au segment (s, σ). On voit sur la figure 2 que l'angle $(\sigma\vec{f}, \sigma\vec{f}')$ est obtus ou encore que le produit scalaire <σf , σf'> est négatif: (on peut, e.g.)

Figure 2



considérer le point r ou le rayon s0σ coupe le cercle de centre s et rayon |sh|, l'arc f r f' vaut moins d'une demi-circonférence ; donc l'angle (r f , r f') est obtus, et *a fortiori* (σf , σf').

2.3.2 On a démontré au § 2.2.3 que les points t et t' sont extérieurs au disque de centre s et rayon |sh| : on voit donc sur la figure que l'angle (σt , σt') comprend à son intérieur l'angle obtus (σf , σf') ; par conséquent (σt , σt') est obtus et le produit scalaire $<\vec{\sigma t}, \vec{\sigma t}'>$ est négatif.

2.3.3 Il est clair que le point d'intersection 0 des segments (s , s') et (t , t') appartient à l'un au moins des deux segments (s , σ) et (s' , σ) (éventuellement aux deux si σ = 0) ; donc soit en considérant le cercle de centre s et rayon |sh| , soit le cercle de centre s' et rayon |s'h| , on démontre que $<\vec{\sigma t}, \vec{\sigma t}'>$ est négatif. Dès lors t et t' ne peuvent être voisins réciproques en présence de σ.

2.4.1 En appliquant la formule de m v rappelée dans l'énoncé, il vient :

$$\text{niv}(r,t) = ((m_r\, m_t)/(m_r + m_t))\ |\, r\, t|^2 = 1/2 \ ;$$

$$\text{niv}(s\ ,\ t) = (\varepsilon^6/(1+\varepsilon^6))\varepsilon^{-2} = \varepsilon^4/(1+\varepsilon^6) \approx \varepsilon^4 ;$$

$$\text{niv}(s\ ,\ s') = (\varepsilon^{12}/(2\varepsilon^6))\,(4\,\varepsilon^{-2}) = 2\varepsilon^4 \quad ;$$

$$\text{niv}(t\ ,\ t') = (1/2)\,(4\varepsilon^2/(1+\varepsilon^2)^2) = 8\varepsilon^2/(1+\varepsilon^2)^2) \approx 8\varepsilon^2 \ ;$$

$$\text{niv}(r\ ,\ s) = (\varepsilon^6/(1+\varepsilon^6))(1+\varepsilon^{-2}) = \varepsilon^4(1+\varepsilon^2)/(1+\varepsilon^6) \approx \varepsilon^4 \ ;$$

compte tenu de la symétrie de la figure, et de ce que les masses restant les mêmes, niv s'accroît avec la distance (e.g. niv(s,t) niv(s,t')) il vient :

$$vois(r)=\{s,s'\} \; ; \; vois(s)=t \; ; \; vois(t)=s \; ; \; vois(s')=t' \; ; \; vois(t')=s' ;$$

il y a donc deux paires de plus proches voisins réciproques (s,t) et (s',t').

2.4.2 On a :

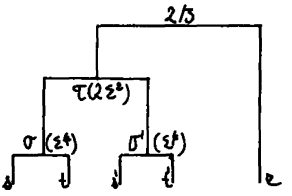
$$X(\sigma) = (m_t X(t) + m_s X(s))/(m_t+m_s)$$

$$= ((2\varepsilon/(1+\varepsilon^2))+\varepsilon^5)/(1+\varepsilon^6)$$

$$= 2\varepsilon((1+\varepsilon^5(1+\varepsilon^2)/2)/((1+\varepsilon^2)(1+\varepsilon^6))) \approx 2\varepsilon$$

$$niv(\sigma,\sigma') = (m_\sigma m_{\sigma'}/(m_\sigma+m_{\sigma'}))\,|2X(\sigma)|^2 \approx 8\varepsilon^2$$

2.4.3 La C.A.H. prend la forme ci-dessous, où auprès de chaque noeud, on a indiqué le premier terme du développement en ε de son niveau :



on a agrégé d'abord les paires de voisins réciproques reconnus au § 2.4.1. Restent alors trois sommets σ,σ', r ; comme σ diffère très peu de t aussi bien en masse qu'en position, $niv(\sigma,r) \approx niv(t,r) = (1/2)$. On agrège donc σ et σ' ; le noeud τ ainsi produit a pour masse 2 et son centre est (à des infiniment petits près) le point (0,2): d'où la valeur approchée notée pour niv(τ,r) : $(2\times 1/(2+1))\,|\tau r|^2$.

On voit sur la figure que la classe $\tau = \{t,t',s,s'\}$ a pour enveloppe convexe un trapèze qui contient r à son intérieur : donc non seulement les enveloppes convexes se coupent ; mais l'une est incluse dans l'autre.

Il reste à expliquer comment a été conçu cet exemple. On a pris trois points r,t,t' de masse 1 formant un triangle isocèle dont l'angle en ε est ≈ 4ε. Les points s et s' ont été placés sur des perpendiculaires à rt et rt' ; à une distance très grande (on aurait pu la faire tendre vers l'infini) mais en donnant à s une masse si faible que le centre de gravité de (t,s) soit presque en t (et de même pour s et s'). Les points s et s' ne peuvent s'agréger entre eux parce que l'angle (rs, rs') est obtus (cf. 1-ère partie) ; d'autre part la masse de s est si faible que le niveau d'agrégation entre t et s est très inférieur à celui entre t et t' : donc s s'agrège à t, et s' à t'. Ensuite σ s'agrège à σ' et dès lors est constituée une classe τ qui comprend le point restant r à l'intérieur de son polygone de sutentation.

On notera que l'emploi de masses très inégales (ici dans un rapport de 10^6) n'ôte rien à la généralité de l'exemple : car placer e. g. au point t une masse de 10^6, équivaut à placer 10^6 individus de masse 1 étroitement groupés autour de t. (ou coïncidant avec t).