LES CAHIERS DE L'ANALYSE DES DONNÉES

I. KHARCHAF
R. ROUSSEAU

Reconnaissance de la structure de blocs d'un tableau de correspondance par la classification ascendante hiérarchique

Les cahiers de l'analyse des données, tome 13, n° 4 (1988), p. 439-443

<http://www.numdam.org/item?id=CAD_1988__13_4_439_0>

Les Cahiers de l'Analyse des Données
Vol XIII - 1988 - n° 4 - pp. 439-443

RECONNAISSANCE DE LA STRUCTURE DE BLOCS D'UN TABLEAU DE CORRESPONDANCE PAR LA CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

[REC. BLOC. CAH]

*I. KHARCHAF**
*R. ROUSSEAU***

1 Position du problème :

Par structure de blocs, nous entendrons ici, la décomposition en blocs diagonaux la plus fine, laquelle, comme on le sait, existe et est unique. De façon précise, pour un tableau de correspondance sur I×J, on dit que deux partitions de I et J, indicées par un même ensemble B,

$$I = \cup\{I_b |\ b \in B\}\ ;\ J = \cup\{J_b |\ b \in B\}\ ;$$

définissent une partition en blocs diagonaux compatible avec un tableau k donné si:

$$\forall\ (i,j) \in I \times J : (k(i,j) \neq 0) \Rightarrow \exists\ b : (\ i \in I_b ;\ j \in J_b)\ .$$

Ceci posé, si on effectue sur l'un des deux ensembles en correspondance, (que nous supposerons désormais être I), une classification ascendante hiérarchique, (par quoi on entendra toujours, dans la suite, CAH avec pour critère, celui de l'agrégation suivant la variance des partitions, la métrique étant la distance du chi2), nous dirons que la structure de blocs a été reconnue par la CAH, si chacun des sous ensembles I_b est une classe de la hiérarchie créée ; auquel cas,toutes les classes créées sont, soit des parties d'un sous ensemble I_b convenable, soit une réunion de tels sous ensembles. Nous désignerons par

(*) Faculté des Sciences, Département de Mathématiques et Informatique, B.P. 1014, RABAT.
(**) Université Catholique de l'Ouest, I.M.A., B.P. 808, 49005 ANGERS Cedex.

(HB) la propriété (à démontrer!) que toute structure de blocs est reconnue par la CAH.

On sait que les niveaux d'agrégation des nœuds sont majorés par les valeurs propres issues de l'analyse du tableau de correspondance; donc, en particulier, inférieurs à 1, cette valeur n'étant prise que si, précisément, il y a une décomposition non triviale en blocs diagonaux. La multiplicité de la valeur propre 1 est exactement card B -1; de plus, comme on pourra le vérifier sur les calculs de critère effectués dans la suite, la distance, (pseudo distance), crit(I_b,$I_{b'}$) entre deux des sous ensembles de la partition B, vaut 1.Donc, si la partition est reconnue par la CAH, on a, au sommet de la hiérarchie, une partition définie par les card B -1 nœuds les plus hauts qui n'est autre que la partition B. Cette disposition s'accorde avec le fait que toutes les agrégations se font au sein d'un sous ensemble I_b, excepté celles entre sous ensembles.

2 Distance relative au tableau et distance relative à un bloc:

Afin de faciliter la lecture des formules, nous convenons désormais de ne plus écrire b en indice; et noterons Ib, Jb pour: I_b, J_b.

Tout élément i de I peut être considéré relativement au bloc Ib(i)×Jb(i) auquel il appartient, ou relativement au tableau I×J tout entier. Ainsi, i possède, à la fois, un profil sur Jb(i) et un profil sur J tout entier: les composantes de ce dernier profil sont nulles, en dehors de celles indicées par un j de Jb(i), qui ont mêmes valeurs que celles du profil sur Jb(i). De même, i a une fréquence relative au tableau; et une autre relative au bloc b(i). Il est commode d'utiliser la lettre P majuscule quand on se réfère au tableau tout entier; et la lettre p minuscule, quand on se réfère au bloc. On écrira donc, pour les fréquences:

$$P_i = p_i \,.\, P_{Ib(i)} \,;$$

i.e.: la fréquence, (ou masse), de i relativement au tableau entier est le produit de la masse de i relativement à son bloc, par la masse de ce bloc relativement au tout. Des formules analogues existent pour J.

Pour la distance entre deux éléments i et i' d'un même sous ensemble Ib, on a la formule:

$$|P_I^i, P_I^{i'}|^2 = |p_{Ib}^i, p_{Ib}^{i'}|^2 / P_{Ib} \,; \quad \text{où : } D2(i,i') = d2(i,i')/Pb.$$

Autrement dit, la distance relative au tout est plus grande; parce que, en bref, les composantes des profils étant les mêmes dans les deux cas, on a, relativement au tout, comme dénominateurs des carrés des différences des composantes, des masses des j plus petites.

Pour le calcul du critère d'agrégation entre deux éléments i et i', (ou, plus généralement entre deux parties s et s') d'un même sous ensemble, les effets sur les masses et les distances se compensent, et on peut écrire:

$Crit(i,i') = crit(i,i')$.

Entre deux éléments i et i', (ou deux parties s et s'), inclus dans deux sous ensembles différents b et b', les calculs de distance et de critère sont plus complexes; car il faut tenir compte de composantes non nulles à la fois dans b et dans b'. Voici les résultats:

$D2(i,i') = ((1+d2(i,Ib))/Pb) + ((1+d2(i',Ib'))/Pb')$;

dans cette formule, comme plus haut, Pb est la masse du sous ensemble b relative au tout; Ib désigne le centre de gravité de ce sous ensemble, assimilé au profil marginal du bloc b sur Jb; le dénominateur Pb s'explique comme précédemment; le 1 s'introduit suivant la formule générale:

$$\Sigma\{(p_j^i)^2/p_j \mid j\in J\} = 1+ \Sigma\{(p_j^i-p_j)^2/p_j \mid j\in J\}.$$

On a de même pour le critère:

$$Crit(i,i') = ((1+d2(i,Ib))/Pb) + ((1+d2(i',Ib'))/Pb')/((1/Pi)+(1/Pi'))$$

$$= (pi(1+d2(i,Ib))(1/Pi) + pi'(1+d2(i',Ib))(1/Pi'))/((1/Pi)+(1/Pi'));$$

formule où on a omis de mettre i et i' en indices; et où on a tenu compte de ce que Pi = pi.Pb.

D'autre part, une expression de Crit(i,i'), pour i et i' éléments de I, ne fait intervenir que les variables j des blocs auxquels appartiennent i et i'. Il en résulte que le problème (HB) se réduit à la reconnaissance d'une partition en deux blocs (cardB = 2).

3 Équivalence entre la reconnaissance des blocs par la CAH et une propriété d'un tableau de correspondance I×J usuel:

Sous la forme que nous lui avons donnée, Crit(i,i'), (pour i et i' dans deux sous ensembles différents), se prête à une minoration ne mettant en jeu, finalement, que les propriétés d'un seul bloc. En effet, pour quatre nombres positifs quelconques, x, x', y, y', on a la relation:

$(x + x')/(y + y') \in ((x/y),(x'/y'))$;

autrement dit, un quotient dont le numérateur et le dénominateur sont des sommes de deux termes est compris dans l'intervalle délimité par les quotients des termes de même rang; et l'égalité avec l'un de ces quotients n'a lieu que si ceux-ci sont égaux. Les notations adoptées réduisent notre assertion à une

évidence: il suffit de considérer, dans le plan,le parallélogramme ayant pour sommets les 4 points:

(0,0) ; (x,y) ; (x',y') ; (x+x',y+y');

et de voir que la diagonale issue de l'origine a une pente comprise entre celles des deux côtés issus de ce point.

On applique la relation à Crit(i,i') en posant y=1/Pi et y'=1/Pi'; et il vient:

Crit(i,i') ∈ (pi(1+d2(i,Ib)),pi'(1+d2(i',Ib')).

Il faut maintenant rappeler la propriété (HB), de la CAH, que nous avons en vue : nous désirons que i ne puisse s'agréger à i', sous l'hypothèse b(i)≠b(i'). Il suffit pour celà qu'il existe dans Ib(i) un élément i" tel que Crit(i,i") soit inférieur strict à Crit(i,i'); ou encore qu'il existe dans Ib(i') un élément i'_1 tel que Crit(i',i'_1)<Crit(i,i'). Compte tenu de la relation vérifiée par Crit(i,i'), il nous suffirait que la propriété suivante, désignée dans la suite par (G), fût vraie en toute généralité pour un ensemble I en correspondance (non décomposable en blocs) avec un ensemble J:

$$\forall i \in I, \exists\, i_1 \in I : crit(i,i_1) < p_i\,(1+d^2(i,I)) \qquad (G);$$

dans cette formule, d2 est la distance du chi2; et crit est calculé comme plus haut.

Dans le cas où cardI = 1, $p_i.(1+d^2(i,I)) = 1$; et la propriété (G) est vérifiée. Dans le cas où CardI = 2, la propriété est vérifiée si et seulement si le tableau n'est pas décomposable en blocs. On montre en effet que pour I = {i, i'}, l'inégalité crit(i,i') < $p_i\,(1+d^2(i,I))$ est vérifiée si et seulement si $d^2(i,I) < (1/p_i)-1$; et la dernière inégalité est vraie si et seulement si le tableau n'est pas décomposable en blocs. Des difficultés non résolues se présentent à partir du cas CardI = 3.

On peut, plus explicitement écrire:

$$\forall i \in I, \exists\, i_1 \in I : ((p_i \,.\, p_{i1})/(p_i + p_{i1})) \,.\, |p_J^{\,i} - p_J^{\,i1}|^2 < p_i \,.\, (1 + |p_J^{\,i} - p_J|^2) \qquad (G);$$

ou, en divisant les deux membres par p_i :

$$\forall i \in I, \exists\, i_1 \in I : (p_{i1}/(p_i + p_{i1})) \,.\, |p_J^{\,i} - p_J^{\,i1}|^2 < (1 + |p_J^{\,i} - p_J|^2) = |p_J^{\,i}|^2 \qquad (G);$$

Il importe de voir que, considérées dans toute leur généralité, (i.e. comme valant pour toutes correspondances), les propriétés (HB) et (G) sont équivalentes. On vient de voir que (G) suffit à établir (HB). Réciproquement, une exception à (G) permet de construire une exception à (HB). De façon précise, si il y a exception à (G) pour un élément i relativement à un tableau de correspondance I×J convenable, on construira un tableau de correspondance (I1+I2)×(J1+J2) formé de deux blocs diagonaux identiques à ce tableau; les chiffres 1 et 2 servant à désigner les deux exemplaires d'un même ensemble ou d'un même élément. Il y aura, dans chacun des deux sous ensembles I1 et I2 de (I1+I2), un exemplaire, i1 ou i2, de l'élément i; et on aura:

$$\mathrm{Crit}(i1,i2) = p_i\,(1+d2(i,I)) \;\; ; \;\; \forall i_{a1} \in I1 : \mathrm{Crit}(i1,i2) \leq \mathrm{Crit}(i_{a1},i1)\,.$$

Si l'on procède à la CAH sur I1+I2, une solution possible de la CAH se fait avec, lors d'une étape, agrégation entre les éléments i1 et 12; et ainsi la propriété (HB) de la CAH n'est pas vérifiée. En effet, i1 peut s'agréger avec i2, comme il peut le faire éventuellement avec un autre élément i_{a1} de I1, ou une classe de tels éléments. Car un i_{a1} est, par hypothèse, plus éloigné de i1, (au sens du critère), que ne l'est i2; et la même inégalité vaut pour une classe construite par agrégation de tels éléments, en vertu de l'axiome de la médiane: e.g. si i_{a1} et i_{b1} s'agrègent en présence de i1, la distance de i1 à $(i_{a1} \cup i_{b1})$ est supérieure ou égale à la plus petite des deux distances $\mathrm{Crit}(i1,i_{a1})$ et $\mathrm{Crit}(i1,i_{b1})$ donc à Crit(i1,i2).

Nous concluons en rappelant que la démonstration d'équivalence, qui fait l'objet de la présente note, laisse à démontrer en toute généralité la propriété (HB), selon nous très vraisemblable.