MATHÉMATIQUES ET SCIENCES HUMAINES

ISRAËL CÉSAR LERMAN

Maximisation de l'association entre deux variables qualitatives ordinales

Mathématiques et sciences humaines, tome 100 (1987), p. 49-56

<http://www.numdam.org/item?id=MSH_1987__100__49_0>

Math. Sci. hum. (25^{e} année, n°100, *1987*, p.49-56)

MAXIMISATION DE L'ASSOCIATION ENTRE DEUX VARIABLES QUALITATIVES ORDINALES

Israël César LERMAN*

I. INTRODUCTION - POSITION DU PROBLEME.

La donnée est un couple de variables qualitatives ordinales définissant un couple de préordres totaux (ω, ϖ) sur un ensemble O d'objets. Nous désignerons par $\{A_i / 1 \leq i \leq I\}$ (resp. $\{B_j / 1 \leq j \leq J\}$) la suite ordonnée des classes en nombre de I (resp. J) définie par ω (resp. ϖ) sur O. On notera par a_i (resp. b_j) le cardinal de la classe A_i (resp. B_j), $1 \leq i \leq I$ (resp. $1 \leq j \leq J$). Nous noterons par

$$\{c_{ij} / 1 \leq i \leq I,\ 1 \leq j \leq J\} \qquad (1)$$

la table de contingence de "croisement" de ω et ϖ. De façon précise,

$$c_{ij} = \mathrm{card}(A_i \cap B_j),\ 1 \leq i \leq I,\ 1 \leq j \leq J.$$

On a bien entendu :

$a_i = \sum_{1 \leq j \leq J} c_{ij}$, $b_j = \sum_{1 \leq i \leq I} c_{ij}$ pour tout (i,j) de IxJ, où on note $I = \{1,2,\ldots,I\}$ et $J = \{1,2,\ldots,J\}$.

D'autre part,

$$n = \sum_{1 \leq i \leq I} a_i = \sum_{1 \leq j \leq J} b_j = \mathrm{card}(O).$$

On suppose -sans aucunement restreindre la généralité- que pour tout i de I (resp. j de J) $a_i \neq 0$ (resp. $b_j \neq 0$).

En d'autres termes, (1) définit ce que nous appellerons également sans risque d'ambiguité un "croisement" entre les deux partages ordonnés $\{a_i / 1 \leq i \leq I\}$ et $\{b_j / 1 \leq j \leq J\}$ de l'entier n ; lesquels définissent les marges du tableau (1).

---

* I.R.I.S.A - Rennes.

** L'auteur remercie les rapporteurs anonymes pour leurs utiles remarques, concernant notamment l'algorithme du "coin Nord-Ouest".

Considérons l'indice 'brut' suivant entre deux préordres totaux ω et ϖ sur O:

$$s(\omega, \varpi) = \Sigma\{ c_{ij} \times c_{lk} / 1 \leq i < l \leq I,\ 1 \leq j < k \leq J\}. \qquad (2)$$

$s(\omega, \varpi)$ est égal à $\text{card}[R(\omega) \cap R(\varpi)]$ où, en notant par ω (resp. ϖ) la relation de préordre total :

$$R(\omega) = \{ (x,y) \in O \times O / \omega(x,y) \text{ et } \neg[\omega(y,x)]$$
$$= \Sigma\{A_i \times A_l / 1 \leq i < l \leq I\}, \qquad (3)$$

où le symbole $\neg$ indique le 'non' logique et où la somme est ensembliste. De même,

$$R(\varpi) = \{(x,y) \in O \times O / \varpi(x,y) \text{ et } \neg[\omega(y,x)]\}$$
$$= \Sigma\{B_j \times B_k / 1 \leq j < k \leq J\}. \qquad (3')$$

Beaucoup de coefficients d'association entre variables qualitatives ordinales intègrent dans leur formation l'indice $s(\omega, \varpi)$. Ainsi en est-il des coefficients τ_a et τ_b de M.G. Kendall [1970] ou de celui γ de W.H. Kruskal [1958].

Divers modes de normalisation de $s(\omega, \varpi)$ peuvent être considérés pour aboutir à un indice réduit [Giakoumakis et Monjardet (1987)]. Nous avons nous-mêmes considéré une normalisation de type statistique où nous centrons par référence à la moyenne et où nous réduisons au moyen de l'écart type d'un indice brut $s(\omega^*, \varpi^*)$, où ω^* et ϖ^* sont deux préordres aléatoires indépendants sur O, respectivement associés à ω et ϖ [Lerman (1973),(1981)].

Beaucoup d'indices d'association entre deux préordres ω et ϖ -dont le nôtre- peuvent formellement se ramener à une fonction affine dont la définition ne dépend que des marges, de l'indice 'brut' $s(\omega, \varpi)$.

Le problème d'une forme de normalisation de ces indices passe par celui de la maximisation de l'indice brut $s(\omega, \varpi)$. Certains auteurs considèrent ce problème indépendamment de toute contrainte où il a une solution évidente (utilisée par Kendall pour définir τ_a) définie par $n(n-1)/2$. Nous pensons quant à nous qu'il est plus précis de considérer ce problème de maximisation sous contraintes de marges fixées :

$$\left.\begin{array}{l} \max\ [\Sigma\{c_{ij}\ c_{lk} / 1 \leq i < l \leq I,\ 1 \leq j < k \leq J\}] \\ \text{sous les contraintes} \\ \sum_j c_{ij} = a_i \text{ pour tout } i = 1,2, \ldots, I \\ \sum_i c_{ij} = b_j \text{ pour tout } j = 1,2, \ldots J \end{array}\right\} \qquad (4)$$

L'idée de base pour la solution de ce problème est de même nature que celle qui a servi à proposer une solution pour le problème de la maximisation de $\Sigma\{ c_{ij}^2 / (i,j) \in I \times J\}$, sous les mêmes contraintes [Lerman et

Peter (1986)] . Il s'agit de remplacer la notion de formule mathématique numérique par celle, algorithmique.

En toute rigueur, la solution proposée du problème mentionné repose sur une conjecture qu'il est difficile de ne pas admettre et qui a été démontrée dans certains cas importants. Toutefois des problèmes de complexité calcul peuvent se présenter.

C'est donc un algorithme récursif qui nous conduira à la solution du problème posé. Mais ici, la solution sera moins difficile et ne nécessitera pas -pour les cas courants- l'écriture d'un programme informatique.

Comme nous l'avons mentionné, nous avons rencontré ces problèmes de maximisation dans le cadre de la normalisation de coefficients d'association entre variables qualitatives. Ils peuvent en fait s'exprimer comme des problèmes particuliers de recherche opérationnelle de maximisation de transport quadratique.

En réalité cet algorithme se rattache à celui dit du 'coin nord ouest' qui est généralement introduit dans les problèmes de transport linéaire pour fournir une solution de départ qui n'est pas optimale [Simmonard (1962)] . Par contre, dans notre cas, au paragraphe III, un théorème général établira que l'algorithme fournit bien la solution optimale exacte à notre problème.

II. ALGORITHME.

Nous emprunterons le même langage que dans [Lerman et Peter (1986)] . On part du tableau de contingence vide à l'intérieur, mais ayant ses marges remplies qu'il s'agit de répartir de façon compatible et optimale. On sera conduit -à chaque pas- à installer le contenu d'une marge ligne α_i dans une colonne j (auquel cas $\alpha_i \leq \beta_j$: contenu de la marge colonne j), ou bien le contenu de la marge colonne β_j dans une ligne i (auquel cas $\alpha_i > \beta_j$). Au départ $\alpha_i = a_i$ pour $1 \leq i \leq I$ (resp. $\beta_j = b_j$ pour $1 \leq j \leq J$). On dira qu'on "vide" ("déverse" ou "décharge") α_i dans la colonne j, ou bien β_j dans la ligne i. Dans l'un ou l'autre des deux cas ($\alpha_i \leq \beta_j$ ou $\alpha_i > \beta_j$) on pourra dire qu'on procède à la "résolution" du couple (α_i, β_j).

Si on désigne par K l'entier (I+J), après chaque "résolution", la dimension K du problème diminue d'une unité ; puisque c'est soit une marge ligne, soit une marge colonne qui se vide.

A partir de là, l'expression de l'algorithme qui -encore une fois- est celui du "coin nord ouest" est très simple :

A chaque pas résoudre le couple origine le plus à gauche et en haut (i et j minimums).

Ainsi imaginons que $a_1 > b_1$ et $(a_1-b_1) < b_2$. Le premier couple résolu est (a_1,b_1). Cette première résolution vide la première colonne et laisse au niveau de la première ligne une nouvelle marge $\alpha_1 = (a_1-b_1)$. Le nouveau couple origine est alors (α_1,b_2). La résolution de ce dernier vide la première ligne car $\alpha_1 < b_2$. La nouvelle première marge colonne non vidée est $(b_2-\alpha_1) = b_1+b_2-a_1$ qu'il s'agit de comparer avec a_2 ; et ainsi de suite ...

Considérons pour chacune des marges $(a_1,a_2,\ldots,a_i,\ldots,a_I)$ et $(b_1, b_2,\ldots,b_j,\ldots,b_J)$, la suite des sommes des sections commençantes :

$$\left.\begin{array}{l}(a_1,\ a_1+a_2,\ \ldots,\ a_1+a_2+\ldots+a_i,\ldots,a_1+a_2+\ldots+a_I)\\ \text{et}\\ (b_1,\ b_1+b_2,\ldots,\ b_1+b_2+\ldots+b_j,\ldots,\ b_1+b_2+\ldots,b_J)\end{array}\right\} \quad (1)$$

La solution de l'algorithme et donc du problème (4) ci-dessus dépend uniquement du préordre total intercalant les sommes de la première suite par rapport aux sommes de la deuxième suite. Ainsi, considérons la situation suivante où I = 3 et J = 4 :

$$b_1 < a_1 < (b_1+b_2) < (a_1+a_2) < (b_1+b_2+b_3) < (a_1+a_2+a_3) = (b_1+b_2+b_3+b_4). \quad (2)$$

La première résolution concerne (a_1,b_1) et décharge b_1 au niveau de la première ligne. Comme $a_1 < (b_1+b_2)$, $(a_1-b_1) < b_2$ et $\alpha_1 = (a_1-b_1)$ est déversé au niveau de la deuxième colonne. Considérons alors $b_2 - (a_1-b_1) = (b_1+b_2) - a_1$ qu'il y a lieu de comparer avec a_2. On a [cf. (2)] $[(b_1+b_2) - a_1] < a_2$, $\beta_2 = [(b_1+b_2) - a_1]$ est déchargé au niveau de la deuxième ligne dont la nouvelle marge est $\alpha_2 = (a_1+a_2) - (b_1+b_2)$ qui est [cf. (2)] inférieur à b_3. α_2 est donc vidé dans la troisième colonne dont la nouvelle marge est $\beta_3 = (b_1+b_2+b_3) - (a_1+a_2)$, laquelle, plus petite que a_3, est déchargée au niveau de la troisième ligne. La dernière marge $\alpha_3 = (a_1+a_2+a_3) - (b_1+b_2+b_3)$ est égale à b_4 et se trouve nécessairement vidée dans la quatrième colonne.

Dans ces conditions, on peut très bien imaginer un programme donnant à partir du préordre total ci-dessus mentionné [e.g. (2)], la structure optimale du tableau de croisement ainsi que l'expression formelle -utilisant les symboles a_i et b_j, $(i,j) \in I \times J$ - de la valeur maximale de $s(\omega, \varpi)$, associée.

Considérons à présent l'exemple numérique :

25	5	0	0	30
0	5	5	0	10
0	0	9	11	20
25	10	14	11	60

$s(\omega, \sigma) = 25 \times 20 + 5 \times 25 + 5 \times 20 + 5 \times 11 = 780$

III. THEOREME.

Nous allons maintenant établir le théorème qui permet de justifier que l'algorithme présenté au paragraphe II ci-dessus conduit bien à la solution optimale du problème (4) du paragraphe I ci-dessus.

III.1. Préliminaires

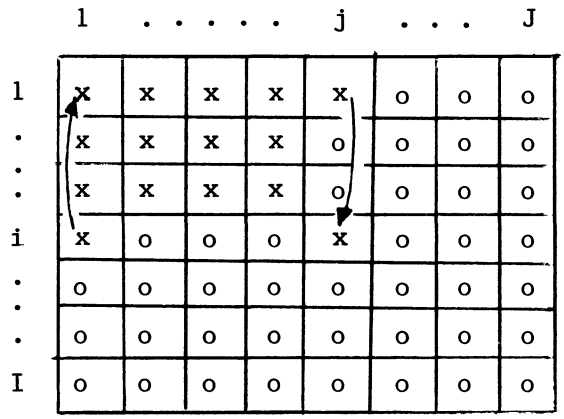

Figure 1 : Représentation d'une opération élémentaire affectant la case (1,1).

Notons

$$D(i,j) = \Sigma\{c_{i',j'} / i' > i \text{ et } j' > j\} ; \quad (1)$$

il s'agit -strictement au-delà de (i,j)- de la charge finissante (droite-bas) de la table de contingence de croisement des deux préordres totaux. Avec la notation (1), on a :

$$s(\omega, \sigma) = \Sigma\{c_{ij} D_{ij} / 1 \leq i \leq (I-1),\ 1 \leq j \leq (J-1)\} \quad (2)$$

Nous avons déjà introduit [Lerman et Peter (1986)], la notion d'opération élémentaire. Elle affectera ici la première case (1,1) du tableau (cf. Figure 1) et correspondra à la transformation suivante :

$c_{11} \rightarrow (c_{11}+1)$ $\qquad$ $c_{1j} \rightarrow (c_{1j}-1)$

$c_{i1} \rightarrow (c_{i1}-1)$ $\qquad$ $c_{ij} \rightarrow (c_{ij}+1)$

Rappelons -ce qui est clair- que cette transformation préserve les marges. Etudions la variation qu'elle entraîne sur $s(\omega, \varpi)$. Nous avons sur la figure noté par une croix x, toutes les cases dont la contribution change et par un petit rond o, toutes les cases dont la contribution est invariable par rapport à l'expression (2). En notant par $\tau_{[(1,1),(i,j)]}$ la transformation élémentaire ci-dessus, on a à calculer, en considérant chacune des cases munie d'une croix

$$\Delta = [s(\omega^\tau, \varpi^\tau) - s(\omega, \varpi)] \qquad (3)$$

où $(\omega^\tau, \varpi^\tau)$ est l'image par τ de (ω, ϖ).

Précisons que l'ensemble des cases munies d'une croix se définit comme suit :

$$\mathcal{C}[(1,1),(i,j)] = \{(i',j')/i' < i, j' < j\} \cup \{(i,1),(1,j),(i,j)\}. \qquad (4)$$

$$\begin{aligned}\Delta = &\{[c(1,1) + 1]\ [D(1,1) + 1)] - c(1,1)\ D(1,1)\} \\ &+ \Sigma\{c(i',j')/i' < i,\ j' < j \text{ et } (i',j') \neq (1,1)\} \\ &+ \{[c(1,j) - 1]\ D(1,j) - c(1,j)\ D(1,j)\} \\ &+ \{[c(i,1) - 1]\ D(i,1) - c(i,1)\ D(i,1)\} \\ &+ \{[c(i,j) + 1]\ D(i,j) - c(i,j)\ D(i,j)\} \end{aligned} \qquad (5)$$

En notant

$$G(i,j) = \Sigma\{c(i',j')/i' < i,\ j' < j\}, \qquad (6)$$

la charge -strictement en deçà de (i,j)- commençante (gauche-haut) de la table de contingence, on obtient :

$$\Delta = D(i,j) + G(i,j) + [D(1,1) - D(1,j) - D(i,1)] + 1 \qquad (7)$$

Comme on montre aisément que

$$D(i,j) + D(1,1) - D(i,j) - D(i,1) = \Sigma\{c_{i',j'}/1 < i' \leq i,\ 1 < j' \leq j\}, \qquad (8)$$

on a le

LEMME. L'accroissement Δ résultant de l'opération élémentaire $\tau_{[(1,1),(i,j)]}$ est positif strictement.

III.2. Théorème

THEOREME. L'algorithme du paragraphe II conduit à la solution optimale du problème (4) de maximisation du paragraphe I.

La démonstration devient à présent simple et se fait par récurrence sur K = I + J.

Il est facile de vérifier la propriété pour K = 4 et d'ailleurs de façon analogue à l'établissement du lemme.

Supposons la propriété vraie pour tout $K < (I+J)$ et démontrons la pour K = (I+J).

Relativement au tableau de type

$$t = (a_1,a_2,\ldots,a_i,\ldots,a_I \; ; \; b_1,b_2,\ldots,b_j,\ldots,b_J), \tag{9}$$

le lemme précédent nous montre que l'on a nécessairement

$$c_{11} = \min(a_1,b_1) \; ; \tag{10}$$

sinon, il existerait nécessairement une transformation élémentaire $\tau_{[\,(1,1),(i,j)\,]}$, laquelle -on l'a vu- améliore strictement le critère $s(\omega,\varpi;t)$ attaché au type (9) ci-dessus. En supposant comme ci-dessus (cf. § II) et sans restreindre la généralité $a_1 \geq b_1$, la première colonne est -pour la configuration optimale- vide en dehors de sa première ligne qui contient b_1. Dans ces conditions, la valeur maximale du critère se met sous la forme

$$b_1(n-a_1) + s_m[\,(a_1-b_1),a_2,\ldots,a_I \; ; \; b_2,\ldots,b_J\,], \tag{11}$$

où nous avons noté $s_m[\,.\,]$ la valeur maximale pouvant être atteinte par $s(\omega,\varpi)$ pour un couple (ω,ϖ) de préordres totaux de type $[\,.\,]$. Mais alors ce dernier type correspond à $K = (I + J - 1)$, C.Q.F.D.

BIBLIOGRAPHIE

[1] V. Giakoumakis et B. Monjardet (1987) ; *"Coefficients d'accord entre deux préordres totaux"*, Statistique et Analyse de données, à paraître.

[2] M.G. Kendall (1970) ; *"Rank Correlation Methods"*, 4ème édition (1ère édition en 1948), Charles Griffin, London.

[3] W.H. Kruskal (1958) ; *"Ordinal measures of association"*, J.A.S.A. vol. 53, Déc. 1958.

[4] I.C. Lerman (1973) ; *"Etude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique"*, Cahiers du B.U.R.O. n° 19, Paris.

[5] I.C. Lerman (1981) ; *"Classification et analyse ordinale des données"*, Dunod, Paris.

[6] I.C. Lerman et Ph. Peter (1986) ; *"Structure maximale pour la somme des carrés d'une contingence aux marges fixées. Une solution algorithmique programmée"*, Pub. Int. n° 318, IRISA-Rennes, Oct. 86.

[7] M. Simmonard (1962) ; *"Programmation linéaire"*, Dunod, Paris.