LES CAHIERS DE L'ANALYSE DES DONNÉES

A. EL MOUSSAOUI

Qualité de la représentation des classes d'une CAH sur un tableau construit en cumulant des blocs de variables

Les cahiers de l'analyse des données, tome 12, n° 2 (1987), p. 237-242

<http://www.numdam.org/item?id=CAD_1987__12_2_237_0>

Les Cahiers de l'Analyse des Données
Vol. XII - 1987 - n° 2 pp. 237-242

QUALITE DE LA REPRESENTATION DES CLASSES D'UNE CAH SUR UN TABLEAU CONSTRUIT EN CUMULANT DES BLOCS DE VARIABLES

[CAH VAR CUM]

A. El Moussaoui (*)

1 Rappel : QLT et CO2 dans les listages d'aide à l'interprétation

On utilise communément pour l'interprétation d'une CAH, deux listages désignés par les sigles FACOR et VACOR. Dans les deux listages, on considère une partition en n classes extraite de la CAH; la partition définie en coupant l'arbre à un niveau déterminé, et donc avec au dessus des classes une structure hiérarchique dont les noeuds ne sont autres que les (n-1) noeuds les plus hauts de la CAH; ou, plus généralement, partition définie en retenant certains noeuds avec pour seule restriction que si un noeud est retenu, (si ce noeud n'est pas le sommet), son père et son frère, doivent l'être également.

Le listage, (FACOR OU VACOR), comprend 3 parties: un tableau relatif aux profils des (n-1) noeuds considérés comme classes de la hiérarchie; un tableau relatif aux n classes de la partition retenue; et enfin un tableau des (n-1) dipôles associés à chaque noeud.

Sans reprendre un exposé déjà souvent fait avec plus ou moins de détails (cf. Eco Ling Méd) nous rappellerons, que le tableau FACOR des classes ne diffère en rien d'un tableau d'éléments supplémentaires adjoint à une analyse de correspondance; le tableau principal étant ici le tableau des individus i de l'ensemble I soumis à la CAH. On a pour chaque classe des coordonnées, ou valeurs des facteurs; des CO2, ou $\cos^2$ des angles fait avec l'axe par le rayon joignant le profil de c à l'origine (au centre de gravité de I); des CTR (contribution relative à l'inertie du nuage projeté sur l'axe). La qualité QLT de la représentation étant la somme des CO2 afférents aux axes factoriels retenus; (le plus souvent 7; sans qu'il y ait d'autres raisons à ce choix précis que la place disponible sur une ligne!).

(*) Docteur 3° cycle en statistique.

En termes géométriques, la qualité QLT n'est autre que le $\cos^2$ de l'angle fait par le rayon $f_J f^c_J$ avec le sous espace L engendré par les axes factoriels retenus; ou encore c'est une fraction dont le dénominateur est le carré du rayon; et dont le numérateur est le carré de la projection du rayon sur L.

Pour les dipôles, on procède de façon semblable, à la seule différence qu'il s'agit non du rayon joignant le profil d'une classe à l'origine, mais du "dipôle" ayant pour extrémité les profils des deux descendants A(n) et B(n) d'un même noeud n.

Le listage VACOR est une généralisation de FACOR, en ce sens qu'au lieu de considérer l'espace des profils rapporté au système orthonormé des axes factoriels (plus exactement le sous espace L engendré par les axes retenus), on considère le système orthogonal de coordonnées constitué par les variables elles-mêmes; de façon précise, à la formule :

$$\|f^c_J - f_J\|^2 = \Sigma\{(F_\alpha(c))^2 \mid \alpha = 1,...\};$$

correspond la formule :

$$\|f^c_J - f_J\|^2 = \Sigma\{(f^c_j - f_j)^2/f_j \mid j \in J\} ;$$

Entre cette deuxième formule et la première on note toutefois plusieurs différences :

1°) le centre de gravité du nuage, a des coordonnées nulles sur chacun des axes factoriels; tandis que les composantes f_j de son profil ne sont pas nulles.

2°) relativement à la distance du χ^2, les facteurs constituent un système non seulement orthogonal mais orthonormé; c'est-à-dire que non seulement il n'y a pas dans la distance de termes rectangles (e.g. en F_1 (c) F_2 (c)), mais encore les carrés $(F_\alpha(c))^2$ ont tous pour coefficient 1; tandis que les composantes du profil constituent un système orthogonal mais non orthonormé: la présence des coefficients $(1/f_j)$ étant requise par le principe d'équivalence distributionnelle.

3°) initialement, on avait prévu de pouvoir ne prendre en compte dans le listage VACOR qu'une partie des variables (de même que le listage FACOR est restreint à quelques facteurs); mais cette possibilité a été peu utilisée; en sorte qu'il n'y a pas de projection sur un "espace L", mais représentation dans l'espace tout entier, avec, par conséquent, une qualité toujours égale à 1 (1000 sur le listage, où les valeurs sont exprimées en millièmes).

2 Nécessité de cumuler les variables par blocs

Pourquoi conserve-t-on toutes les composantes f^c_j sur le listage VACOR? Parce qu'on ne sait lesquels éliminer; l'expérience ayant souvent montré que l'explication d'une dichotomie intéressante se faisait par une variable qu'on avait

écartée *a priori* ... Cependant la multiplicité des variables est gênante non seulement parce que le tableau de VACOR s'étale sur des colonnes dont certaines sont inutiles, mais encore parce que souvent aucune de ces colonnes ne donne avec un CO2 élevé l'explication de l'écart entre f^c_J et f_J (ou encore, s'il s'agit d'un dipôle, entre $f_J^{A(n)}$ et $f_J^{B(n)}$). En effet même si la somme des CO2 est 1000 (QLT = 1000), ces termes étant nombreux, il se peut qu'aucun d'eux ne dépasse 100 (i.e. 1/10).

Du point de vue de l'interprétation on conçoit par exemple que tandis que l'importance relative des dépenses alimentaires caractérise une classe de consommateurs, aucun des postes de ces dépenses (éclatées suivant une nomenclature fine), n'apporte de contribution massive. Cet exemple explique que dans bien des cas, la clé de l'interprétation naît dans une agrégation des variables: en cumulant par blocs q les colonnes j du tableau initial I x J, on réduit, assurément, l'étendue du listage VACOR; et si les classes q ont été bien choisies, l'interprétation apparaît. Il va sans dire que le choix des classes, résulte naturellement de la CAH effectuée sur l'ensemble J d'après le même tableau de base I x J.

Mais ici une question se pose: quelle est la qualité de la représentation de la CAH sur I, par le tableau I x Q (où Q désigne la partition de J en un système de classes q)? Une première réponse peut être donnée en considérant le tableau I x Q sans référence au tableau I x J (à partir duquel celui-là a été construit): un individu i est une ligne; le cumul des lignes afférentes aux individus i d'une classe c ($c \in I$)) a un profil f_Q^c dont la distance au profil moyen f_Q s'exprime par la formule usuelle :

$$\|f_Q^c - f_Q\|^2 = \Sigma\{f_q^c - f_q)^2/f_q \mid q \in Q\} ;$$

dans cette formule chaque terme q apporte une contribution; soit:

$$CO2_q = ((f_q^c - f_q)^2/f_q)/ \|f_Q^c - f_Q\|^2.$$

Et comme précédemment, la qualité, somme des $CO2_q$ vaut toujours 1000 millièmes.

Cette conclusion réconfortante laisse pourtant planer un doute. Du point de vue de l'interprétation, il est indubitable que la partition Q (dont l'origine n'est aucunement prise en compte dans le calcul des $CO2_q$ proposé ci-dessus) pourrait avoir été très mal choisie: agrégeant, pour reprendre l'exemple évoqué ci-dessus, des dépenses alimentaires de première nécessité avec des dépenses de luxe... Une estimation numérique de la QLT devrait permettre d'apprécier l'adéquation de la partition Q retenue (sur J), à la partition C (sur I) qu'on doit expliquer. Au fond adopter une partition Q sur J, ou retenir quelques facteurs issus de l'analyse du tableau I x J sont des démarches analogues: dans les deux cas, que l'on cumule des colonnes ou que l'on élimine certains facteurs, on écarte une partie de l'information initialement disponible, en se fiant à l'analyse des données pour le choix de l'information prise en compte. Si la représentation d'une classe c ($c \in I$) dans l'espace des p premiers facteurs a une qualité QLT généralement inférieure à

1000, il doit en être de même pour la représentation d'une classe c après cumul des variables j suivant une partition Q.

Une formule se présente naturellement:

$$CO2_q = ((f_q^c - f_q)^2/f_q)/ \|f_J^c - f_J\|^2 ;$$

d'où pour la qualité:

$$QLT = \Sigma\{CO2_q \mid q \in Q\} = \|f_Q^c - f_Q\|^2/ \|f_J^c - f_J\|^2 .$$

Pour que cette formule réponde à notre attente, il faut que QLT ainsi définie soit toujours 1; et il est souhaitable qu'on retrouve l'interprétation géométrique du § 1: le numérateur de la fraction QLT n'est autre que le carré de la distance à l'origine de la projection du profil f_J^c de la classe c sur un certain " sous espace L" de R_J associé à la partition Q, comme l'espace engendré par les p premiers axes est associé au choix des p premiers facteurs. L'objet du § 3 est de montrer qu'une telle représentation existe en effet.

3 Cumul des variables par blocs et projection orthogonale

En général, une projection p d'un espace vectoriel X sur un de ses sous espaces L, est une application linéaire p de X dans lui-même telle que

1°) $x \in X : p(x) \in L$;

2°) $x \in L : p(x) = x$;

i.e. l'image de de tout x par p est dans L; et tout élément de L est invariant par p. Ainsi L est l'image p(X) de l'application linéaire p; et le noyau N de p (un ensemble x tels que p(x) = 0) est un supplémentaire" de L; tout vecteur x s'écrit de manière unique sous la forme d'une somme de deux composantes:

$$x = x_L + x_N \; ; \; x_L \in L \; ; \; x_N \in N ;$$

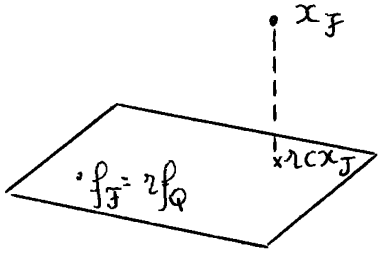
$$x_L = p(x) \; ; \; x_N = x - p(x).$$

On parle de projection orthogonale, si relativement à une métrique définie positive (associée à un produit scalaire bilinéaire sur X) les sous espaces N et L sont orthogonaux entre eux; ou encore si quel que soit x, le vecteur x-p(x), différence entre x et sa projection sur L est orthogonal à L, c'est-à-dire à tout vecteur de L.

Pour rejoindre ce schéma dans le cas présent, il faut que le profil f_Q^c d'une classe c soit identifié à un profil sur J, généralement distinct de f_J^c, qui serait la projection $p(f_J^c)$. A la vérité ici le fait que f_Q^c soit le profil d'une classe c (c'est-à-dire le profil d'une ligne obtenue en faisant la somme des lignes afférentes aux individus $i \in c$) importe peu. Seul compte le fait que partant d'une distribution de masse x_J sur J on construit, par cumul suivant la partition Q de J une distribution de masse x_Q sur Q:

$q \in Q : x_q = \Sigma\{ x_j \backslash j \in J\}$;

$x_Q =$ cum(x_J) ; ou encore : $x_Q = p(x_J)$.



On doit identifier l'espace des R_Q des distributions de masse sur Q, à un sous espace L de l'espace R_J des distributions de masse sur J. Cette identification ne peut se faire qu'en utilisant la loi marginale f_J définie sur J. En bref, étant donnée $y_Q \in R_Q$, il faut pour chaque classe q *répartir* le nuage y_q qui lui est affecté entre les éléments j de cette classe q: ce que l'on fait par une formule de proportionalité:

$$\text{rep}(y_Q) = y_J = \{y_j \mid j \in J$$

$$\forall j \in q : y_j = y_q(f_j/f_q) = f_j(y_q/f_q);$$

où on a noté $f_q = \Sigma\{f_j \mid j \in q\}$. Cette formule s'interprète encore en disant que la densité de $\text{rep}(y_Q) = y_J$ relativement à f_J est constante sur chaque classe q, et y a pour valeur la densité (y_q/f_q) de la mesure y_Q relativement à f_Q.

Il reste maintenant à montrer que l'application rep o cum de R_J est une projection orthogonale (relativement à la métrique du χ^2 de centre f_J). Afin de simplifier l'écriture, nous noterons seulement r et c pour rep et cum . On a les 4 propositions suivantes:

<u>proposition 1</u> : $\forall \quad y_Q \in R_Q$, $\text{cry}_Q = y_Q$;

en d'autres termes, si partant d'une mesure y_Q sur Q, on répartit sa masse sur J (suivant les coefficients donnés pour f_J), on obtient une mesure ry_J qui par cumul redonne y_Q.

<u>preuve</u> : $\forall \ j \in J : ry_j = y_{q(j)}(f_j/f_{q(j)})$;

(où on a noté q(j) la classe de la partition Q, à laquelle appartient j)

$$\text{cryq} = \Sigma\{ry_j \mid j \in q\} = \Sigma\{y_q(f_j/f_q) \mid j \in q\}$$

$$= (y_q/f_q)\,\Sigma\{f_j \mid j \in q\} = (y_q/f_q)f_q = y_q.$$

Il résulte de la proposition 1 que rc est un projecteur;

proposition 2 : $\forall\ x_J \in R_J\ \ rc(rc\ x_J) = rc\ x_J$.

preuve : on peut écrire en associant les applications successives r et c:

$rc(rc\ x_J) = r(cr(c\ x_J)) = r(c\ x_J) = rc\ x_J$.

Pour l'orthogonalité, on a la

proposition 3 : $\forall\ x_J \in R_J, y_Q \in R_Q : < x_J, ry_Q > = < c\ x_J, y_Q >$.

Dans cette formule, le premier produit scalaire (entre mesures sur J) est calculé pour la métrique du χ^2 de centre f_J ; et le second (entre mesures sur Q) pour la métrique du x^2 de centre f_Q.

preuve : on développe le premier produit scalaire et on retrouve le deuxième:

$$\begin{aligned}
<x_J, ry_Q> &= \Sigma\{x_j\ .\ (ry)_j / f_j \mid j \in J\} \\
&= \Sigma\{\Sigma\{x_j\ (ry)_j / f_j \mid j \in q\}\ q \in Q\} \\
&= \Sigma\{\Sigma\{x_j\ y_q\ (f_j/f_q)/\ f_j \mid j \in q\} \mid q \in Q\} \\
&= \Sigma\{\Sigma\{x_j\ y_q/f_q \mid j \in q\} \mid q \in Q\} \\
&= \Sigma\{\Sigma\{x_j \mid j \in q\ \ (y_q\ /fq) \mid q \in Q\} \\
&= \Sigma\{(cx)_q(y_q/f_q) \mid q \in Q\} = <c\ x_J, y_Q>.
\end{aligned}$$

La proposition 3 permet de démontrer que le noyau N de l'application p = rc est orthogonal à l'image L ; c'est la

proposition 4 : $\forall\ x_J \in R_J,\ \ y_Q \in R_Q : (rc\ x_J = 0) \Longrightarrow <x_J, ry_Q> = 0$

preuve : en effet, $rc\ x_J = 0$ est équivalent à $c\ xJ = 0$ (car la répartition des masses r ne peut produire une mesure nulle à partir d'une masse non nulle); et si $c\ x_J = 0$, est nul le produit scalaire $< cx_J, y_Q > = <cx_J, ry_Q>$.

Aussi le cumul des colonnes par blocs est équivalent à une projection orthogonale; la norme $\|f_Q^c - f_Q\|^2$ est la norme de la projection de la différence $(f_J^c - f_J)$; le quotient $\|f_Q^c - f_Q\|^2 / \|f_J^c - f_J\|^2$ est bien le $\cos^2$ d'un vecteur avec un sous espace, ou dans l'interprétation statistique une qualité de représentation; et les termes de ce quotient afférents aux différentes classes q sont des $CO2_q$, des cosinus carrés d'angles formés avec les axes, d'un système orthogonal dont est muni l'espace $L \approx R_Q$. Les formules introduites au § 2 ont été légitimées par les propositions démontrées au § 3.