CONSTRUCTION ASCENDANTE HIERARCHIQUE RAPIDE : LE PROGRAMME CAHVR

M. Roux
CEPE-CNRS
B P 5051
34033 Montpellier-Cedex

Résumé :
Cet article décrit les bases théoriques et donne des recommandations pratiques d'utilisation du programme CAHVR. Ce programme construit la hiérarchie du moment d'ordre deux par la méthode des voisins réciproques.

Mots-clés :
Construction ascendante hiérarchique, moment d'ordre deux, voisins réciproques.

1.- Introduction

Le programme CAHVR construit une hiérarchie par agrégations successives, selon le critère du moment d'ordre deux d'une partition. Son originalité tient en deux points :

- il travaille directement sur le tableau rectangulaire des données, et ne calcule les distances qu'au fur et à mesure des besoins.

- il utilise l'algorithme rapide "des voisins réciproques".

Ces deux particularités lui permettent de traiter en un temps raisonnable de grands tableaux de données quantitatives, à condition, toutefois, que ces tableaux tiennent en mémoire centrale de l'ordinateur.

Dans les présentes notes nous donnons d'abord quelques rappels théoriques, à la fois sur la méthode du moment d'ordre deux et sur le principe des voisins réciproques (Parag. 2), puis nous examinons les modalités pratiques d'utilisation de ce programme (Parag. 3). Enfin nous concluons (Parag. 4) en rappelant les principales caractéristiques de ce programme.

2.- Rappels théoriques

2.1.- Agrégation par le moment d'ordre deux

Le moment centré d'ordre deux d'un ensemble I d'éléments i, i', i" ... est défini par :

$$M^2(I/G) = \Sigma_i \, m_i \, d^2(i,G) \qquad (1)$$

où m_i désigne la pondération, ou masse, affectée au point i, et G le centre de gravité de l'ensemble I; $d^2(i,G)$ représente le carré de la distance de i au centre de gravité. Lorsqu'on calcule le moment d'ordre 2 non centré, celà signifie qu'on remplace le point G par un point P quelconque. On peut démontrer le théorème de Huyghens :

$$M^2(I/P) = M^2(I/G) + m \, d^2(P,G) \qquad (2)$$

Cette formule, dans laquelle m désigne la masse totale de l'ensemble I, montre que tout autre point que G fournit un moment d'ordre 2 plus élevé, ou encore que G réalise le minimum des moments d'ordre 2. Lorsqu'on parle de moment d'ordre 2, sans autre précision, il est sous-entendu que le point de référence est le centre de gravité.

Le moment d'ordre 2 est une sorte de généralisation de ce qu'est la variance dans le cas univarié, à ceci près que l'on ne divise pas par la masse totale, masse qui n'est autre que l'effectif quand les masses individuelles sont égales à l'unité (absence de pondération). Le théorème de Huyghens permet à son tour d'établir la formule suivante sur laquelle repose la construction ascendante du moment d'ordre 2 :

$$M^2(qUq') = M^2(q) + M^2(q') + [m_q m_{q'}/(m_q + m_{q'})] \, d^2(q,q') \qquad (3)$$

Dans cette formule q et q' représentent des individus, ou des groupes d'individus, candidats pour une fusion donnant le groupe qUq' (réunion de q et de q'); m_q et $m_{q'}$ désignent les masses de ces deux groupes (somme des masses de leurs individus). Cette écriture montre que la réunion de deux groupes a un moment d'ordre 2 supérieur à la somme des moments de ces deux groupes. En

analyse univariée on dirait que la variance totale est égale à la somme des variances intra-groupes augmentée de la variance inter-groupe. De la même façon le dernier terme de (3), dans lequel $d^2(q,q')$ désigne le carré de la distance entre les centres de gravité des deux classes q et q', représente le moment inter-groupe.

Le principe de la méthode consiste précisément à agréger, à chaque pas de l'algorithme, les deux classes pour lesquelles ce moment inter-groupe est minimum, de façon à constituer une nouvelle classe aussi homogène que possible. C'est donc la quantité :

$$D(q,q') = [m_q m_{q'}/(m_q + m_{q'})]\, d^2(q,q')$$

qui tient lieu de critère d'agrégation et que nous appelons *pseudo-distance* entre les classes q et q'.

2.2.- L'algorithme des voisins réciproques

Cet algorithme repose sur la condition suivante que doit vérifier la pseudo-distance :

$$D(qUq', k) \geq \text{Min}\,[D(q,k), D(q',k)] \qquad (4)$$

elle signifie que la pseudo-distance entre un groupe k et une classe nouvellement formée, qUq', doit être plus grande que la plus petite des pseudo-distances entre k et l'une ou l'autre des anciennes classes q et q'.

On appelle *voisins réciproques* des paires de classes q et q' pour lesquelles q est la classe la plus proche de q' et q' est la classe la plus proche de q, au sens de la pseudo distance D. Si la propriété (4) est vérifiée, alors on montre que toute paire de voisins réciproques constitue obligatoirement un noeud de la hiérarchie, quelle que soit la valeur de la pseudo-distance entre ces classes. On profite donc de cette observation pour agréger, à chaque étape de l'algorithme ascendant, toutes les paires de voisins réciproques, au lieu de la seule paire qui présente la plus petite pseudo-distance.

Montrons que la condition (4) permet bien de faire celà. Si elle est vérifiée alors l'agrégation de q et q' ne modifie pas le fait que deux autres classes k et k' soient en situation de voisins réciproques. En effet si k et k' sont mutuellement plus proches voisins on a :

$$D(k,k') \leq D(q,k) \quad \text{et} \quad D(k,k') \leq D(q',k)$$

et, d'après (4) il résulte que :

$$D(k,k') \leq D(qUq',k) \quad \text{et} \quad D(k,k') \leq D(qUq',k')$$

c'est à dire que la formation du groupe qUq' ne change pas le fait que D(k,k') est la plus petite des pseudo-distances relatives à k et k'.

Reste à démontrer que la pseudo-distance utilisée dans la méthode du moment d'ordre 2 vérifie la condition (4). Celà se fait à l'aide de la formule de récurrence qu'on utilise habituellement lorsqu'on travaille sur la matrice des pseudo-distances (Cf. Benzécri et coll. 1973) :

$$D(k,qUq') = [(m_k + m_q)D(q,k) + (m_k + m_{q'})D(q',k) - m_k D(q,q')]/(m_k + m_q + m_{q'})$$

En effet, si l'on agrège q et q' c'est que l'on a à la fois :

$$D(q,q') \leq D(q,k) \quad \text{et} \quad D(q,q') \leq D(q',k)$$

on peut donc remplacer D(q,q') par [D(q,k) + D(q',k)]/2 dans la formule précédente, ce qui donne :

$$D(qUq',k) \geq [(m_q + m_k/2)D(q,k) + (m_{q'} + m_k/2)D(q',k)]/(m_k + m_q + m_{q'})$$

La nouvelle pseudo-distance D(qUq',k) apparait donc supérieure à une moyenne pondérée des deux précédentes pseudo-distances D(q,k) et D(q',k), elle est donc supérieure à la plus petite des deux, ce qui est la condition (4).

3.- Considérations pratiques

3.1.- Particularités de la programmation actuelle de CAHVR

Comme nous l'avons annoncé le programme travaille directement sur le tableau rectangulaire des données. Pour celà, après chaque agrégation on remplace les 2 lignes agrégées par les coordonnées du centre de gravité de ces deux lignes. Lorsqu'on a besoin de la distance inter-groupe on la calcule par la formule euclidienne usuelle :

$$d^2(q,q') = \Sigma_j (x_{qj} - x_{q'j})^2 \quad (6)$$

ou x_{qj} représente la j-ème coordonnée du centre de gravité de la classe q

3.2.- Conséquences pratiques

Les dispositions précédentes ont deux conséquences évidentes : le tableau des données est limité par la taille de la mémoire disponible, et les données doivent être quantitatives pour que les notions de centre de gravité et de moment d'ordre deux aient un sens.

La limitation de la taille du tableau est, en fait, peu contraignante même sur micro-ordinateur. En effet un PC standard avec 640 K-octets de mémoire centrale laisse environ 400 K-octets utiles, une fois chargés le système d'exploitation et le programme, soit un tableau de 100.000 cases. Avec une vingtaine de variables, par exemple, celà autorise théoriquement un échantillon de 5000 observations !

La deuxième condition, qui s'applique à tout programme basé sur le moment d'ordre deux, est beaucoup plus contraignante. En effet pour que les distances ne soient pas biaisées par les unités choisies pour les différentes variables, il est nécessaire que les données aient reçu une forme de normalisation préalable. La plus classique, sinon la plus naturelle, consiste à réduire les données, en divisant les mesures relatives à une variable par l'écart-type de cette variable (le centrage ne change rien car les coordonnées n'interviennent que par leurs différences).

Voyons maintenant ce que l'on peut faire lorsque les données ne sont pas quantitatives mais sont toutefois passibles de l'analyse factorielle des correspondances. Nous entendons par là les données purement qualitatives (sous forme disjonctive) ou bien les tableaux de contingence (comptages, et plus généralement les tableaux de grandeurs additives). Pour de telles données la distance à utiliser devrait être la métrique du khi-carré :

$$d^2(i,i') = \Sigma_j (x_{ij}/x_{i.} - x_{i'j}/x_{i'.})^2/x_{.j} \qquad (7)$$

$x_{i.}$ et $x_{.j}$ représentant, respectivement, la somme des termes de la ligne i et de la colonne j

Il est possible de faire une transformation préalable des données pour obtenir, avec la distance euclidienne usuelle (Formule 6), les mêmes résultats que si on avait utilisé la formule du Khi-carré (7) Il suffit de remplacer les valeurs brutes x_{ij} par :

$$y_{ij} = x_{ij} / (x_{i.} \text{Rac}(x_{.j}))$$

où l'écriture Rac() se lit "Racine carrée de" (Cf. Lebart, Morineau, Tabart 1977).

Une autre possibilité, également valable, est de réaliser l'analyse factorielle des correspondances du tableau brut, puis d'appliquer le programme au tableau des coordonnées factorielles issues de cette analyse. On sait, en effet, que la distance euclidienne usuelle sur ces coordonnées est approximativement égale à la distance du Khi-carré sur les données brutes. L'approximation est d'autant meilleure qu'on utilise un plus grand nombre d'axes factoriels.

Le nombre d'axes à retenir constitue précisément le problème à résoudre dans cette stratégie qui a par ailleurs de nombreux avantages (en particulier la stabilité des résultats, Cf. Roux 1985). Selon nous deux critères sont à prendre en compte pour le choix du nombre d'axes : l'interprétabilité des graphiques correspondants, et la décroissance des pourcentages d'inertie. On conserve les premiers axes factoriels jusqu'à ce que cette décroissance devienne "négligeable". En aucun cas la somme des pourcentages extraits n'est à prendre en considération puisque celle-ci dépend de la taille du tableau (Cf. Lebart, Morineau, Tabart 1977).

3.3.- Autres remarques pratiques

Les niveaux successifs des noeuds de la hiérarchie fournie par CAHVR ne sont pas forcément en ordre croissant. En effet l'algorithme agrège les paires de voisins réciproques même si ceux-ci sont relativement éloignés, alors qu'un certain nombre de groupes, de niveaux moins élevés, ne peuvent se former car leurs sous-groupes n'ont pas encore été créés.

En ce qui concerne les temps de calcul, J.P. Benzécri a démontré que, si n est le nombre d'objets, alors le temps est de l'ordre de n^2, à nombre de variables constant (Cf. Benzécri, 1982). Celà est vérifié dans les expériences de Juan (Cf. Juan 1982) dont le programme présente, à peu de choses près, les mêmes caractéristiques que CAHVR. Nous avons également fait des essais avec les fameux IRIS de Fisher, qui comportent 4 variables, et en augmentant progressivement le nombre d'observations. Voici les résultats obtenus sur un micro-ordinateur de type AT cadencé à 8 Mhz :

50 obs. : 8" ; 100 obs : 18" ; 150 obs 35"

Ces valeurs sont fournies à titre indicatif car le temps de lecture des données est inclus.

4.- Conclusion

Rappelons les principales caractéristiques de ce programme Son avantage certain réside dans sa vitesse de calcul, qui autorise le traitement

d'effectifs importants, sans perte de qualité par rapport à l'algorithme usuel du moment d'ordre deux. Ses contraintes sont qu'il ne peut traiter que des données quantitatives (comme tout algorithme de ce type) et que celles-ci divent pouvoir se loger en mémoire centrale de l'ordinateur.

BIBLIOGRAPHIE

Benzécri J.P. et coll. 1973 - L'analyse des données. Tome 1: La Taxinomie. Dunod, Paris, 615p.

Benzécri J.P. 1982 - Sur la proportion des paires de voisins réciproques pour une distribution uniforme dans un espace euclidien. Cahiers de l'Analyse des Données, Vol. 7, 2, pp185-188.

De Rham C. 1980.- La classification hiérarchique selon la méthode des voisins réciproques. Cah. Ana. des données, Vol 5, 2, pp 135-144.

Juan J. 1982.- Le programme HIVOR de classification ascendante hiérarchique selon les voisins réciproques et le critère de la variance. Cahiers Ana. des données, VOL. 7, 2, pp 173-184.

Lebart L., A. Morineau et N. Tabart 1977.- Techniques de la description statistique, méthodes et logiciels pour l'analyse des grands tableaux. Dunod, Paris.

Roux M. 1985 - Algorithmes de classification. Masson, Paris, 151p.