LES CAHIERS DE L'ANALYSE DES DONNÉES

V. CHOLAKIAN

Sur l'information associée aux différents tableaux binaires issus d'un tableau ternaire

Les cahiers de l'analyse des données, tome 8, n° 1 (1983), p. 7-9
<http://www.numdam.org/item?id=CAD_1983__8_1_7_0>

Les Cahiers de l'Analyse des Données
Vol. VIII - 1983 - n° 1 - p. 7-9

SUR L'INFORMATION ASSOCIÉE AUX DIFFÉRENTS TABLEAUX BINAIRES ISSUS D'UN TABLEAU TERNAIRE

[INF. TERN.]

par V. Cholakian (*) (1)

Le but de cette note est de montrer en utilisant l'information de Shannon, les liens qui existent entre les différents tableaux binaires (rectangulaires) issus d'un tableau ternaire. A la fin nous suggérons des généralisations au cas d'un tableau multiple de rang quelconque (r>3). Il est à souligner que les relations démontrées ici ne s'étendent pas directement aux calculs de χ^2.

1 *Information et correspondance ternaire*

Soit k_{IJT} un tableau ternaire dont tous les éléments $k(i,j,t)$ sont positifs ; on peut présenter ce tableau ternaire de trois manières différentes comme un tableau rectangulaire : soit $k_{(IJ)T}$, $k_{(IT)J}$, $k_{(JT)I}$; et de plus il y a 3 tableaux rectangulaires de marge notés k_{IJ}, k_{IT}, k_{JT}. A tous ces tableaux sont associés des lois de probabilités qu'on note classiquement en utilisant la lettre p au lieu de k.

On reprend les résultats et les notations de [*INF. TAB.*] (TIB n° 5) :

$H(p_I) = -\Sigma\{p_i \log_2 p_i \mid i \in I\}$ = information correspondant à la variable aléatoire i dont la loi est p_I.

$H(p_{IJ}) = -\Sigma\{p_{ij} \log_2 p_{ij} \mid i \in I\ ,\ j \in J\}$ = information correspondant à la variable aléatoire (i,j) dont la loi est p_{IJ}.

$H(p^i{}_J)$ = information conditionnelle apportée par j quand i est connu (avec $p^i{}_J = \{p^i{}_j \mid j \in J\}$; $p^i{}_j = p_{ij}/p_i$).

(1) $H(J/I) = \Sigma\{p_i\, H(p^i{}_J) \mid i \in I\} = H(p_{IJ}) - H(p_I)$ = espérance mathématique de l'information conditionnelle apportée par j sur I.

(2)
$$\begin{aligned} H(p_{IJ}\,;p_I\,p_J) &= H(p_J) - H(J/I) = H(p_I) - H(I/J) \\ &= H(p_I) + H(p_J) - H(p_{IJ}) \\ &= \Sigma\{p_{ij}\log_2(p_{ij}/(p_i\,p_j)) \mid i\in I; j\in J\} \\ &= \text{information mutuelle entre I et J.} \end{aligned}$$

(1) Université de Moncton ; dept de phys. mathématique ; Moncton Nouveau Brunswick E1A 3E9.

On va démontrer le résultat suivant :

(3) $$H(p_{IJ} ; p_I p_J) + H(p_{(IJ)T} ; p_{(IJ)} p_T)$$

$$= H(p_{IT} ; p_I p_T) + H(p_{(IT)J} ; p_{(IT)} p_J)$$

$$= H(p_{JT} ; p_J p_T) + H(p_{(JT)I} ; p_{(JT)} p_I) ;$$

démonstration : dans ces formules il s'agit seulement d'information mutuelle associée à divers tableaux rectangulaires : ou bien des lois marginales (sur IJ, IT, JT), ou bien des tableaux ternaires présentés sous forme binaire : par exemple $p_{(IT)J}$ n'est autre que le tableau p_{ITJ} avec en lignes les (i , t), en colonnes les j .

Les diverses informations mutuelles associées à ces 6 tableaux se calculent toutes en fonction des informations correspondant à une seule variable, à une paire de variables, ou au triplet . On a par exemple (cf. (2))

(4) $$H(p_{(IJ)T} ; p_{(IJ)} p_T) = H(p_{IJ}) + H(p_T) - H(p_{IJT}) ,$$

où on a tenu compte de ce que $H(p_{(IJ)T})$ n'est autre que $H(p_{IJT})$. Ceci posé, les trois sommes dont (3) affirme l'égalité, apparaissent comme ayant toutes la même valeur qui est :

(5) $$H(p_I) + H(p_J) + H(p_T) - H(p_{IJT}) ;$$

on peut dire que (5) est "l'information mutuelle relative au triplet (I,J,T)", en ce sens que (5) exprime le déficit de l'information apportée par (i,j,t) relativement à la somme de celles apportées séparément par i par j et par t.

On démontrerait de même :

(6) $$H(I/(JT)) - H(I/T) = H(J/(IT)) - H(J/T) .$$

On sait (cf. [*INF. TAB.*]) que l'information mutuelle est équivalente (au voisinage de l'indépendance) à une distance du χ^2 :

$$(2 \operatorname{Log} 2) . H(p_{IJ} ; p_I \, p_J) \approx \| p_{IJ} - p_I \, p_J \|^2 = \text{trace } (I \times J) .$$

Ceci suggère de comparer nos relations (3) aux valeurs des χ^2 correspondants. Par exemple en se référant à l'article de Ch. Arbache (C.A.D. Vol VII n° 1, p. 71, 1982) où l'auteur analyse un tableau ternaire, $k_{P \times R \times T}$, on trouve :

$$\text{trace}(P \times T) + \text{trace}((PT) \times R) = 0{,}088 + 0{,}854 = 0{,}942 ;$$
$$\text{trace}(R \times T) + \text{trace}((RT) \times P) = 0{,}023 + 0{,}985 = 1{,}008 ;$$
$$\text{trace}(R \times P) + \text{trace}((RP) \times T) = 0{,}736 + 0{,}196 = 0{,}932 .$$

Ces quantités diffèrent assez peu, mais elles ne sont pas égales; toutefois il peut être intéressant de considérer leur moyenne. Une comparaison approfondie avec les résultats de A. Bener ([*INTER. CORR. MULT.*] in C.A.D. Vol VII n° 1, pp 25-32 ; 1982) serait sans doute fructueuse.

2 *Remarque : essai de généralisation à une correspondance multiple*

Considérons des ensembles finis J_q indicés par $q \in Q$; notons pour toute partie a de Q :

(7) $Ja = \Pi\{Jq | q \in a\}$.

Soit p_{JQ} une loi sur le produit JQ de tous les Jq. On peut sans confusion noter p_{Ja} la loi marginale de p_{JQ} sur le produit Ja ; en particulier p_{Jq} est la loi marginale à une variable sur Jq. Et pour l'information on notera, quelle que soit la partie a de Q :

(8) $H(a) = H(p_{Ja})$;

et d'après la formule (5) on définira aussi :

$$HM(a) = -H(a) + \Sigma\{H(\{q\}) | q \in a\} ;$$

ceci posé si a et b sont deux parties de Q, d'intersection vide ($a \cap b) = \emptyset$) on peut s'intéresser à :

(9) $$H(p_{Ja\cup b} ; p_{Ja}\, p_{Jb}) = H(a)+H(b)-H(a\cup b)$$
$$= HM(a\cup b)-HM(a)-HM(b),$$

quantité qui sera notée: $H(a\cup b\, ; a,b)$,

et qui est une information mutuelle usuelle associée à un tableau rectangulaire.

Or il est remarquable que les $H(a\cup b\, ; a,b)$ sont très nombreux, autant que les paires (a,b) de parties de Q d'intersection vide ; alors que les HM(a) le sont beaucoup moins (autant que les parties de a) ; plus précisément il faut tenir compte de ce que $H(a\cup b\, ; a,b)$ est non identiquement nul seulement pour a et b non vides ; et de même HM(a) n'est non nul que si Card a ≥ 2. Donc si Card Q = 3, cas vu ci-dessus, on a 6 termes $H(a\cup b\, ; a,b)$ non nuls ; qui correspondent aux 6 tableaux binaires que nous avons construits, i.e. :

$$(a,b) = (\{I,J\}T) ; (\{I,T\},J) ; (\{J,T\},I) ; (I,J) ; (J,T) ; (T,I)$$

et on les exprime en combinaisons linéaires des HM(a) non nuls qui ne sont que 4 :

$$a = \{I,J,T\} ; \{I,J\} ; \{J,T\} ; \{I,T\}.$$

C'est pourquoi il y a entre les $H(a\cup b\, ; a,b)$ deux (2 = 6 - 4) relations linéaires que nous avons démontrées en (3).

Soit maintenant Card Q = 4 : Q = {1,2,3,4}. Il y a 11 parties a de cardinal ≥ 2 ; et il y a 19 paires (a,b) de paires de parties (a,b) non vides d'intersection vide (il y a 4 paires de la forme (123 ,4); 3 de la forme (12 , 34); 6 de la forme 12 , 3); 6 de la forme (1,2)). Ainsi 19 quantités $H(a\cup b\, ; a,b)$ s'expriment en combinaison de 11 quantités H(a) : il y donc entre les $H(a\cup b\, ; a,b)$ (informations mutuelles de tableaux binaires associés au tableau quaternaire) huit (8 = 19 - 11) relations linéaires.

Le nombre des relations linéaires croît rapidement avec Card Q.

Il importe de noter que bien que les $H(a\cup b\, ; a,b)$ soient (comme on l'a dit à propos du cas ternaire) équivalentes à des χ^2, les χ^2 eux-mêmes $\| p_{Ja\cup b} - p_{Ja}\, p_{Jb} \|^2$ ne satisfont pas à toutes ces relations linéaires.