LES CAHIERS DE L'ANALYSE DES DONNÉES

A. ALAWIEH

Cas modèle de l'analyse d'une variable continue unique codée barycentriquement

Les cahiers de l'analyse des données, tome 16, nº 1 (1991), p. 29-34

<http://www.numdam.org/item?id=CAD_1991__16_1_29_0>

Les Cahiers de l'Analyse des Données
Vol XVI - 1991 - n° 1 - pp. 29-34.

CAS MODÈLE DE L'ANALYSE D'UNE VARIABLE CONTINUE UNIQUE CODÉE BARYCENTRIQUEMENT

[MOD. CODE BARY.]

*A. ALAWIEH**

Il semble vain d'appliquer l'analyse multidimensionnelle à l'étude d'un tableau à une seule colonne! Que ce tableau puisse être codé sur plusieurs colonnes, ne change rien au fait qu'une variable unique est adéquatement étudiée par de simples tracés d'histogrammes. Mais, dans la pratique (cf., e.g., [ESP. VIE MONDE], in *CAD*, Vol XV, n°4, 1990), il se peut que l'analyse d'un ensemble redondant de variables codées barycentriquement, produise d'abord, dans l'espace des premiers axes factoriels, une représentation semblable à celle qu'on obtiendrait à partir d'une seule variable. Il est donc légitime de s'intéresser, comme à un cas modèle, à l'analyse d'une variable unique codée barycentriquement.

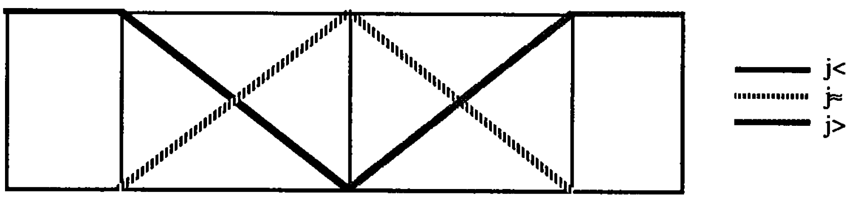
Nous démontrerons que, quels que soient la distribution de la variable étudiée et le découpage choisi, les modalités sont toujours rangées sur l'axe1 dans leur ordre naturel. Ce résultat général est d'autant plus satisfaisant qu'il semblerait, d'après certains exemples, qu'un choix mal équilibré des valeurs pivot puisse empêcher de retrouver l'ordre de celles-ci. Nous considérerons donc d'abord des cas particuliers afin d'acquérir une vue du problème général qui sera traité ensuite.

1 Codage barycentrique d'une variable continue de distribution uniforme

Du point de vue géométrique, il est plus simple que les modalités créées aient toutes le même poids: dans ce cas, en effet, il résulte de la formule de la distance du chi2 que le simplexe ayant pour sommets les profils purs (i.e. concentrés sur une seule modalité) a toutes ses arêtes égales. Pour obtenir ce résultat dans le codage barycentrique d'une variable de distribution uniforme, il faut, (ce que l'on fait d'ailleurs généralement,) placer les valeurs pivot extrêmes

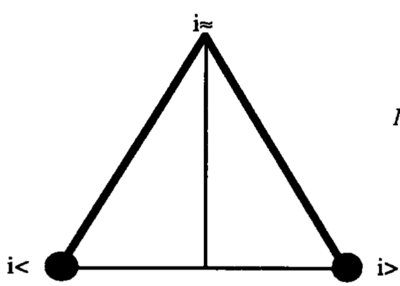
(*) Laboratoire de Statistique, Université Pierre et Marie Curie, Paris.

en retrait des extrémités de l'intervalle où est comprise la variable: ainsi, il y a, à droite et à gauche, un segment de masse finie dont tous les points sont codés suivant le même profil pur; les autres points étant représentés sur les segments joignant deux profils purs de rang consécutifs.



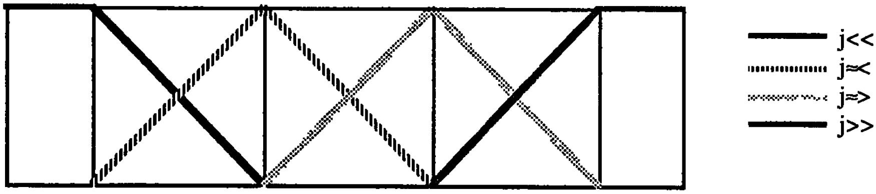
codage d'une variable suivant 3 modalités de même poids {j<, j≈, j>}

De façon précise, dans le cas d'un codage suivant 3 modalités, il faut réserver de chaque côté un sixième de l'intervalle. Entre deux valeurs pivot consécutives, il y a un tiers de la masse totale, que le codage attribue pour moitié à chacun des pivots: la figure indique par des lignes brisées de tracés différents la part de chacune des modalités dans le profil codé des valeurs de la



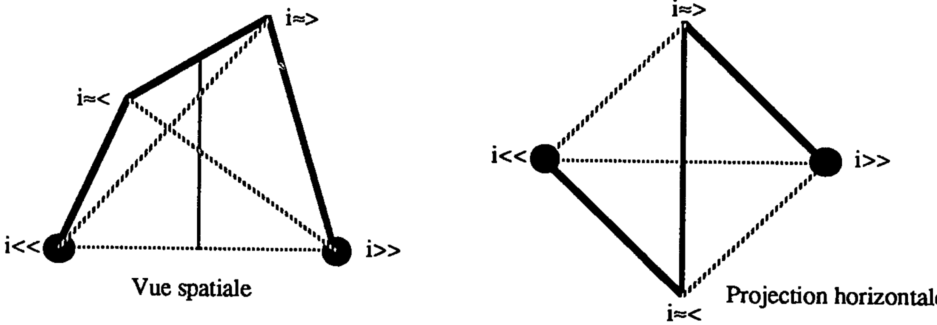
Nuage des individus pour une variable continue codée suivant trois modalités de même poids: on a noté {i<, i≈, i>} les profils purs

variable. Ainsi, le nuage, inscrit sur le périmètre d'un triangle équilatéral, comporte deux masses ponctuelles {i<} et {i>} égales à 1/6; et deux arêtes ayant chacune la masse 1/3.



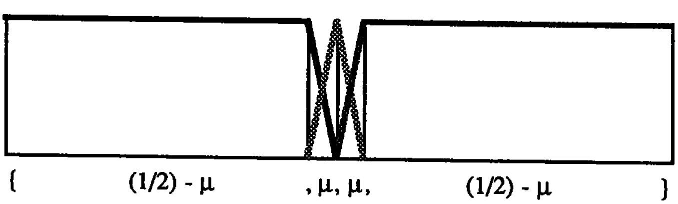
codage d'une variable suivant 4 modalités de même poids {j<<, j≈<, j≈>, j>>}

De même, dans le cas d'un codage suivant 4 modalités, il faut réserver, de chaque côté, un huitième de l'intervalle. Le nuage, porté par un tétraèdre régulier, comporte deux masses ponctuelles {i<<} et {i>>} égales à 1/8; et trois



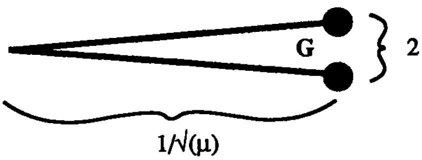
Nuage des individus pour une variable continue codée suivant quatre modalités de même poids; on a noté les profils purs: {i<<, i≈<, i≈>, i>>}

arêtes ayant chacune la masse 1/4. Dans l'espace, la perpendiculaire commune (verticale sur la figure) aux deux arêtes {i<<, i>>} et {i≈<, i≈>} est un axe de symétrie pour le nuage. C'est donc un axe principal d'inertie; et l'on peut vérifier que c'est l'axe de rang 2. Notre projection horizontale est le plan des axes (1,3); dans ce plan, l'axe 1 coupe le segment (projeté) {i<<, i≈<} en un point plus proche de i<< que de i≈<; ce qui assure que, sur l'axe 1, les modalités se projettent dans leur ordre naturel.



Codage suivant 3 modalités; la modalité centrale est de poids μ infinitésimal.

Avec des modalités de poids inégal, le simplexe ayant pour sommets les profils purs n'a pas toutes ses arêtes égales. Dans l'exemple simple figuré ici, la modalité centrale a un poids infinitésimal μ. Il en résulte que le nuage est un triangle isocèle dont la base est constituée des modalités extrêmes, lourdes et proches (distance ≈ 2); tandis que la modalité centrale, très éloignée, tient lieu de sommet. Assurément, suivant la dimension principale de la figure, la modalité centrale s'oppose aux deux extrêmes; mais le premier axe principal d'inertie n'est pas cet axe d'allongement maximum. De façon précise, le centre de gravité, G, est très proche de la base; dans la direction verticale, l'inertie provient essentiellement des deux modalités extrêmes et est ≈1; dans la direction horizontale, au contraire, l'inertie provient des deux segments de masse totale 2μ; elle est équivalente à l'inertie, par rapport à l'une de ses extrémités, d'un segment unique dont la masse est 2μ et le carré de longueur est (1/μ), soit (2/3).



représentatiuon du nuage:
la masse totale des segments est 2μ

D'où il résulte que sur le premier axe principal d'inertie (qui est, sur la figure, la verticale passant par G) les modalités se projettent suivant leur ordre naturel.

Dans les exposés d'introduction à l'analyse des correspondances, afin de soutenir l'intuition, on assimile "premier axe principal d'inertie du nuage" à "axe principal d'allongement de la figure". Cette assimilation ne vaut pas toujours; on a trouvé ici une exception d'importance, mettant en jeu les modalités d'une structure très simple.

Il n'est pas utile de considérer d'autres exemples plus complexes; car, ainsi que nous l'avons annoncé, nous démontrerons au §2 que, quels que soient la distribution de la variable étudiée et le découpage choisi, les modalités du codage barycentrique sont toujours rangées sur l'axe1 dans leur ordre naturel.

2 Ordre latéral entre profils et codage barycentrique général

Dans la leçon [Ord. Lat.], TIB, n°8, du Traité sur *L'Analyse Des Données*, est définie la notion d'"ordre latéral entre loi de probabilités sur un ensemble ordonné". D'où découle la notion de "translation latéralement croissante d'un ensemble ordonné vers un autre"; et le théorème que "si sont latéralement croissantes les deux transitions, de I vers J et de J vers I, associées à une correspondance entre deux espaces ordonnés I et J, alors il existe au moins un facteur relatif à la plus grande valeur propre qui soit croissant sur I et J." Nous reprendrons ici brièvement ces notions afin d'appliquer le théorème à l'analyse du codage barycentrique d'une variable unique.

Soit a_I, b_I, deux lois de probabilités (ou deux mesures positives de même masse) sur un ensemble ordonné I: on dit que b_I est à droite de a_I, ce que l'on note $Dr(a_I, b_I)$, s'il existe une mesure g_{II} (*g* comme *graphe*) sur I × I, telle que la première et la deuxième projection de g_{II} sont respectivement a_I et b_I, et que le support de la mesure n'a pas de point au dessous de la diagonale.

Cette définition correspond à la notion intuitive que l'on peut décomposer simultanément a_I et b_I en somme de termes élémentaires:

$$a_I = \Sigma\{a_I^n \mid n = 1, 2, \ldots\} \quad ; \quad b_I = \Sigma\{b_I^n \mid n = 1, 2, \ldots\},$$

de telle sorte que les termes de même rang *n* des mesures aient des supports situés l'un (celui de b^n) à droite de l'autre (celui de a^n); (ces supports pouvant être éventuellement réduits à un seul et même point). Dans [Ord. Lat.], on considère le cas général d'ensembles partiellement ordonnés; ici, nous ne rencontrerons que deux ensembles totalement ordonnés: un segment (l'ensemble des individus, ou valeurs possibles de la variable) et un ensemble fini (l'ensemble des modalités); et c'est pourquoi la relation d'ordre entre éléments de ces ensembles pourra être notée simplement par le signe usuel '<'.

Ceci posé, on dit qu'une transition de I vers J est latéralement croissante si:

$$\forall\, i, i' \in I : \quad (i < i') \;\Rightarrow\; \mathrm{Dr}(f_J^{\,i}, f_J^{\,i'}) \;;$$

en d'autres termes: si *i'* est à droite de *i*, le profil de *i'* sur J, est à droite du profil de *i* sur J.

Pour la démonstration du *Théorème* sur l'analyse d'une correspondance dont les deux transitions associées sont latéralement croissantes, nous renvoyons à [Ord. Lat.]; nous bornant à rappeler que, ces transitions ayant la propriété d'envoyer le cône des fonctions croissantes sur I dans le cône des fonctions croissantes sur J et réciproquement, il suffit, pour obtenir un couple de facteurs associés croissants relatif à la première valeur propre, de partir d'une fonction croissante (non contenu dans l'espace engendré par les facteurs relatifs à des valeurs propres autres que la plus grande) et de procéder par itération (selon la méthode usuelle associée, en analyse factorielle, au nom de Hotelling).

Reste à démontrer que sont latéralement croissantes les transitions associées au modèle général de correspondance que nous considérons: avec un ensemble I qui est un segment muni d'une loi de probabilité quelconque (non nécessairement de densité constante, comme au §1); et un ensemble fini J, {j(1), j(2),...}, identifié à l'ensemble ordonné des valeurs pivot {i(1), i(2), ...}, disposées sur I de façon quelconque.

On voit aisément que, si (i < i'), il y a entre les profils sur J obtenus par codage barycentrique la relation demandée. Il suffit de le démontrer dans le cas, le moins net *a priori*, où *i* et *i'* appartiennent à un même intervalle (i(n), i(n+1)) délimité par deux pivots consécutifs. Les profils sont alors portés par le sous-ensemble à deux éléments {j(n), j(n+1)}; on peut noter:

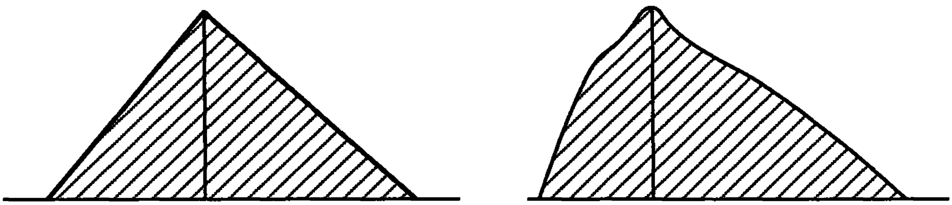
$$k(i, j(n)) = a \;;\; k(i, j(n+1)) = b \quad ; \quad k(i', j(n)) = a' \;;\; k(i', j(n+1)) = b' \;;$$

$$\text{avec:}\quad a' < a \;;\; b < b' \;;\; a + b = a' + b' = 1 \;;\; (a - a') = (b' - b) = c > 0 \;;$$

et on décompose chacun des deux profils sur {j(n), j(n+1)}en une somme de trois termes:

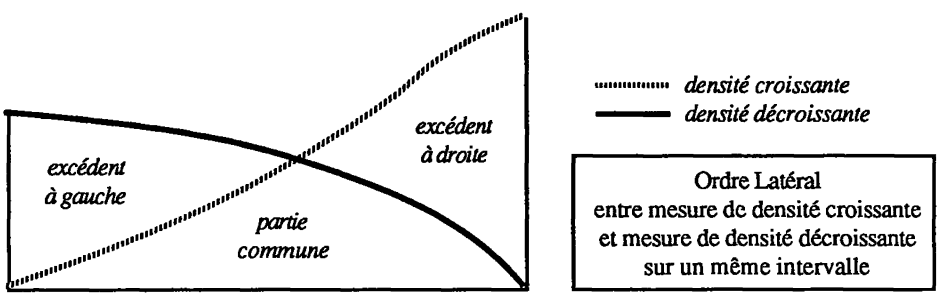
$$f_J^{\,i} = (a', 0) + (c, 0) + (0, b) \quad ; \quad f_J^{\,i'} = (a', 0) + (0, c) + (0, b) \quad ;$$

les premiers et derniers termes sont les mêmes pour *i* et *i'* et le terme médian du profil de *i'* est à droite de celui du profil de *i*: d'où, entre les profils, la relation d'ordre latéral demandée.



Densité du profil d'une modalité j(n) sur l'ensemble continu des individus:
à gauche, échelle initiale de la variable; à droite, échelle avec densité uniforme.

Quant aux profils sur I des modalités j(n), ils ont pour densité, par rapport à la loi de i sur I, les fonctions caractéristiques bien connues en dent de scie. Il est peut-être plus clair de redessiner le segment I de telle sorte que la loi de i ait une densité constante: ainsi l'on voit, tel quel, le profil de j(n).



Pour voir la relation d'ordre latéral entre les profils sur I de deux modalités, il suffit de considérer (ici encore) le cas le moins net de deux modalités consécutives j(n) et j(n+1), dont les supports de distribution empiètent. Le profil de j(n) comprend une partie A(n) portée par (i(n-1), i(n)) et une partie B(n+1), portée par (i(n), i(n+1)); et de même le profil de j(n+1) s'écrit A(n+1) + B(n+2). La figure montre comment décomposer B(n+1) et A(n+1) en une *partie commune* C(n+1) et des *excédents* Bg(n+1) et Ad(n+1); et l'on a:

$$f_I^{j(n)} = A(n) + Bg(n+1) + C(n+1) \; ; \; f_I^{j(n+1)} = Ad(n+1) + B(n+2)) + C(n+1) \; ;$$
$$\text{avec: } Dr((A(n) + Bg(n+1)), (Ad(n+1) + B(n+2)) \; ;$$

d'où la relation latérale demandée, et le théorème annoncé.

NB: la clause "il existe *au moins un* facteur relatif à la plus grande valeur propre...", posée dans le théorème, vise le cas limite d'une v.p. multiple; ce qui se produit, e.g., si la distribution de i est concentrée en trois masses ponctuelles.