Maximisation de l'association entre deux variables qualitatives ordinales

Israël-César Lerman

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
tél. (1) 39 63 55 11

Rapports de Recherche

N° 627

MAXIMISATION DE L'ASSOCIATION ENTRE DEUX VARIABLES QUALITATIVES ORDINALES

Israël - César LERMAN

Février 1987

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042-RENNES CÉDEX
FRANCE
Téléphone: 99 36 20 00
Télex: UNIRISA 950 473 F
Télécopie: 99 38 38 32

janvier 1987

Publication Interne n° 341 - 18 Pages

MAXIMISATION DE L'ASSOCIATION ENTRE DEUX VARIABLES QUALITATIVES ORDINALES

Israël-César LERMAN

RESUME :

On considère un indice 'brut' qui sert de base à la construction d'un large éventail de coefficients d'association ou d'accord entre deux préordres totaux sur un ensemble fini d'objets. A des fins de normalisation de tels coefficients, nous nous posons le problème de la construction algorithmique de la configuration du tableau de contingence du croisement entre deux préordres totaux, maximisant la valeur de l'indice 'brut'. Ce problème est posé en cas de marges fixes ('compositions' fixées des deux préordres totaux), ce qui permet d'obtenir une relation statistique privilégiée de transition de l'un des préordres totaux à l'autre.

MAXIMAL ASSOCIATION BETWEEN TWO ORDINAL QUALITATIVE VARIABLES

Israël-César LERMAN

ABSTRACT :

We consider a raw index which can be used as a basis of the construction of a large family of association coefficients between two total preorders on a finite set of objects. To standardize such coefficients, we study the problem of an algorithmic construction of a specific configuration of the contingency table which crosses the two total preorders. This last is obtained in order to maximise the raw index, under the constraints of fixed margins. The obtained configuration provides an optimal statistical transition relationship from one preorder to the other.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L. A. 227)
UNIVERSITÉ DE RENNES 1 I. N. S. A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

MAXIMISATION DE L'ASSOCIATION ENTRE DEUX VARIABLES QUALITATIVES ORDINALES.

I- INTRODUCTION ; POSITION DU PROBLEME.

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

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

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

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

On a bien entendu :

$$a_i = \sum_{1 \leqslant j \leqslant J} c_{ij} \quad , \quad b_j = \sum_{1 \leqslant i \leqslant I} c_{ij} \quad \text{pour tout } (i,j) \text{ de}$$

$\mathbb{I} \times \mathbb{J}$, où on note $\mathbb{I} = \{1, 2, \ldots, I\}$ et $\mathbb{J} = \{1, 2, \ldots, J\}$.

D'autre part,

$$n = \sum_{1 \leqslant i \leqslant I} a_i = \sum_{1 \leqslant j \leqslant J} b_j = \text{card}(\mathcal{O}).$$

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

En d'autres termes, (1) définit un croisement entre les deux partages ordonnés $\{a_i / 1 \leqslant i \leqslant I\}$ et $\{b_j / 1 \leqslant j \leqslant J\}$ de l'entier n, lesquels définissant les marges du tableau (1).

Beaucoup d'indices d'association entre deux préordres totaux ω et Θ [Giakoumakis et Monjardet (1987)] dont le nôtre [Lerman (1973), (1981)] peuvent formellement se ramener à une fonction affine – dont la définition ne dépend que des marges – de l'indice 'brut' :

$$s(\omega, \Theta) = \sum \{ c_{ij} c_{lk} / 1 \leqslant i < l \leqslant I, 1 \leqslant j < k \leqslant J \}. \quad (2)$$

$s(\omega, \varpi)$ représente card $[R(\omega) \cap R(\varpi)]$ où, en notant par ω (resp. ϖ) la relation de préordre total :

$$R(\omega) = \{(x,y) \in O \times O / \omega(x,y) \text{ et } \neg[\omega(y,x)]\}$$

$$= \subset \{ A_i \times A_\ell / 1 \leqslant i < \ell \leqslant I \}, \quad (3)$$

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

$$R(\varpi) = \{(x,y) \in O \times O / \varpi(x,y) \text{ et } \neg[\varpi(y,x)]\}$$

$$= \subset \{ B_j \times B_k / 1 \leqslant j < k \leqslant J \}. \quad (3')$$

Le problème de la 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 et n'arrivent d'ailleurs qu'à des résultats très parcellaires [Giakoumakis et Monjardet (1987)]. Nous pensons quant à nous qu'il est plus juste et plus précis de considérer ce problème de maximisation sous contraintes de marges fixées :

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

L'idée de base pour la solution de ce problème est de même nature que celle qui a servi à résoudre le problème de la maximisation de $\sum \{ c_{ij}^2 / (i,j) \in I \times J \}$, sous les mêmes contraintes [Lerman et Peter (1986)]. Il s'agit de remplacer la notion de formule mathématique numérique par celle, algorithmique. C'est donc un algorithme récursif qui nous conduira à la solution du problème posé. Mais ici, la solution sera moins difficile et ne necessitera pas – pour les cas courants – l'écriture d'un programme informatique. Nous expliciterons l'algorithme au paragraphe II en l'illustrant par un exemple. 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 $\alpha_i \leqslant \beta_j$: 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 $1 \leqslant i \leqslant I$ (resp. $\beta_j = b_j$ pour $1 \leqslant j \leqslant J$). 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 ($\alpha_i < \beta_j$ ou $\alpha_i > \beta_j$) on pourra dire qu'on procède à la "résolution" du couple (i, 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 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 a 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, $(i,j) \in \mathbb{I} \times \mathbb{J}$ – de la valeur maximale de $s(w, \varpi)$, associée.

Considérons à présent l'exemple numérique :

25	5	0	0	30
0	10	0	0	10
0	0	12	8	20
25	15	12	8	

$$s(w, \varpi) = 25 \times 20 + 5 \times 20 + 10 \times 20 = 800$$

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.

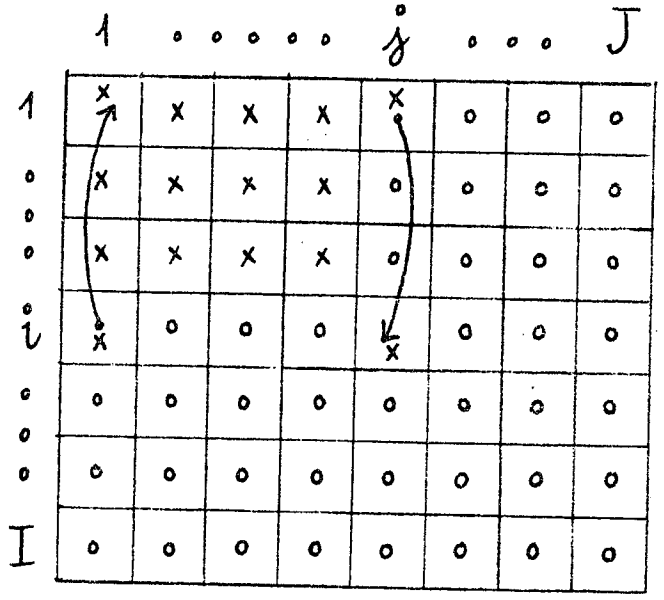

Figure 1: Représentation d'une opération élémentaire affectant la case (1,1).

Notons

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

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 :

$$s(\omega, \varpi) = \sum \{ c_{ij} D_{ij} / 1 \leqslant i \leqslant (I-1), 1 \leqslant j \leqslant (J-1) \}. \quad (2)$$

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 ∘, 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 :

$$\Delta = \left\{ \tau_{[(1,1),(i,j)]} [\, s(\omega, \varpi)] - s(\omega, \varpi) \right\} \quad . \quad (3)$$

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

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

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

En notant

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

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

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

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

Le lemme devient aisé à voir dès lors qu'on dispose de l'expression (7) ci-dessus. Le mieux pour le faire est un schéma où on commence par voir ce que représente $[D(1,1) - D(1,j) - D(i,1)]$. On marquera ⊕ les cases dont les charges sont à prendre positivement et ⊖ les cases dont les charges sont à prendre négativement (cf. Figure 2). La somme des charges des cases à prendre négativement est exactement $D(i,j)$; de sorte que l'accroissement Δ est exactement :

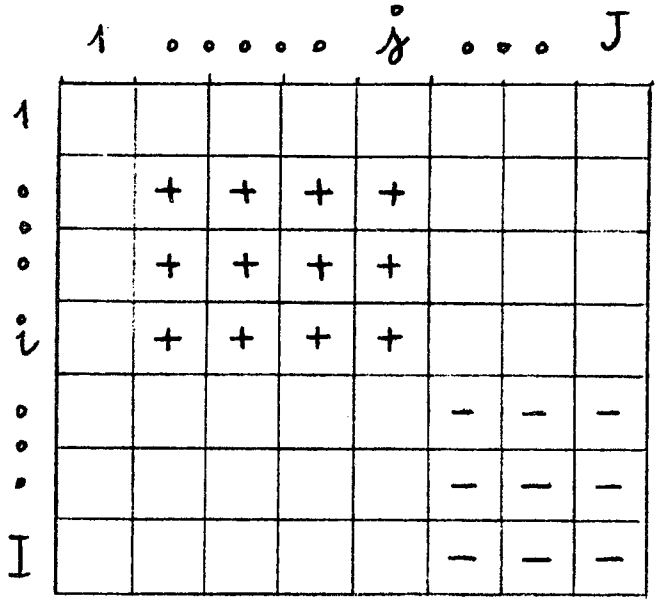

Figure 2 : $[D(1,1) - D(1,j) - D(i,1)]$

$$\Delta = 1 + G(i,j) + \{c_{i'j'} / 1 < i' \leq i,\ 1 < j' \leq j\} \quad , (8)$$

qui est strictement positif.

III.2. Théorème. 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 $K = 1 + J$.

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

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

le lemme précédent nous montre que l'on a nécessairement

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

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 $a_1 \geqslant b_1$, 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

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

où nous avons noté $s_m [\,\cdot\,]$ la valeur maximale pouvant

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

Références

[1] V. Giakoumakis et B. Monjardet (1987) ; "*Coefficients d'accord entre deux préordres totaux*", Texte ronéotypé, Centre d'Analyse et de Mathématique Sociales, Paris.

[2] 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.

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

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

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique