LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
P. CAZES

Problème sur la classification

Les cahiers de l'analyse des données, tome 3, n° 1 (1978), p. 95-101

<http://www.numdam.org/item?id=CAD_1978__3_1_95_0>

Les Cahiers de l'Analyse des Données
Vol. III 1978 - n° 1 p. 95 101

PROBLÈME SUR LA CLASSIFICATION

Solution par J.-P. Benzécri (1)
et P. Cazes (2)

1 *Enoncé du problème*

On sait (cf e.g. c. mutuelles) qu'à toute classification hiérarchique binaire sur un nuage $N(I)$ est associée une décomposition de l'inertie du nuage en parts afférentes aux noeuds : en bref la part ou niveau du noeud n n'est autre que l'inertie du nuage réduit à deux points qui sont les centres de gravité des deux classes A(n) et B(n) en lesquelles se subdivise n. L'objet de la première partie du problème*est de montrer que le niveau du noeud le plus haut est toujours inférieur ou égal au plus fort moment d'inertie λ_1 , et que, plus généralement, $\lambda_1 + \lambda_2 + .. + \lambda_p$ est supérieur à la somme des niveaux des p noeuds les plus hauts. Dans la deuxième partie , on établit au contraire que le niveau du noeud le plus haut peut n'être qu'une fraction arbitrairement petite du moment d'inertie λ_1 : résultat d'où l'on pourra conclure que l'efficacité d'une classification peut être arbitrairement faible relativement à celle d'une analyse factorielle ; la réciproque n'étant jamais vraie (cf partie I).

La partie I très élémentaire, ne suppose même rien connu de l'étudiant quant à la classification hiérarchique et la décomposition de la variance. La partie II, au contraire, requiert des démonstrations d'analyse, des majorations, qu'on s'est efforcé de rendre claires et faciles par des notations explicites et un choix judicieux des paramètres.

I

Dans cette première partie, on considère un espace euclidien, E ; ou espace vectoriel muni d'une forme quadratique définie positive ; un point de E est désigné par une lettre affectée de l'indice E : x_E , y_E , i_E etc ; le vecteur d'origine x_E et d'extrémité y_E est noté $(y_E - x_E)$; le carré de norme de ce vecteur, ou carré de la distance de x_E à y_E est noté $\|y_E - x_E\|^2$; le produit scalaire de deux vecteurs sera noté $\langle u,v \rangle$; par exemple on a par définition $\|y_E - x_E\|^2 = \langle (y_E - x_E), (y_E - x_E) \rangle$.

On s'intéresse à un ensemble de points de E, munis de masse, ou nuage $N(I)$, indicé par un ensemble fini I : le point d'indice i est noté i_E, et sa masse m_i . Soit a une partie de I;on note :

(*) Ce problème a été proposé aux étudiants du D.E.A. de statistique de l'université Pierre et Marie Curie (Paris 6) à la session de Juin

(*) Ce problème a été proposé aux étudiants du D.E.A. de statistique de l'Université Pierre et Marie Curie (Paris VI) à la session de Juin 1977.

(1) Professeur de statistique à l'Université Pierre et Marie Curie

(2) Maître-assistant à l'Université Pierre et Marie Curie

$$m_a = \Sigma\{m_i \mid i \in a\} ,$$

la masse totale de cette partie ; son centre de gravité est noté a_E ; a_E peut être caractérisé par la propriété classique :

$$\Sigma\{m_i(i_E - a_E) \mid i \in a\} = 0.$$

L'inertie de la partie a est notée In(a) :

$$In(a) = \Sigma\{m_i \| i_E - a_E\|^2 \mid i \in a\} .$$

Ainsi, en particulier, on note m_I la masse totale du nuage $N(I)$; I_E est le centre de gravité de $N(I)$; In(I) est l'inertie du nuage.

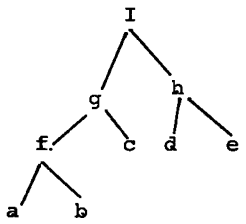
1°) Soit x_E , y_E deux points de E, affectés respectivement des masses m_x et m_y : exprimer en fonction de m_x , m_y et $\|x_E - y_E\|^2$ l'inertie du nuage formé par ces deux points.

2°) Plus généralement l'inertie In(a) d'une partie a du nuage $N(I)$ est donnée par une formule telle que :

$$In(a) = \Sigma\{r_{ii'} \|i_E - i'_E\|^2 \mid i \in a, i' \in a\} ;$$

dans cette formule la somme double comprend $(\text{Card } a)^2$ termes ; on demande de donner la valeur exacte du coefficient $r_{ii'}$ en fonction de m_i , $m_{i'}$ et de m_a . (On pourra se borner à énoncer le résultat sans démonstration).

3°) On suppose que I est muni d'un système de classes suivant le schéma ci-dessous :



{a,b,c,d,e,f,g,h} sont huit parties de I possédant les propriétés suivantes : a,b,c,d,e sont deux à deux d'intersection vide et leur réunion est I (i.e. {a,b,c,d,e} est une partition de I en cinq classes); de plus on a :

$$f = a \cup b \ ; \ g = f \cup c \ ; \ h = d \cup e \ ; \ I = g \cup h.$$

Etant données deux parties, par exemple a et b, de I, on note D(a,b) l'inertie du nuage formé des deux points a_E et b_E munis des masses respectives m_a et m_b (cf 1°) et on pose :

$$\nu(I) = D(g,h) \ ; \ \nu(h) = D(d,e) \ ;$$

$$\nu(g) = D(f,c) \ ; \ \nu(f) = D(a,b) .$$

On supposera dans la suite que : $\nu(f) < \nu(g) < \nu(I)$; $\nu(h) < \nu(I)$.

Exprimer l'inertie In(I) en fonction de $\nu(I)$, In(g), In(h).

4°) Exprimer l'inertie In(I) en fonction de $\nu(I)$, $\nu(h)$, In(g), In(e), In(d). Exprimer l'inertie In(I) en fonction de $\nu(I)$, $\nu(h)$, $\nu(g)$, $\nu(f)$, In(a), In(b), In(c), In(d), In(e).

5°) Soit Dr la droite joignant g_E à h_E. Exprimer en fonction de un ou de plusieurs des nombres $\nu(I)$, In(g), In(h) une borne inférieure de l'inertie du nuage obtenu en projetant $N(I)$ orthogonalement sur Dr ; à quelle condition cette borne est-elle atteinte?

6°) On note Pl le plan défini par les trois points g_E , d_E , e_E. Les points I_E et h_E sont-ils dans Pl? Exprimer en fonction de $\nu(I)$ et de $\nu(h)$ une borne inférieure de l'inertie du nuage obtenu en projetant $N(I)$ orthogonalement sur Pl ; à quelle condition cette borne est-elle atteinte?

7°) On rappelle que le premier axe principal d'inertie Δ_1 du nuage $N(I)$ est, parmi toute les droites passant par le centre de gravité I_E , celle sur laquelle le nuage obtenu en projetant $N(I)$, a l'inertie maxima λ_1. De même le plan engendré par les deux premiers axes d'inertie Δ_1 et Δ_2 est, de tous les plans passant par I_E , celui sur lequel le nuage obtenu en projetant $N(I)$ a l'inertie maxima $\lambda_1 + \lambda_2$; etc. Utiliser ces propriétés et les résultats de 5° et 6°, pour donner en fonction de $\nu(I)$ $\nu(h)$, $\nu(g)$, $\nu(f)$, une borne inférieure pour λ_1 ; une borne inférieure pour $\lambda_1 + \lambda_2$; une borne inférieure pour $\lambda_1 + \lambda_2 + \lambda_3$; une borne inférieure pour $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$.

II

La première partie (cf 7°) a permis de borner inférieurement λ_1 , $\lambda_1 + \lambda_2$ etc, en fonction du niveau des noeuds d'une classification hiérarchique sur I. Dans cette deuxième partie, on cherche s'il est possible de borner inférieurement les niveaux en fonction des valeurs propres. Pour cela on se place non dans un espace multidimensionnel E, mais simplement sur la droite ; et on considère une distribution de masse particulière, qui comprend une masse ponctuelle placée à l'origine, et une distribution continue s'étendant sur la demi-droite $[a,\infty]$(où $a > 0$). La distribution dépend d'un paramètre b, en plus du paramètre a qui caractérise le support ; on considère les séparations du système des masses en deux classes $[0,s]$ et $[s,\infty]$ par un point s ($s \in [a,\infty]$) ; et on joue sur les valeurs de a,b,s pour établir des inégalités qui répondent au problème qu'on se pose.

On note :

$m[0] = a^{-3-b}$, la masse ponctuelle placée en 0 ;

$m(x)dx = x^{-3-b}dx$: la densité de la distribution continue sur $[a,\infty]$; on pourra supposer, pour simplifier les majorations, que l'on a:

$0 < a < 0,1$; $0 < b < 0,1$.

Soit [u,v] une partie de la droite (intervalle fermé, ou demi-droite fermée ; i.e. origine comprise) : par exemple [0,∞], [0,s], [s,∞] etc ; nous considérerons la masse, le centre de gravité, l'inertie, de la partie du système des masses qui est portée par [u,v] ; i.e. :

m[u,v] : la masse sur l'intervalle [u,v] : e.g. $m[0,\infty] = m[0] + \int_a^\infty x^{-3-b}\,dx$

g[u,v] : l'abscisse du centre de gravité du système des masses portées par [u,v].

I([u,v];c) : l'inertie du système des masses portées par [u,v] par rapport au point c de la droite ; e.g. : $I([(a,\infty];c) = \int_a^\infty x^{-3-b}(x-c)^2\,dx$;

les quantités m[u,v], g[u,v], I([u,v];c) dépendent évidemment des paramètres a et b de la distribution de masse choisie ; mais afin d'alléger on ne note pas $m^{a,b}[u,v]$ etc...

8°) Soit $s \in [a,\infty]$, calculer :

m[s,∞] ; g[s,∞] ; I([s,∞];0) ; g[0,∞] ;

et placer les uns par rapport aux autres les trois nombres

$g[0,\infty]$, $g[0,s]$, a^2.

(par placer plusieurs nombres tels que A,B,C,D,E on veut dire : préciser les égalités et les inégalités par une formule telle que :
A < B < D = E < C)

9°) Placer les unes par rapport aux autres les cinq quantités :

I([0,∞],0) ; I([0,∞];g[0,∞]) ; I([a,∞];0) ; I([a,∞];g[0,∞]) ; $(1-a)^2$ I([a,∞];0)

10°) On partage le système de masse en deux parties : l'une [0,s] à gauche de s ; l'autre [s,∞] à droite de s. Utiliser les résultats de 8° pour majorer par une expression simple de la forme $A\,s^{-B}$ l'inertie interclasse Inter(s), ou inertie du nuage formé des deux points g[0,s] et g[s,∞] affectés respectivement des masses m[0,s], m[s,∞]. Utiliser les résultats de 9° pour majorer le rapport Inter(s)/I([0,∞];g[0,∞]). Que peut-on conclure de cette majoration relativement au problème posé en tête de II ?

2 Solution du problème

2.1 Posons a = {x,y} ; le centre de gravité a_E de a s'écrivant :

$$a_E = (m_x\, x_E + m_y\, y_E)/(m_x + m_y)$$

on a :

$$x_E - a_E = m_y(x_E - y_E)/(m_x+m_y)\ ;\ y_E - a_E = -m_x(x_E - y_E)/(m_x+m_y)$$

d'où l'on déduit :

$$\begin{aligned} In(a) &= m_x(x_E - a_E)^2 + m_y(y_E - a_E)^2 \\ &= m_x m_y (x_E - y_E)^2/(m_x + m_y) \end{aligned} \qquad (1)$$

2.2 On a :

$$In(a) = \Sigma\{(m_i\, m_{i'}/(2\,m_a))\ \|i_E - i'_E\|^2 \mid i \in a,\ i' \in a\} \qquad (2)$$

Pour démontrer cette formule, il suffit dans le deuxième membre de (2) d'écrire $i_E - i'_E$ sous la forme $(i_E - a_E) - (i'_E - a_E)$, et de développer le

carré ; l'on obtient alors en effectuant les sommations (1/2) In(a) + (1/2) In(a) + 0 = In(a) c.q.f.d.

2.3 On a :

$$In(I) = D(g,h) + In(g) + In(h) = \nu(I) + In(g) + In(h) \qquad (3)$$

En effet les égalités ci-dessus correspondent à la décomposition classique de l'inertie totale In(I) en inertie interclasse $\nu(I)$ (ou inertie associée aux centres de gravité des classes) et inertie intra-classe In(g) + In(h) (ou somme des inerties associées à chaque classe), décomposition qui se démontre aisément à l'aide du théorème de Huyghens.

2.4 Réappliquant la formule (3) non plus sur I, mais sur h, puis sur g et sur f, l'on obtient :

$$\begin{aligned} In(I) &= \nu(I) + In(g) + \nu(h) + In(d) + In(e) \qquad (4)\\ &= \nu(I) + \nu(g) + In(f) + in(c) + \nu(h) + In(d) + In(e)\\ &= \nu(I) + \nu(g) + \nu(f) + \nu(h) + In(a) + In(b) + In(c) + In(d) + In(e) \end{aligned}$$

2.5 Soit N'(I) le nuage projeté de N(I) sur la droite Dr joignant g_E à h_E, droite qui contient le centre de gravité I_E du nuage N(I).

La formule (3) appliquée au nuage N'(I) s'écrit :

$$In'(I) = \nu'(I) + In'(g) + In'(h)$$

où le prime indique qu'il s'agit d'inerties prises relativement à N'(I).

Le centre de gravité g'_E (resp. h'_E) de g (resp. h) pour le nuage N'(I) étant égal à g_E (resp. h_E), on déduit des résultats du 2.1 et compte tenu de la définition de $\nu(I)$ et $\nu'(I)$, que $\nu(I) = \nu'(I)$. On a donc:

$$In'(I) = \nu(I) + In'(g) + In'(h) \geq \nu(I)$$

l'égalité n'étant atteinte que si In'(g) = In'(h) = 0, i.e. si tout i_E de g (resp. h) se projette sur Dr en g_E (resp. h_E), ce qui revient encore à dire que tous les points i_E de g (resp. h) appartiennent à l'hyperplan affin passant par g_E (resp. h_E), et orthogonal à la droite $g_E\ h_E$.

2.6 h_E étant au centre de gravité de d_E et e_E (affectés respectivement des masses m_d et m_e) se trouve sur la droite $d_E\ e_E$, donc dans le plan P_1. De même, I_E qui est au centre de gravité de g_E et h_E appartient à P_1.

Caractérisant par l'indice seconde le nuage projeté de N(I) sur P_1, ainsi que toutes les caractéristiques associées, à ce nuage, on a, d'après (4), et compte-tenu de ce que $\nu''(I) = \nu(I)$ et $\nu''(h) = \nu(h)$ puisque $g''_E = g_E$, $h''_E = h_E$, $d''_E = d_E$, $e''_E = e_E$: $In''(I) = \nu(I) + \nu(h) + In''(g) + In''(d) + In''(e) \geq \nu(I) + \nu(h)$

l'égalité n'ayant lieu que si In''(g) = In''(d) = In''(e) = 0, soit si tout point i_E de g (resp. e ; d) se projette sur P_1 en g_E (resp. e_E ; d_E).

2.7 On a :

$$\lambda_1 \geq In'(I) \geq \nu(I)$$

$$\lambda_1 + \lambda_2 \geq In''(I) \geq \nu(I) + \nu(h)$$

et en réitérant la procédure :

$$\lambda_1 + \lambda_2 + \lambda_3 \geq \nu(I) + \nu(h) + \nu(g)$$

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 \geq \nu(I) + \nu(h) + \nu(g) + \nu(f)$$

2.8

$$m[s,\infty] = \int_s^\infty x^{-3-b}\,dx = ((2+b)\ (s^{2+b}))^{-1} \ ;$$

$$g[s,\infty] = \left(\int_s^\infty x^{-3-b}\ \ x\,dx\right) / m[s,\infty]$$

$$= ((1+b)\,(s^{1+b}))^{-1} / ((2+b)\ \ (s^{2+b}))^{-1}$$

$$= ((2+b) / (1+b))\ s \ ;$$

$$I([s,\infty];O) = \int_s^\infty x^{-3-b}\ \ x^2 dx = (b\ s^b)^{-1} \ ;$$

$$g[O,\infty] = g[a,\infty]\ x\ m[a,\infty] / m[O,\infty]$$

$$= ((2+b)/(1+b))\ a\ x\ ((2+b)\,(a^{2+b}))^{-1} / (m[a,\infty] \ + a^{-3-b})$$

$$< (1+b)^{-1}\ a^2 < a^2 \ ;$$

puisque $[O,s]$ s'obtient en amputant à droite l'intervalle $[O,\infty]$ le centre de gravité $g[O,s]$ est à gauche de $g[O,\infty]$, d'où le placement :

$$O < g[O,s] < \ g[O,\infty] < a^2.$$

2.9 Le placement est :

$$(1-a)^2\ x\ I([a,\infty];O) < I([a,\infty];g[O,\infty]) < I([O,\infty];\ g[O,\infty]) < I([O,\infty];O) = I([a,\infty];O).$$

En effet, les systèmes sur $[a,\infty]$ et $[O,\infty]$ ne différant que par la masse ponctuelle $m[O]$ placée en O, ont même inertie par rapport à O : i.e. $I\ ([O,\infty];O) = I([a,\infty];O)$. Tout système, e.g. $[O,\infty]$, a une inertie plus faible relativement à son centre de gravité $g[O,\infty]$ que relativement à tout autre point : i.e. $I([O,\infty];\ g[O,\infty]) < I([O,\infty];O)$. L'inertie d'une partie est moindre que celle du tout : $I([a,\infty];g[O,\infty]) < I([O,\infty];g[O,\infty])$. Reste à comparer les deux intégrales $I_o = I([a,\infty];\ O)$ et $I_g = I([a,\infty];\ g[O,\infty])$: on a :

$$I_o = \int_a^\infty x^{-3-b}\ x^2\ dx \ ; \ I_g = \int_a^\infty x^{-3-b}\ (x - g[O,\infty])^2 dx \ ; \text{ or :}$$

$$((x - g[O,\infty])/x) = 1 - (g[O,\infty]/x)$$

et puisque (cf 8°) $g[O,\infty]$ est compris entre O et a^2 et que $a \le x$, on a :

$$\forall\ x \in [a,\infty] \ : ((x - g[O,\infty])/x) = 1 - (g[O,\infty]/x) > 1 - a$$

d'où il reste que $I_g > (1-a)^2\ I_o$; ce qui achève de démontrer le placement annoncé.

2.10 D'après 1° et 8° on a :

$$Inter(s) = (g[s,\infty] - g[O,s])^2 \ \ m[s,\infty]\ m[O,s] / m[O,\infty]$$

$$< (g[s,\infty])^2\ x\ m[s,\infty]$$

$$< ((2+b)/(1+b))^2 \ \ s^2\ x\ ((2+b)\,(s^{2+b}))^{-1}$$

$$< (2+b)\ (1+b)^{-2}\ s^{-b} < 2s^{-b} \ ;$$

Pour l'inertie totale on a d'après 2.9 et 2.8

$$I([O,\infty]\ ;g[O,\infty]) > (1-a)^2\ I([a,\infty];O)$$

$$> (1-a)^2\ b^{-1}\ a^{-b}$$

D'où pour le rapport (en tenant compte de ce que $(1-a)^{-2} < 1,5$, car $a < 0,1$) :

$$\text{Inter}(s)/I(]0,\infty];g[0,\infty]) < 2(1-a)^{-2}\, b\, (a/s)^{b}$$

$$< 3\, b$$

Or le maximum de l'inertie interclasse (où ce qui est équivalent : le minimum de la variance intraclasse) pour une partition en deux classes d'une distribution de masse portée par la droite, s'obtient en séparant la distribution par un point s convenablement placé ; (les deux classes étant $[-\infty,s]$ et $[s,\infty]$). En effet pour toute séparation d'une autre forme, en deux classes qui s'enchevêtrent, c1 et c2 ayant pour centres de gravité g_1 et g_2 , avec e.g. g_1 à gauche de g_2 , on peut diminuer la variance intraclasse en échangeant entre c1 et c2 des points tels que x1 et x2 si x1 est à droite de x2 (enchevêtrement). Pour la distribution linéaire étudiée ici, l'inertie totale n'est autre que l'unique moment principal d'inertie non-nul $\lambda_1 = I([0,\infty];g[0,\infty])$; le niveau ν du noeud le plus haut d'une classification sur cette distribution de masse ne dépassera pas Inter(s) ; son rapport à λ_1 , sera donc nécessairement inférieur à 3 b ; c'est à dire arbitrairement petit.