

J. HAYEK

A. LOEMBA

Un cas modèle pour l'analyse d'un ensemble de découpages d'une variable unique

Les cahiers de l'analyse des données, tome 16, n° 2 (1991), p. 201-214

http://www.numdam.org/item?id=CAD_1991__16_2_201_0

© Les cahiers de l'analyse des données, Dunod, 1991, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

UN CAS MODÈLE POUR L'ANALYSE D'UN ENSEMBLE DE DÉCOUPAGES D'UNE VARIABLE UNIQUE

[MOD. DÉC. VAR.]

J. HAYEK*

1 Scalogramme de Guttman et questionnaire redondant

Afin de justifier les heureux résultats obtenus communément dans l'analyse des réponses fournies par un ensemble I de sujets à un questionnaire, l'article [MOD. VAR. CLASS.] (in *CAD*, Vol XV, n°4) considère "un modèle de stricte redondance; où il n'y a véritablement qu'une seule question plusieurs fois présentée mais toujours avec le même découpage en modalités et le même codage; la diversité des réponses fournies par un même sujet ne résultant que d'un comportement aléatoire".

Le même article se termine par une suggestion:

"On peut imaginer encore qu'une question de base unique soit plusieurs fois présentée avec des découpages différents: en ce cas, même avec des réponses non aléatoires, on aura des corrélations révélant la structure de l'ensemble des modalités. Le modèle le plus simple serait celui d'une variable continue sous-jacente unique dont chaque question réalise le codage suivant un découpage différent".

C'est un tel modèle que nous nous proposons de construire ici. À la vérité, un modèle existe déjà, qu'il suffira d'interpréter et de généraliser pour réaliser convenablement la suggestion de [MOD. VAR. CLASS.]: c'est le *scalogramme* de Guttman.

Reprenons, en effet, sous sa forme initiale, le modèle de l'échelle de Guttman. Soit Q un ensemble de n questions, numérotées de 1 à n , et admettant chacune un ensemble $J_q = \{q_-, q_+\}$ de deux modalités de réponse, que nous appellerons, respectivement, modalité *inférieure*, q_- , et modalité *supérieure*, q_+ . Supposons, de plus, que tout individu i qui rentre dans la modalité inférieure

(*) Étudiant en doctorat d'Université, Université Pierre et Marie CURIE.

d'une question q rentre nécessairement dans la modalité inférieure de toute question q' dont le rang est supérieur à celui de q : Il y aura alors $n+1$ systèmes de réponses permis; et, si l'on convient de disposer l'ensemble J des modalités de réponse non, comme on le fait ordinairement, en plaçant successivement les blocs J_q afférents à chaque question, mais en plaçant d'abord toutes les modalités inférieures, et ensuite toutes les modalités supérieures, on aura, pour l'ensemble des lignes de réponse possibles, le tableau classique en parallélogramme.

	1-	2-	3-	4-	5-	1+	2+	3+	4+	5+
i0	1	1	1	1	1	0	0	0	0	0
i1	0	1	1	1	1	1	0	0	0	0
i2	0	0	1	1	1	1	1	0	0	0
i3	0	0	0	1	1	1	1	1	0	0
i4	0	0	0	0	1	1	1	1	1	0
i5	0	0	0	0	0	1	1	1	1	1

exemple de scalogramme: les questions sont désignées par les nombres de 1 à 5; et les six systèmes de réponse possibles sont indicés de 0 à 5.

On interprète les questions de ce modèle comme des découpages en deux classes d'une unique variable x prenant ses valeurs dans l'intervalle $(0, 1)$: à la question q est associée la borne q/n ; la modalité de réponse d'un individu i à la question q est 'q-', si $x(i)$ tombe dans l'intervalle $(0, q/n)$; et, si $x(i)$ est dans $(q/n, 1)$, la réponse est 'q+'.

De ce point de vue, il s'agit donc bien "d'une variable continue sous-jacente unique dont chaque question réalise le codage suivant un découpage différent". Et c'est ainsi, au fond, que L. Guttman a conçu le scalogramme; car il considérait un univers de questions comme suscitant, sous une forme redondante, l'expression, plus ou moins forte, d'une tendance unique. Toutefois, le modèle doit être généralisé, parce qu'il ne considère que des variables découpées en deux classes. Pour introduire des découpages en trois classes, ou davantage, il sera commode de considérer un modèle continu.

2 Modèles continus et découpages d'une variable en classes

On sait (cf., e.g., *L'Analyse Des Données*, TIIB n°7, §3.1) que l'échelle de Guttman admet pour limite, quand $n \rightarrow \infty$, une correspondance continue (dont les facteurs pour l'ensemble des individus s'expriment directement en terme de polynômes de Legendre).

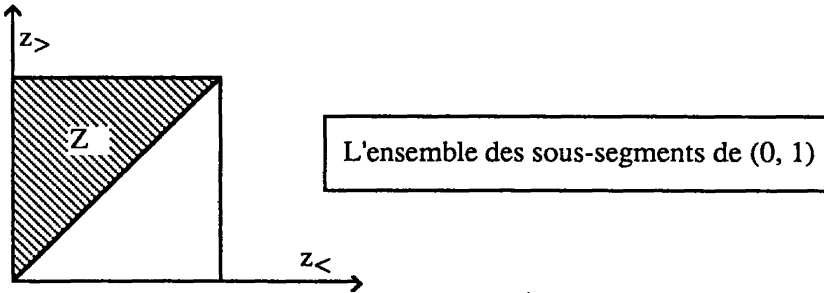
L'ensemble des individus (ou, ce qui revient au même, l'ensemble des systèmes de réponse) est assimilé à l'intervalle $X = (0, 1)$, sur lequel la loi

marginale a la densité constante $p(x) = 1$. L'ensemble des questions, noté U , s'identifie également à l'intervalle $(0, 1)$; l'abscisse u d'une question étant distribuée équiprobablement. Ainsi, l'ensemble Y des modalités de réponse peut être paramétré par l'intervalle $(0, 2)$. De façon précise, la modalité u - reçoit l'abscisse $y(u-) = u$; de même, $y(u+) = 1+u$; et on a la correspondance continue p_{XY} définie, ainsi que ses marges, par les formules:

$$p(x, y) = \begin{cases} 1 & \text{si } y \in (x, x+1) \\ 0 & \text{sinon} \end{cases}$$

$$p(x) dx = dx ; p(y) dy = \begin{cases} y dy & \text{si } y \in (0, 1) \\ (2-y) dy & \text{sinon} \end{cases}$$

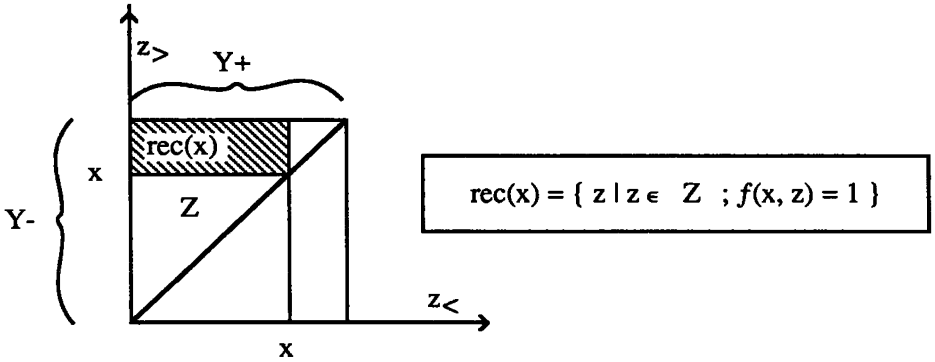
Pour généraliser ce modèle, conservons la forme adoptée pour l'ensemble X des invidus: $(0, 1)$; p_X pouvant être la loi uniforme, ou une autre loi. L'ensemble, Z , des modalités de réponse s'identifiera, nécessairement, à tout ou partie de l'ensemble des sous-intervalles de l'intervalle $(0, 1)$. Dans le scalogramme, on se restreint aux sous-intervalles dont l'une des extrémités a pour abscisse 0 ou 1: cette condition doit être écartée si l'on considère des découpages de $(0, 1)$ en plus de 2 classes. On a pu considérer Y comme formé de deux segments consécutifs: $(0, 1)$, et $(1, 2)$; avec $y \in (0, 1)$, pour représenter l'intervalle $(0, y)$; et $y \in (1, 2)$, pour représenter l'intervalle $(y-1, 1)$. Dans le cas général, on posera:



$$Z = \{ \{z_<, z_>\} \mid 0 \leq z_< < z_> \leq 1 \} ;$$

où une inégalité stricte a été introduite pour qu'il n'y ait pas de sous-intervalle réduit à un point. En termes géométriques, Z est un triangle rectangle, moitié du carré $(0, 1) \times (0, 1)$ formée des points situés au-dessus de la première bissectrice. Pour le scalogramme, Y n'est autre que la réunion des deux côtés de l'angle droit du triangle Z .

Afin de fixer p_{XY} , on a considéré explicitement l'ensemble U des questions, c'est-à-dire des partitions de $(0, 1)$ en deux segments consécutifs. Dans le cas général, il semble peu commode de partir de l'ensemble des questions; car



celles-ci peuvent avoir un nombre quelconque de modalités. On cherchera donc à caractériser la loi p_{XZ} directement.

On sait d'abord que le support de cette loi est compris dans l'ensemble Ω :

$$\Omega = \{ (x, \{z_<, z_>\}) \mid x, z_<, z_> \in (0, 1) ; z_< \leq x < z_> \} ;$$

dont la fonction caractéristique est:

$$f_{\Omega}(x, z) = \text{si } x \in (z_<, z_> \text{ alors } 1 \text{ sinon } 0 ;$$

Ω étant le graphe de la relation d'appartenance de x au segment z . Pour une valeur fixée de x , l'ensemble des points z où $f(x, z)$ vaut 1 n'est autre qu'un rectangle $\text{rec}(x)$ inscrit dans le triangle Z :

$$\text{rec}(x) = \{ z \mid z \in Z ; f(x, z) = 1 \} = \{ (z_<, z_>) \mid z_< \in (0, x) ; z_> \in (x, 1) \}.$$

Dans la définition de Ω , on a choisi d'introduire des intervalles semi-ouverts de telle sorte qu'un point situé à la limite de deux modalités soit compris dans la modalité inférieure.

Ensuite, au lieu de considérer la masse relative des partitions de $(0, 1)$, on attribue directement une masse aux modalités de réponses (c'est-à-dire aux segments $z \in Z$); dont chacune, même dans le cas fini, peut intervenir dans plusieurs questions, autrement dit dans plusieurs des partitions retenues pour le segment $(0, 1)$. Cette distribution de masse sera notée q_Z . (Dans le cas du scalogramme, on a $q_Y = dy$, l'élément de longueur usuel.) On pose alors:

$$p_{XZ} = f(x, z) p_X \otimes q_Z ;$$

il importe de remarquer que la mesure q_Z est nécessairement singulière, i.e. qu'elle n'a pas de densité continue par rapport à l'élément d'aire usuel dz de Z :

en effet, toute partition comporte des classes extrêmes; l'une, à gauche, représentée par un point $z=(0, z_>)$; l'autre, à droite, de la forme $(z_<, 0)$; en sorte que les côtés de l'angle droit du triangle Z portent des distributions de masse non nulle, (plus précisément chacune de masse 1, si l'on adopte la normalisation proposée ci-après,) auxquelles se restreint q_Z , dans le cas où toute question n'a que 2 modalités. On notera Z° l'intérieur de Z ; et $\{Y_-, Y_+\}$, les deux côtés:

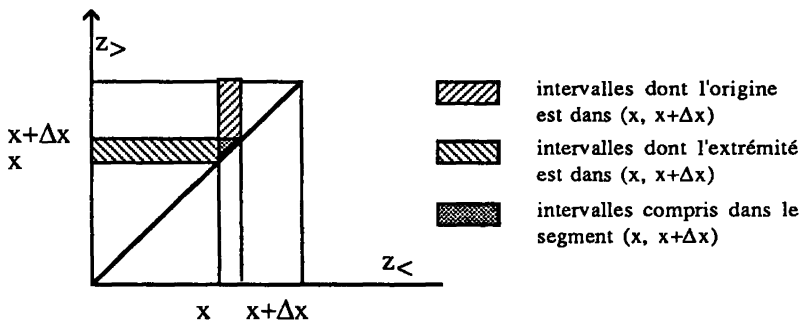
$$Y_- = \{(0, z_>) \mid z_> \in]0, 1[\} ; \quad Y_+ = \{(z_<, 1) \mid z_< \in]0, 1[\} .$$

Reste à déterminer à quelles conditions doit être assujettie la mesure positive q_Z . Pour cela, sans se soucier d'abord de grouper les modalités en questions ou partitions, (ce qui, même dans le cas fini, peut se faire de plusieurs façons,) on imposera que soit respectée une généralisation appropriée de la forme disjonctive complète; en ce sens que le nombre des modalités dans lesquelles rentre un individu soit indépendant de x . Si l'on normalise à 1 ce nombre, (lequel dans le cas présent résulte d'une intégration,) on a la formule:

$$\forall x \in X : \text{masse } (f(x, \cdot)) q_Z = 1 ;$$

en termes géométriques, on dira que tout rectangle $\text{rec}(x)$ a la masse 1 pour la mesure q_Z .

$$\forall x \in X : \int \{f(x, z) dq_Z \mid z \in \text{rec}(x)\} = 1 .$$



Comme on le voit sur la figure, affirmer que les deux rectangles $\text{rec}(x)$ et $\text{rec}(x+\Delta x)$ contiennent la même masse équivaut, en bref, à affirmer qu'ont même masse les intervalles z qui ont leur extrémité $z_>$ dans $(x, x+\Delta x)$ et ceux qui y ont leur origine $z_<$. Cette condition s'interprète aisément par référence à des questions, ou partitions: toute partition ayant dans $(x, x+\Delta x)$ un point de séparation comprend nécessairement, à la fois, un intervalle qui y a son extrémité et un autre qui y a son origine.

En éliminant les sous-intervalles de longueur nulle, on a imposé que l'hypoténuse du triangle Z (qui est aussi la diagonale du carré de côté unité) ait masse nulle pour la mesure q_Z . De plus, si, comme il est naturel, on élimine l'intervalle (0,1) tout entier, qui définit une modalité dans laquelle tous les x sont compris, on ajoutera la condition qu'ait masse nulle le point $\{0, 1\}$ (sommet de l'angle droit du triangle Z); d'où une décomposition de q_Z en trois termes, dont les supports sont les ensembles disjoints $\{Z^0, Y_-, Y_+\}$:

$$q_Z = q_{Z^0} + q_{Y_-} + q_{Y_+} \quad ; \quad \text{masse}(q_{Y_-}) = \text{masse}(q_{Y_+}) = 1 \quad ;$$

relativement à cette décomposition, la condition imposée à q_Z s'écrit comme une égalité entre distributions de l'extrémité et de l'origine des intervalles:

$$\text{pr}_{Y_+}^{Z}(q_{Z^0}) + \text{rot}_{Y_+}^{Y_-}(q_{Y_-}) \approx \text{pr}_{Y_-}^{Z}(q_{Z^0}) + \text{rot}_{Y_-}^{Y_+}(q_{Y_+}) \approx q_X \quad ;$$

dans cette formule, le signe ' \approx ' désigne l'isomorphisme naturel entre segments; 'rot' est une rotation autour du centre du carré; et apparaît la mesure q_X de la distribution sur $]0, 1[$ des points de division entre segments.

On remarquera que la mesure q_X a une masse totale inférieure d'une unité à celle de q_Z et supérieure d'une unité à celle de q_{Z^0} ; ce qui correspond au fait que, dans une partition, le nombre des points de division est inférieur d'une unité au nombre total de segments; et dépasse d'une unité le nombre des segments internes. Cette mesure q_X peut être mise à la base de constructions de modèles: si l'on impose, comme il est naturel, que soient éliminés du modèle le segment (0,1) ainsi que les segments réduits à un point unique, les seules conditions auxquelles doit satisfaire q_X sont d'avoir masse totale supérieure ou égale à 1, et de ne pas comporter de masse ponctuelle strictement supérieure à 1.

Introduire la mesure q_X permet de préciser ici les conditions imposées à la mesure p_Z . Partons, pour construire le modèle, de p_{Z^0} : de q_X , sont déjà connus les termes de projection, que nous noterons comme des mesures sur le segment $(0,1) \approx X$, lois des origines et extrémités des intervalles inclus dans $]0, 1[$:

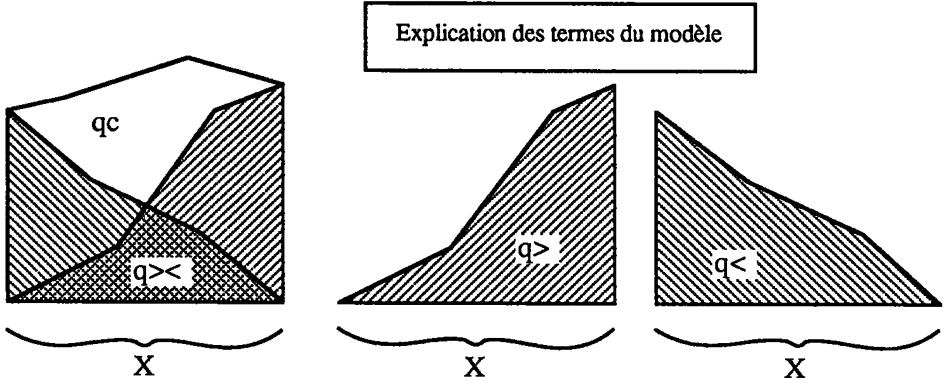
$$\text{pr}_{Y_+}^{Z}(q_{Z^0}) \approx q_X^< \quad ; \quad \text{pr}_{Y_-}^{Z}(q_{Z^0}) \approx q_X^> \quad ; \quad \text{d'où: } q_X = q_X^< + q_{Y_+} = q_X^> + q_{Y_-} \quad ;$$

et, en notant: $q_X^>< = \inf(q_X^<, q_X^>)$, il vient:

$$q_X = q_X^>< + (q_X^< - q_X^><) + q_{Y_+} = q_X^>< + (q_X^> - q_X^><) + q_{Y_-} \quad ;$$

$$(q_X^< - q_X^><) + q_{Y_+} = (q_X^> - q_X^><) + q_{Y_-} \quad ;$$

$$q_{Y_-} - (q_X^> - q_X^><) = q_{Y_+} - (q_X^< - q_X^><) = q_X^c \quad ;$$



parce que les différences $(q^>_X - q^<_X)$ et $(q^<_X - q^>_X)$ sont des mesures étrangères entre elles, q^c_X doit être une mesure ≥ 0 ; d'où une inégalité, condition nécessaire et suffisante pour compléter le modèle:

$$\text{masse } (q^<_X - q^>_X) = \text{masse } (q^>_X - q^<_X) = \mu \leq 1 ;$$

on fixe q_{Y_-} et q_{Y_+} , (de masse normalisée à 1,) par le choix d'une mesure positive quelconque, q^c_X de masse $(1-\mu)$:

$$q_{Y_-} \approx (q^<_X - q^>_X) + q^c_X \quad ; \quad q_{Y_+} \approx (q^>_X - q^<_X) + q^c_X .$$

Dans ces formules, $(q^<_X - q^>_X)$ (resp. $(q^>_X - q^<_X)$) s'interprète comme la distribution des origines (resp. extrémités) de segments inclus dans $]0,1[$ qui ne peuvent prolonger (resp. être prolongés par) un segment satisfaisant à la même condition, et doivent donc prolonger (resp. être prolongés par) un segment issu de 0 (resp. se terminant en 1).

3 Modèle général de classes et ensembles de partitions

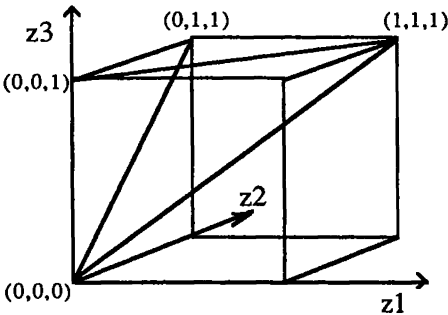
C'est l'occasion de revenir à la construction d'un modèle à partir de la notion de partition. Supposons d'abord que toutes les questions soient codées par découpage en 3 modalités. Dans ce cas, une partition est complètement définie par la donnée de son intervalle central, qui est un point z intérieur au triangle Z . La mesure q_Z a pour masse totale 3; elle est la somme de trois mesures: la mesure q_{Z^o} , portée par l'intérieur de Z , et assujettie à la seule condition d'être positive de masse totale 1; et les deux mesures, q_{Y_-} et q_{Y_+} , déterminées par q_{Z^o} et portées l'une par Y_- , l'autre par Y_+ , qui correspondent respectivement aux intervalles initial et final des partitions. Par exemple, la

mesure portée par Y_- s'obtient (cf. *supra*) en projetant q°_Z sur Y_+ puis faisant tourner la mesure obtenue d'un angle droit autour du centre du carré.

En particulier, si q°_Z a une densité constante:

$$q^{\circ}_Z = 2.dz, \text{ où } dz \text{ désigne l'élément d'aire usuel de } Z,$$

la mesure portée par Y_- est: $2.(1-z_>).dz_>$; sur Y_+ on a: $2.z_<.dz_<$; et la distribution q_X des points de séparation est uniforme: $q_X = 2.dx$.



Les quatre sommets qui ont été marqués définissent le tétraèdre: $\{z \mid z_1 \leq z_2 \leq z_3\}$

De même, une partition en 4 classes est donnée par une suite de trois points de division (z_1, z_2, z_3) , satisfaisant aux inégalités: $0 < x_1 < x_2 < x_3 < 1$. Ce qui correspond à une partie tétraédrique du cube unité. Si l'on attribue à la distribution de la partition sur ce tétraèdre une densité uniforme, on a, pour q_Z , une mesure de masse totale 4 (4 est, en effet, le nombre de modalités, ou classes, de chaque question) dont on peut, comme dans le cas de 3 classes, préciser les trois mesures composantes, portées, respectivement, par l'intérieur Z° de Z , par Y_- et par Y_+ :

$$q_{Z^{\circ}} = 6.(1 + z_< - z_>).dz \quad ; \quad q_{Y_-} = 3.(1 - z_>)^2.dz_> \quad ; \quad q_{Y_+} = 3.(z_<)^2.dz_< .$$

Dans le cas général d'une partition en m classes, on doit considérer les suites croissantes de $m-1$ nombres de l'intervalle $(0,1)$, c'est-à-dire un simplexe de l'hypercube d'arête 1; et si, sur ce simplexe, la densité est constante (et égale à $(m-1)!$), on a pour q_Z la somme de trois mesures dont les densités sont données par des formules analogues à celles déjà écrites dans le cas $m = 4$:

$$q_{Z^{\circ}} = (m-1).(m-2).(1 + z_< - z_>)^{m-3}.dz \quad ;$$

$$q_{Y_-} = (m-1).(1 - z_>)^{m-2}.dz_> \quad ; \quad q_{Y_+} = (m-1).(z_<)^{m-2}.dz_< \quad ;$$

sur ces formules, on peut faire diverses vérifications: la masse de q_Z est $(m-2)$, tandis que celle des deux autres mesures, q_Y et q_{Y+} , est 1; la masse totale de q_Z étant m ; enfin, la distribution, notée, q_X , des points de séparation entre segments, a la densité constante $(m-1)$: $q_X = (m-1).dx$.

Dans le cas d'un modèle général, la masse totale de la mesure q_Z s'interprète comme le nombre moyen de modalités des questions ou partitions que l'on peut, d'une manière ou d'une autre, mettre à la base du modèle. Il faut noter que cette masse totale peut être infinie sans sortir de la condition du §2 (avoir masse 1 sur tout rectangle $rec(x)$). En voici la preuve sur un exemple.

Notons q_Z^m le modèle (construit ci-dessus) correspondant à l'ensemble des partitions en m classes affecté de la densité constante $(m-1)!$. On peut, par combinaison linéaire infinie à coefficients positifs, construire des modèles de la forme:

$$q_Z = \sum \{a(m).q_Z^m \mid m = 3, 4, \dots, \infty\}; \quad \text{où: } \sum \{a(m) \mid m = 3, 4, \dots, \infty\} = 1;$$

la masse totale de q_Z étant donnée par la somme $\sum \{m.a(m)\}$, on voit que cette masse sera infinie, notamment, si l'on pose: $a(m) = 3/(m.(m-1))$. Le propre de ces modèles est d'avoir une densité tendant vers l'infini quand on tend, dans Z , vers la diagonale du carré: car des partitions en un grand nombre de classes introduisent nécessairement des sous-segments de très petite longueur ($z_{<} \approx z_{>}$).

Selon un résultat classique, que l'on retrouvera au §5, la trace de la correspondance, ou somme des valeurs propres, n'est autre que le nombre moyen des modalités diminué de 1; c'est-à-dire la masse totale de la mesure q_X , ou celle de q_Z diminuée de 1. Ceci restreint l'intérêt des cas modèle pour lesquels $\int dq_Z = \infty$.

Au terme de ce §, nous nous bornerons à évoquer le problème, non traité ici, de la construction d'un modèle de partition à partir de la seule loi q_Z , assujettie aux conditions imposées ci-dessus au §2.

4 Transitions probabilistes associées au modèle général

À la loi de probabilité p_{XZ} (loi de correspondance entre X et Z) définie au §2, sont associées des transitions p_Z^X et p_X^Z . On a:

$$p_Z^X = f(x, z) q_Z; \quad p_X^Z = (1 / \int \{p_X \mid x \in (z_{<}, z_{>})\}) \cdot f(x, z) \cdot p_X;$$

nous expliquerons successivement ces formules.

Le profil p_Z^x de x sur Z , a pour support le rectangle $\text{rec}(x)$, ou ensemble des intervalles comprenant x , selon la définition du §2; on a montré que, dans le modèle général, la restriction à $\text{rec}(x)$ de la distribution de masse q_Z est une loi de probabilité: cette loi est précisément le profil de x .

Le profil p_X^z de z sur X n'est autre que la restriction de la loi marginale p_X à l'intervalle $(z_<, z_>)$ défini par z . Plus précisément, cette restriction doit être normalisée en la divisant par sa masse totale.

Il importe de démontrer que les deux transitions p_Z^X et p_X^Z sont des "transitions latéralement croissantes", au sens de la leçon [Ord. Lat.], TIB, n°8, du *Traité sur L'Analyse Des Données*; cf. aussi [MOD. CODE BARY.], in *CAD*, Vol. XV, n°4, (1990). En effet, selon un théorème de [Ord. Lat.], rappelé dans [MOD. CODE BARY.], ceci implique que le modèle général admet au moins un couple de facteurs associés relatif à la première valeur propre qui est un couple de fonctions croissantes; donc, en particulier, une fonction croissante sur le segment X . [En appendice, §6, ce théorème sera complété par un résultat d'unicité.] Ceci permet d'expliquer, *a posteriori*, comme une analogie avec le modèle, le succès généralement rencontré, dans la pratique, par l'interprétation du 1-er facteur.

De façon précise, on suppose l'intervalle X muni de sa structure d'ordre total usuelle; et Z , muni de la structure d'ordre partiel produit des structures d'ordres de ses coordonnées $z_<$ et $z_>$:

$$\text{Dr}(z, z') \Leftrightarrow (z_< < z'_<) \wedge (z_> < z'_>) .$$

Considérons d'abord la transition p_Z^X . Soit $x < x'$: les profils p_Z^x et $p_Z^{x'}$, associés à ces deux points, sont les restrictions de q_Z aux rectangles $\text{rec}(x)$ et $\text{rec}(x')$. On distinguera trois sous-rectangles:

$$\text{rec}(x - x') = \{z \mid z_< \leq x < z_> \leq x'\} = \text{rec}(x) - \text{rec}(x') ;$$

$$\text{rec}(x \wedge x') = \{z \mid z_< \leq x < x' < z_>\} = \text{rec}(x) \cap \text{rec}(x') ;$$

$$\text{rec}(x' - x) = \{z \mid x < z_< \leq x' < z_>\} = \text{rec}(x') - \text{rec}(x) ;$$

ces trois rectangles comprennent respectivement les intervalles qui contiennent x mais non x' ; les intervalles qui contiennent à la fois x et x' ; les intervalles qui contiennent x' mais non x . Ces trois rectangles sont deux à deux d'intersection vide, et on a les relations:

$$\text{rec}(x) = \text{rec}(x - x') \cup \text{rec}(x \wedge x') \quad ; \quad \text{rec}(x') = \text{rec}(x' - x) \cup \text{rec}(x \wedge x') ;$$

$$\text{Dr}(\text{rec}(x - x'), \text{rec}(x' - x)) ;$$

la dernière relation exprime le fait que, pour la relation d'ordre partiel définie sur Z , tout segment contenant x' mais non x est à droite de tout segment contenant x mais non x' .

Ceci posé, il est clair que la restriction de q_Z à $\text{rec}(x'-x)$ est à droite de sa restriction à $\text{rec}(x-x')$; et donc que le profil de x' , somme des restrictions à $\text{rec}(x-x')$ et $\text{rec}(x \wedge x')$, est à droite du profil de x , somme des restrictions à $\text{rec}(x'-x)$ et $\text{rec}(x \wedge x')$; ce qui exprime que p_X^X est latéralement croissant.

Considérons maintenant la transition p_X^Z . Soit z et z' tels que $\text{Dr}(z, z')$: les profils p_X^z et $p_X^{z'}$, associés aux deux segments, z et z' , sont les restrictions de p_X à ces segments; (plus précisément, ces restrictions normalisées pour avoir masse 1). Si le segment $[z'_<, z'_>[$ est tout entier à droite du segment $[z_<, z_>[$, la relation d'ordre latéral $\text{Dr}(p_X^z, p_X^{z'})$ est évidente: seul est à considérer le cas où ces segments ont une intersection non vide; soit :

$$z_< < z'_< < z_> < z'_> .$$

On distinguera les trois segments successifs:

$$z_1 = [z_<, z'_<[\quad ; \quad z_2 = [z'_<, z_>[\quad ; \quad z_3 = [z_>, z'_>[\quad ;$$

nous désignerons par $\{m^1, m^2, m^3\}$ les masses des restrictions respectives, $\{p_X^1, p_X^2, p_X^3\}$ de la loi marginale p_X à ces trois segments. Raisonnons, e.g., sous l'hypothèse que $(m_1 < m_3)$. Le profil p_X^z n'est autre que le quotient par $(1/(m_1+m_2))$ de la somme des restrictions de p_X à z_1 et à z_2 ; de même, $p_X^{z'}$ est le quotient par $(1/(m_3+m_2))$ de la somme des restrictions de p_X à z_3 et à z_2 ; on peut écrire:

$$p_X^z = (1/(m_1+m_2)) p_X^1 + (1/(m_3+m_2)) p_X^2 + ((1/(m_1+m_2)) - (1/(m_3+m_2))) p_X^2 \quad ;$$

$$p_X^{z'} = (1/(m_3+m_2)) p_X^2 + (1/(m_3+m_2)) p_X^3 \quad ;$$

sous cette forme, il apparaît que $p_X^{z'}$ comprend, d'une part, un terme en p_X^2 qui est aussi dans p_X^z ; et, d'autre part, un terme en p_X^3 qui, relativement au reste de p_X^z , i.e. à la somme:

$$(1/(m_1+m_2)) p_X^1 + ((1/(m_1+m_2)) - (1/(m_3+m_2))) p_X^2 \quad ,$$

a même masse et est distribué sur un support tout entier à la droite de celui de cette somme. D'où il résulte que $p_X^{Z'}$ est à droite de p_X^Z , i.e. que p_X^Z est latéralement croissant; ce qui permet d'appliquer le théorème, rappelé ci-dessus, relatif à l'existence de facteur croissant.

5 Transition composée et équation des facteurs

On sait que les couples de facteurs associés (φ^X, φ^Z) , relatifs à la valeur propre λ , satisfont au système:

$$\varphi^X \circ p_X^Z = \sqrt{(\lambda)} \varphi^Z \quad ; \quad \varphi^Z \circ p_Z^X = \sqrt{(\lambda)} \varphi^X \quad ;$$

d'où résulte, pour chacun des facteurs, une équation:

$$\varphi^X \circ p_X^Z \circ p_Z^X = \lambda \varphi^X \quad ; \quad \varphi^Z \circ p_Z^X \circ p_X^Z = \lambda \varphi^Z \quad ;$$

ceci nous incite à étudier, dans la cadre du modèle général, la transition composée $\pi_X^X = p_X^Z \circ p_Z^X$.

Les résultats obtenus au §4 permettent de calculer le profil π_X^X , associé à un point x de l'intervalle X :

$$\pi_X^X = p_X^Z \circ p_Z^X = \int \{p_X^Z dq_z \mid z \in \text{rec}(z)\} \quad ;$$

en d'autres termes, on combine les profils p_X^Z (obtenus chacun en restreignant la loi marginale p_X au segment $(z_<, z_>)$ défini par z , en se bornant aux z définissant un segment où x est contenu; les pondérations étant fournies par la mesure q_Z , dont on sait que l'intégrale sur $\text{rec}(x)$ vaut 1.

Si on passe de x à x' , situé à droite de x , disparaissent de l'intégrale les segments z qui contiennent x mais non x' ; et s'introduisent ceux qui contiennent x' mais non x . Ceci permet de vérifier que la transition π_X^X est latéralement croissante (fait qui nous est déjà connu parce que π_X^X est défini en composant deux transitions latéralement croissantes): car, en bref, les segments qui s'introduisent sont à droite de ceux qui disparaissent.

Dans le cas où la loi marginale p_X a une densité continue ne s'annulant en aucun point de X , on peut supposer que la coordonnée x est choisie de telle sorte que p_X soit la mesure de densité constante égale à 1; ce qu'on peut écrire: $p_X = dx$. Dans ce cas, le profil p_X^Z n'est autre que la loi $dx/(z_> - z_<)$ limitée au segment défini par z ; d'où, en reprenant les notations du §2:

$$p_X^z = (f(x, z)/(z_> - z_<)) \cdot dx ;$$

$$\pi_X^{x'} = \int \{f(x, z)/(z_> - z_<) dq_Z \mid z \in \text{rec}(x')\} \cdot dx = \pi(x, x') dx ;$$

où on a noté $\pi(x, x')$ la fonction de densité, par rapport à p_X , du profil $\pi_X^{x'}$ de x' ; la valeur $\pi(x, x')$ n'est autre que l'intégrale, étendue à l'ensemble des z qui définissent un segment contenant à la fois x et x' , de la fonction $(1/(z_> - z_<))$, pour la mesure q_Z ; $\pi(x, x')$ est donc une fonction symétrique de ses deux arguments. Pour x' donné, le maximum de $\pi(x, x')$, fonction de x , est atteint en $x=x'$; et plus précisément, la fonction est croissante en x pour $(x < x')$ et décroissante pour $(x > x')$.

Il est facile, sur cette formule, de calculer la trace de l'application linéaire π_X^X ; trace qui, sous l'hypothèse $p_X = dx$, n'est autre que l'intégrale $\int \{\pi(x, x) dx\}$; il vient:

$$\int \{\pi(x, x) dx\} = \iint \{f(x, z)/(z_> - z_<) dx dq_Z \mid x \in X ; z \in Z\} ;$$

en intégrant d'abord par rapport à x , on trouve:

$$\int \{f(x, z)/(z_> - z_<) dx \mid x \in X\} = 1 ;$$

car l'intervalle de X , qu'on délimite par $f(x, z)=1$, a précisément pour longueur $(z_> - z_<)$; (fait qu'on retrouverait aussi, hors de l'hypothèse $p_X = dx$, parce que la masse totale de p_X^z doit être 1); d'où finalement:

$$\int \{\pi(x, x) dx\} = \int \{dq_Z \mid z \in Z\} = \text{masse totale de } q_Z ;$$

on a déjà dit au §3 *in fine*, que cette masse totale s'interprète comme le nombre moyen de modalités des questions; et, compte tenu du facteur trivial constant relatif à la valeur propre 1, la masse de q_Z , diminuée de 1, donne la somme des valeurs propres (comptées avec leur multiplicité); cette somme est encore égale à la masse totale de q_X , c'est-à-dire au nombre moyen de points de séparation entre modalités.

6 Appendice: unicité du facteur croissant sur le segment X, associé à des transitions latéralement croissantes

L'occasion s'offre ici de compléter le théorème cité au début du §4. On a dit que le modèle général admet au moins un couple de facteurs associés relatif à la première valeur propre qui est un couple de fonctions croissantes; on montrera que, plus précisément, dans tout système orthogonal de facteurs sur X , il y a, au plus, un facteur croissant.

En effet, supposons qu'existent deux facteurs $\{\varphi_\alpha, \varphi_\beta\}$ croissants, de moyenne nulle et orthogonaux entre eux; ce qu'on écrira simplement:

$$\langle \varphi_\alpha, \varphi_\beta \rangle = \langle \varphi_\alpha, 1 \rangle = \langle 1, \varphi_\beta \rangle = 0 .$$

Parce que φ_β est croissante et de moyenne nulle, il existe au moins un point $x \in X$ tel que, pour $x' < x$, $\varphi_\beta(x')$ est ≤ 0 ; et pour $x' > x$, $\varphi_\beta(x')$ est ≥ 0 . En retranchant de la fonction φ_α sa valeur au point x' , on obtient une fonction φ'_α qui, comme combinaison de fonctions orthogonales à φ_β , est orthogonale à cette dernière fonction:

$$\langle \varphi'_\alpha, \varphi_\beta \rangle = 0 ;$$

or ce produit scalaire est nécessairement positif, car c'est l'intégrale d'un produit de fonctions non identiquement nulles qui sont de même signe aussi bien à gauche qu'à droite de x ; (le produit étant nul en x).

L'on aboutit donc à une contradiction d'où il résulte que ne peuvent exister deux facteurs $\{\varphi_\alpha, \varphi_\beta\}$ croissants et orthogonaux entre eux. Si, de plus, comme c'est le cas général, la première valeur propre est de multiplicité 1, le facteur φ_1 est l'unique facteur croissant; car tout autre facteur, lui étant orthogonal, ne peut être croissant.

Nous estimons que ce résultat éclaire les observations faites communément dans le plan (1,2) issu de l'analyse d'un système redondant de variables découpées en classes.

Référence bibliographique

J.-P. BENZÉCRI : Ordre latéral entre lois de probabilités sur un ensemble ordonné: applications à l'analyse factorielle et aux critères de classification, [Ord. Lat.]; in *L'Analyse Des Données*;, TIB, n°8; 1-ère édition, DUNOD, Paris, (1973).