Comparaison de partitions avec des marges fixées ; développements récents

Israël-César Lerman

To cite this version:

Israël-César Lerman. Comparaison de partitions avec des marges fixées ; développements récents. [Rapport de recherche] RR-0701, INRIA. 1987. inria-00075852

HAL Id: inria-00075852
https://inria.hal.science/inria-00075852
Submitted on 24 May 2006

INRIA

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° 701

COMPARAISON DE PARTITIONS AVEC DES MARGES FIXEES ; DEVELOPPEMENTS RECENTS

Israel-César LERMAN

JUILLET 1987

IRISA INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE

Publication Interne n° 361 - Mai 1987 - 14 Pages

COMPARAISON DE PARTITIONS AVEC DES MARGES FIXEES ; DEVELOPPEMENTS RECENTS

Israël-César LERMAN

RESUME

L'analyse formelle de la plupart des coefficients de comparaison entre deux partitions, montre que la contrainte résultant de la structure de la relation 'partition', n'intervient absolument pas. Cette contrainte doit se manifester au niveau de la normalisation ; c'est-à-dire, du dénominateur du coefficient. Cette normalisation peut être de nature formelle ou statistique. Pour la première, on montre comment -en remplaçant la notion de formule mathématique par celle d'algorithme récursif- on résoud un problème d'optimisation combinatoire jusque là réputé très difficile. Comme le montre l'analyse formelle et asymptotique du coefficient obtenu, la normalisation statistique -par rapport à une hypothèse adéquate d'absence de liaison- tient étroitement compte de la contrainte.

I.C. LERMAN ; 'Classification et Analyse ordinale des données', Dunod, Paris (1981).
I.C. LERMAN et Ph. PETER ; 'Structure maximale pour la somme des carrés d'une contingence aux marges fixées. Une solution algorithmique programmée, Publ. Int. n°318, IRISA-RENNES, Octobre 1986, 90 pages.
L.J. HUBERT et P. ARABIE ; 'Comparing partitions', Journal of Classification, 2,2-3 (1985) 193-218.

COMPARING PARTITIONS WITH FIXED MARGINS ; RECENT DEVELOPEMENTS

Israël-César LERMAN

ABSTRACT

The formal analysis of the most comparison coefficients between two partitions, shows the non intervention of the relational constraint which results of the partition structure. This constraint must appear in the standardization of the considered coefficient ; that is to say, at the level of the coefficient denominator. This standardization has 'formal' or 'statistical' nature. For the first, we show how -by replacing the notion of mathematical formulae by this one of recursive algorithm- we resolve a combinatorial optimization problem which has been considered as yet as very difficult. On the other hand, we study the asymptotic formal expression of the coefficient obtained by statistical standardization with respect to an adequat hypothesis of no relation. This last coefficient takes closely into account the relational constraint.

HUBERT, L.J., ARABIE, P., Comparing Partitions, Journal of Classification, 2,2-3 (1985) 193-218.
LERMAN, I.C., Classification et Analyse Ordinale des Données, Dunod, Paris (1981).
LERMAN, I.C., PETER, Ph., Structure maximale pour la somme des carrés d'une contingence aux marges fixées. Une solution algorithmique programmée, Publ. Int. n°318, IRISA-RENNES, Octobre 1986, 90 pages.

I. INTRODUCTION

Depuis l'origine de nos recherches, le point de vue que nous adoptons correspond à considérer une variable qualitative descriptive d'un ensemble O d'objets, comme définissant une relation sur l'ensemble O qu'on peut -selon sa complexité- représenter de façon ensembliste au niveau de O, de OxO ou même de (OxO)x(OxO) [Lerman (1970),(1973),(1981),(1987)]. Le cas d'une variable qualitative nominale c est un cas particulier d'une variable relationnelle qui induit une partition $\pi=\{A_i/1\leq i\leq I\}$ en I classes non vides. On peut représenter une telle partition au niveau de l'ensemble $O^{\{2\}}=P_2(O)$ des paires d'objets distincts, par le sous ensemble $R(\pi)$ des paires d'objets qu'elle réunit, plus précisément

$R(\pi)=\sum\{A_i^{\{2\}}=P_2(A_i)/1\leq i\leq I\}$ (somme ensembliste)

Si on introduit le sous ensemble $S(\pi)$ des paires d'objets que la partition sépare:

$S(\pi)=\sum\{A_i*A_{i'}/1\leq i\leq i'\leq I\}$,

où $A_i*A_{i'}=\{\{x,y\ /x\in A_i$ et $y\in A_{i'}\}$, $R(\pi)$ et $S(\pi)$ forment une partition en deux classes de $O^{\{2\}}$.

La comparaison de deux partitions π et χ peut alors s'exprimer en termes de comparaison de parties d'un même ensemble qui est l'ensemble $O^{\{2\}}$ des paires d'objets distincts. On peut en effet associer à $\chi=\{B_j/1\leq j\leq J\}$, les ensembles de représentation $R(\chi)$ et $S(\chi)$. Un certain nombre de coefficients dont on parle dans la littérature Anglophone, utilisent les indices bruts suivants : $s=card[R(\pi)\cap R(\chi)]$, $u=card[R(\pi)\cap R^c(\chi)|$, $v=card[R^c(\pi)\cap R(\chi)]$ et $t=card[R^c(\pi)\cap R^c(\chi)]$ où $R^c(\pi)$ [resp. $R^c(\chi)$] désigne la partie complémentaire dans $O^{\{2\}}$ de $R(\pi)$ [resp. $R(\chi)$]. Citons par exemple :

Rand(1971): $[(s+t)/(s+u+v+t)]$. (1)
Jaccard(1908): $[s/(s+u+v)]$. (2)
Fowlkes and Mallows (1983): $\{s/\sqrt{[(s+u)(s+v)]}\}$. (3)

Si l'indice (2) est correctement attribué à Jaccard, il n'en est pas tout à fait de même pour chacun des deux autres indices qui ont été considérés bien avant, mais dans un tout autre contexte. Vers les années 50 à 60 sont apparus de nombreux indices de similarité entre individus (taxons) décrits -dans le cadre d'une table d'incidence- par des attributs logiques de présence-absence. Si C est l'ensemble des caractéristiques, on peut représenter un même individu x par le sous ensemble C_x des attributs qu'il possède. La comparaison de deux individus x et y se ramène à celle de deux parties C_x et C_y de C, à partir des paramètres : $s=card(C_x\cap C_y)$, $u=card(C_x\cap C_y^c)$, $v=card(C_x^c\cap C_y)$ et $t=card(C_x^c\cap C_y^c)$ où C_x(resp.C_y^c) désigne la partie complémentaire dans C de C_x (resp.C_y).

Dans ces conditions, on voit très vite que le coefficient de Rand n'est autre que celui de Sokal et Michener (1958) qui se met également sous la forme $\{1-[(u+v)/c]\}$ où $c=card(C)=s+u+v+t$. D'autre part, le coefficient de Fowlkes and Mallows n'est autre que celui d'Ochiai(1957). De plus, un coefficient de type 'Goodman et Kruskal'(1954) $\{[(s+t)-(u+v)]/(s+t+u+v)\}$ peut s'écrire $\{1-[2(u+v)/c]\}$ et correspond exactement à l'indice de Hamann(1961). Si maintenant, on regarde l'objet x (resp.y) comme représenté par un préordre total à deux classes C_x^c et C_x, avec $C_x^c<C_x$ (resp. C_y^c et C_y avec $C_y^c<C_y$), le coefficient de Goodman et Kruskal n'est autre que celui de Yule [(1911),(1912)]: $[(st-uv)/(st+uv)]$.

Ainsi, si on réduit la comparaison de deux partitions π et χ à la comparaison de deux parties $R(\pi)$ et $R(\chi)$ d'un même ensemble fini $O^{\{2\}}$, sans s'interroger plus avant sur la nature des structures que l'on compare, on peut considérer n'importe lequel des indices de similarité considérés par les taxinomistes pour comparer des taxons décrits par des attributs logiques. Quitte ensuite à expliciter ce qu'on obtient par rapport au tableau de contingence $\{c_{ij}/1\leq i\leq I,\ 1\leq j\leq J\}$ de croisement des deux partitions π et χ ; $c_{ij}=\mathrm{card}(A_i \cap B_j)$, $1\leq i\leq I$, $1\leq j\leq J$. On posera aussi $a_i=\mathrm{card}(A_i)$, $b_j=\mathrm{card}(B_j)$, $1\leq i\leq I$, $1\leq j\leq J$. On a en effet,

$$\mathrm{card}[O^{\{2\}}]=\binom{n}{2}, \text{ où } n \text{ est le cardinal de l'ensemble des objets,}$$
$$\mathrm{card}[R(\pi)]=\sum\{\tbinom{a_i}{2}/1\leq i\leq I\},\quad \mathrm{card}[R(\chi)]=\sum\{\tbinom{b_j}{2}/1\leq j\leq J\}$$
$$s=\mathrm{card}[R(\pi)\cap R(\chi)]=\sum\{\tbinom{c_{ij}}{2}/1\leq i\leq I,\ 1\leq j\leq J\},$$
$$u=\mathrm{card}[R(\pi)\cap S(x)]=\sum\{c_{ij}c_{ij'}/1\leq i\leq I,\ 1\leq j<j'\leq J\}, \qquad (4)$$
$$v=\mathrm{card}[S(\pi)\cap R(\chi)]=\sum\{c_{ij}c_{i'j}/1\leq i<i'\leq I,\ 1\leq j\leq J\},$$
$$t=\mathrm{card}[S(\pi)\cap S(\chi)]=\sum\{c_{ij}c_{i'j'}/1\leq i<i'\leq I,\ 1\leq j<j'\leq J\},$$

où nous notons $S(\pi)$ pour $R^c(\pi)$ et $S(\chi)$ pour $R^c(\chi)$.

Tous les indices ainsi obtenus -par transposition de coefficients connus de similarité entre parties d'un même ensemble fini- ne tiennent pas complètement compte de la nature particulière des structures de partition à comparer. On peut d'ailleurs considérer ces indices pour comparer deux relations binaires quelconques, dès lors qu'on les aura représentés par leurs graphes respectifs au niveau de OxO. Toutefois, l'ensemble de représentation n'est pas abstrait ; il s'agit de l'ensemble des paires. Les sous ensembles $R(\pi)$ et $R(\chi)$ sont fermés transitivement et la contrainte de TRANSITIVITE n'intervient pas dans la comparaison. F. Marcotorchino [Marcotorchino (1984)] a montré que beaucoup de coefficients d'association entre partitions qui sont proposés dans la littérature, correspondent à la comparaison de deux codages 'linéaires' de l'ensemble OxO des couples. Ils ne sont donc pas spécifiques à la comparaison de deux partitions. Ils peuvent être formellement considérés pour comparer deux relations binaires quelconques, ou même deux parties d'un ensemble abstrait fini.

II. SCHEMA GENERAL DE COMPARAISON ENTRE VARIABLES QUALITATIVES RELATIONNELLES

Pour comparer deux variables qualitatives relationnelles, nous avons au cours de notre recherche fait émerger le diagramme suivant :

$$(\alpha,\beta)\in A\times B \longrightarrow [R(\alpha),R(\beta)]\in \Omega_\alpha \times \Omega_\beta$$
$$\longrightarrow s=s(\alpha,\beta)=\mathrm{card}[R(\alpha)\cap R(\beta)]$$

- Hypothèse d'absence de lien (h.a.l.) (ou d'indépendance) tenant en compte de façon stricte ou 'floue' les caractéristiques de cardinalité de α et de β.

$$\longrightarrow S=s(\alpha^*,\beta^*)=\mathrm{card}[R(\alpha^*)\cap R(\beta^*)]$$
$$\longrightarrow Q(\alpha,\beta)=[s-\mathcal{E}(S)]/\sqrt{[\mathrm{var}(S)]}$$

Dans ce schéma α et β sont les deux relations sur l'ensemble des objets, respectivement déterminés par les deux variables à comparer. A(resp.B) est l'ensemble de toutes les relations du "même type" que α(resp.β). $R(\alpha)$ [resp.$R(\beta)$] est la représentation ensembliste de α(resp.β). $R(\alpha)$ [resp.$R(\beta)$] est un sous ensemble de O, ou bien de OxO, ou même de (OxO)x(OxO). Ω_α(resp.Ω_β) est l'ensemble de tous les sous ensembles possibles de représentation d'une relation de même type que α(resp.β). $s=s(\alpha,\beta)$ est appelé indice "brut". α^* et β^* sont deux variables aléatoires relationnelles indépendantes, respectivement associées à α et β, conformément à l'hypothèse d'absence de liaison, qui tient compte -de façon stricte ou floue- des caractéristiques cardinales de α et de β. S est l'"indice brut aléatoire" dont l'espérance mathématique et la variance sont notées $\mathcal{E}(S)$ et var(S). $Q(\alpha,\beta)$ est l'indice "centré et réduit".

Nous avons extensivement utilisé ce schéma dans l'élaboration de nos coefficients d'association totale ou partielle entre variables qualitatives [Lerman(1973),(1981),(1983a)(1983b)]. Pour le rendre plus clair, nous allons l'illustrer dans le cas qui nous concerne ici de la comparaison de deux partitions, définies en l'occurence par deux variables qualitatives nominales.

α et β sont deux partitions dont on supposera -sans restreindre la généralité- qu'elles sont en classes étiquetées. Nous les désignerons -conformément à ci-dessus- par π et χ. $t(\pi)$ [resp.$t(\chi)$] indiquera le type de la partition π (resp.χ) ; c'est-à-dire, la suite ordonnée des cardinaux de ses classes : $t(\pi)=(a_i/1\leq i\leq I)$ [resp.$t(\chi)=(b_j/1\leq j\leq J)$]. Dans ces conditions A(resp.B) est l'ensemble des partitions -en classes étiquetées sur O, de type $t(\pi)$ [resp.$t(\chi)$]. $R(\pi)$ [resp.$R(\chi)$] est l'ensemble des paires d'objets dont les deux composantes sont réunies dans une même classe de la partition π (resp.χ) (cf.ci-dessus). Ω_π (resp.Ω_χ) peut être défini comme étant l'ensemble des parties de $O^{\{2\}}$ dont chacune correspond à la représentation d'une partition de type $t(\pi)$ |resp. $t(\chi)$]. Nous avons déjà exprimé [cf.(4) ci-dessus], $s=\mathrm{card}[R(\pi)\cap R(\chi)]$.

Il y a trois formes fondamentales de l'h.a.l. [Lerman(1981) chap.2]. Nous allons considérer ici celle stricte où π^* (resp.χ^*) est une partition aléatoire dans l'ensemble A(resp.B) muni d'une probabilité uniformément répartie. La moyenne et la variance de l'indice brut aléatoire $S=s(\pi^*,\chi^*)$ sont respectivement donnés par [Lerman(1973),(1981)]:

$$\mathcal{E}(S)=\lambda\mu \text{ et } \mathrm{var}(S)=\lambda\mu+\rho\sigma+\theta\zeta-\lambda^2\mu^2,$$

où

$$\lambda=\sum\{a_i(a_i-1)/\sqrt{[2n(n-1)]}/1\leq i\leq I\} \qquad (1)$$

$$\rho=\sum\{a_i(a_i-1)(a_i-2)/\sqrt{[n(n-1)(n-2)]}\ 1\leq i\leq I\}$$

$$\theta=\{[\sum_i a_i(a_i-1)]^2-2\sum_i a_i(a_i-1)(2a_i-3)\}/2\sqrt{[n(n-1)(n-2)(n-3)]}$$

et où les expressions de μ,σ et ζ ont respectivement la même forme que λ, ρ et θ ; les a_i de $t(\pi)$ étant remplacés par les b_j de $t(\chi)$.

On remarquera que θ peut s'exprimer en fonction de λ et de ρ puisque

$$\sum_i a_i(a_i-1)(2a_i-3)=2\sum_i a_i(a_i-1)(a_i-2)+\sum_i a_i(a_i-1).$$

Nous allons à présent situer deux coefficients classiques par rapport au schéma que nous avons présenté. Le premier -de comparaison de deux attributs logiques a et b- est celui de K. Pearson [Pearson(1928)] et le second -de comparaison entre deux variables 'rang' r et s- est celui de M.G. Kendall [Kendall(1970)].

Pour obtenir le coefficient de K. Pearson, on représente un même attribut logique a(resp.b) par le sous ensemble O(a) [resp.O(b)] de l'ensemble O des objets, formé de ceux qui possèdent a(resp.b). Par conséquent R(a)=O(a) [resp.R(b)=O(b)]. L'h.a.l. est stricte et associe à R(a) [resp.R(b)], une partie aléatoire $R(a^*)$ [resp.(b^*)] dans l'ensemble -muni d'une probabilité uniforme- des parties de O de même cardinal n(a) [resp.n(b)]. La v.a. S est hypergéométrique et Q(a,b) se met sous la forme :

$$Q(a,b)=\sqrt{(n-1)\rho(a,b)}\simeq\sqrt{n\ \rho(a,b)}. \qquad (2)$$

où $\rho(a,b)$ est un coefficient pur, compris entre -1 et +1, dont la limite -pour n tendant vers l'infini et [n(a)/n] (resp.[n(b)/n]) vers une limite finie- est indépendante de n. Dans ces conditions, le coefficient $\rho(a,b)$ peut s'obtenir par l'une ou l'autre des deux expressions suivantes :

$$\text{(i) } \rho(a,b)= \frac{1}{\sqrt{n}}\, Q(a,b), \qquad (3)$$

$$\text{(ii) } \rho(a,b)= \frac{Q(a,b)}{\sqrt{Q(a,a)Q(b,b)}} . \qquad (4)$$

Considérons à présent la manière d'obtenir -dans le cadre du schéma- le coefficent τ de M.G. Kendall de comparaison de deux variables 'rang' r et s. Chacune des variables définit un ordre total et strict sur l'ensemble O des objets. R(r) [resp. R(s)] est le graphe dans OxO de la relation d'ordre total définie par r (resp.s) qu'on notera également r(resp.s) r^* (resp.s^*) est un ordre aléatoire dans l'ensemble -muni d'une probabilité uniforme- des n! ordres totaux et stricts sur O. Dans ces conditions, l'indice τ de M.G. Kendall se met sous la forme

$$\tau\,(r,s)= \frac{\{s(r,s)-\mathcal{E}[s(r^*,s^*)]\}}{\{\max[s(r',s')]-\mathcal{E}[s(r^*,s^*)]\}} , \qquad (5)$$

où $\max[s(r',s')]$ est le maximum possible de l'indice brut de comparaison de deux ordres totaux r' et s', il s'agit ici de n(n-1)/2.

Certains chercheurs tels que L. Hubert et Ph. Arabie [Hubert & Arabie (1985)] considèrent qu'un coefficient d'association entre deux variables qualitatives doit nécessairement avoir la même forme que (5), à savoir et dans le cas général

$$\tau\,(\alpha,\beta)= \frac{\{s(\alpha,\beta)-\mathcal{E}[s(\alpha^*,\beta^*)]\}}{\{\max[s(\alpha',\beta')]-\mathcal{E}[s(\alpha^*,\beta^*)]\}} , \qquad (6)$$

où $\max[s(\alpha',\beta')]$ est le maximum possible de card $[R(\alpha')\cap R(\beta')]$ pour deux structures, α' de même type que α et β' de même type que β.

D'autre part, ces chercheurs considèrent qu'une statistique telle que $Q(\alpha,\beta)$ doit être consacrée à tester l'hypothèse d'indépendance entre α et β . Mais nous avons montré [Lerman(1984)] la non pertinence des tests statistiques d'indépendance entre variables descriptives en analyse des données. Rien n'interdit donc de considérer un coefficient $\rho(\alpha,\beta)$ directement déduit de $Q(\alpha,\beta)$, de la même façon que le coefficient $\rho(a,b)$ de K. Pearson peut être déduit de $Q(a,b)$ [cf. expressions (3) et (4)].

Notre objectif dans la suite de ce papier est de présenter un aperçu sur les plus récents résultats que nous venons d'obtenir sur la possibilité et l'analyse de formules telles que (3),(4) ou surtout (5) en cas de comparaison de deux partitions π et χ. Contrairement aux indices présentés au paragraphe I, la normalisation de nos coefficients tiendra intimement compte des contraintes de structure dans la comparaison de deux partitions.

III. FORME LIMITE DE $Q(\pi,\chi)$. NORMALISATION PAR L'ECART TYPE

Nous allons en réalité donner la solution d'un problème sensiblement plus général pour la comparaison de deux 'codages' symétriques (resp. antisymétriques) de OxO. Si $\{\phi(x,y)/(x,y)\in X\mathrm{x}X\}$ et $\{\psi(x,y)/(x,y)\in X\mathrm{x}X\}$ -où $X=\{1,2,\ldots,x,\ldots,n\}$ indexe O- désignent les deux codages, on suppose avoir

$$(\forall(x,y)\in X\mathrm{x}X)\ [\phi(x,y)=\phi(y,x) \text{ et } \psi(x,y)=\psi(y,x)]$$
$$\text{ou } (\forall(x,y)\in X\mathrm{x}X)\ [\phi(x,y)=-\phi(y,x) \text{ et } \psi(x,y)=-\psi(y,x)],$$

d'autre part,

$(\forall x\in X)$ $[\phi(x,x)=\psi(x,x)=1]$ en cas de codage symétrique,

ou

$(\forall x \in X)\ [\phi(x,x)=\psi(x,x)=0]$ en cas de codage antisymétrique.

L'indice brut ignore la diagonale de XxX et se met sous la forme

$$s(\phi,\psi)=\sum\{\phi(x,y)\psi(x,y)/(x,y)\in X^{[2]}\} \qquad (1)$$

L'indice brut aléatoire peut prendre la forme suivante

$$S=s(\psi^*,\psi^*)=\sum\{\phi[\sigma(x),\sigma(y)]\psi[\tau(x),\tau(y)]/(x,y)\in X^{[2]}\}, \qquad (2)$$

où σ et τ sont deux permutations aléatoires indépendantes prises dans l'ensemble G_n -muni d'une probabilité uniforme- des n! permutations sur X.

La comparaison de deux partitions π et $\varkappa$ correspond à celle de deux codages symétriques à valeurs 0 ou 1 et valant 1 sur la diagonale de XxX.

On a

$$\mathcal{E}(S)=\frac{1}{n^{[2]}}\Big[\sum_{[x,y]}\phi(x,y)\Big]\Big[\sum_{[x,y]}\psi(x,y)\Big], \qquad (3)$$

où $n^{[2]}=n(n-1)$ et où $[x,y]$ est un élément courant de $X^{[2]}$.

Nous allons maintenant donner de la variance de S deux expressions ; la première est dûe à Mantel [Mantel(1967)] et la seconde mise en évidence par nous-mêmes [Lerman(1976)], tout à fait indépendamment -car nous ignorions le papier de Mantel- mais à la suite d'une tentative de G. Lecalvé [Lecalvé(1976)] qui s'inspirait d'un vieux papier de H.E. Daniels [Daniels(1944)]. Ces expressions seront données dans le contexte particulier de la comparaison de deux codages symétriques (resp. antisymétriques).

Pour présenter l'expression de la variance selon Mantel introduisons les paramètres classiques suivants :

$$A_1=\Big[\sum_{X^{[2]}}\phi(x,y)\Big]^2,\ A_2=\sum_{x\in X}\Big[\sum_{y\in X-\{x\}}\phi(x,y)\Big]^2 \text{ et } A_3=\sum_{X^{[2]}}[\phi(x,y)]^2,$$

B_1,B_2 et B_3 qui ont respectivement les mêmes expressions que A_1,A_2 et A_3, au remplacement près de ϕ par ψ. En notant par $n^{[r]}=n(n-1)\ldots(n-r+1)$, la r-ième puissance factorielle de n, on a :

$$\begin{aligned}\mathrm{var}(S)=&\frac{2}{n^{[2]}}A_3B_3+\frac{4}{n^{[3]}}(A_2-A_3)(B_2-B_3)\\&+\frac{1}{n^{[4]}}(A_1-4A_2+2A_3)(B_1-4B_2+2B_3)-\frac{1}{(n^{[2]})}A_1B_1. \qquad (4)\end{aligned}$$

Notre expression de la variance dans la situation concernée est la suivante :

$$\begin{aligned}\mathrm{var}(S)=&\frac{2}{n^{[2]}}A_3B_3+\frac{4}{n^{[3]}}\Big[\sum_G\phi(x,y)\phi(x,z)\Big]\Big[\sum_G\psi(x,y)\psi(x,z)\Big]\\&+\frac{1}{n^{[4]}}\Big[\sum_H\phi(x,y)\phi(z,t)\Big]\Big[\sum_H\psi(x,y)\psi(z,t)\Big]\\&-\frac{1}{(n^{[2]})}A_1B_1, \qquad (5)\end{aligned}$$

où G(resp.H) est l'ensemble des tri-uples $[x,y,z]$ (resp. quadruplets $[x,y,z,t]$) à composantes mutuellement distinctes.

La correspondance entre notre expression (5) et celle (4) de Mantel est alors claire à partir de l'identification qu'indique les coefficients $\frac{2}{n^{[2]}}$, $\frac{4}{n^{[3]}}$, $\frac{1}{n^{[4]}}$ et $\frac{1}{(n^{[2]})^2}$.

On se rend en particulier compte que l'expression de la variance dépend d'une part de paramètres A_1 et A_3 (resp.B_1 et B_3) qui -pour la comparaison de deux relations symétriques (resp. antisymétriques)- s'expriment au niveau de l'ensemble $X^{\{2\}}$ des paires et d'autre part, du terme irréductible $\sum\{\phi(x,y)\phi(x,z)/[x,y,z]\in G\}$ (resp. $\sum\{\psi(x,y)\psi(x,z)/[x,y,z]\in G\}$). Dans ces conditions on peut appeler cette quantité -qui fait intervenir la transitivité- la "caractéristique particulière de la structure" ϕ (resp. ψ).

Introduisons les moments factoriels suivants :

$$\lambda_1 = \frac{1}{n^{[2]}}\sum\{\varphi(x,y)/(x,y)\in X^{[2]}\},$$

$$\lambda_2 = \frac{1}{n^{[2]}}\sum\{\varphi^2(x,y)/(x,y)\in X^{[2]}\}, \qquad (6)$$

$$\rho = \frac{1}{n^{[3]}}\sum\{\varphi(x,y)\varphi(x,z)/[x,y,z]\in G\},$$

d'autre part, respectivement, μ_1,μ_2 et σ qui sont à ψ, ce que λ_1,λ_2 et ρ sont à ϕ.

Nous établissons [Lerman(1987c)] le résultat suivant de calcul :

Théorème 1. $$\mathrm{var}(S) = \frac{2\,n^{[2]}}{(n-3)}\{2n^{[2]}(\rho-\lambda_1^2)(\sigma-\mu_1^2) + \frac{n^{[2]}}{(n-2)}(\lambda_2-\lambda_1^2)(\mu_2-\mu_1^2) - 4(\rho-\lambda_2)(\sigma-\mu_2)\}. \qquad (7)$$

Introduisons à présent les moments absolus suivants :

$$p_1 = \frac{1}{n^2}\sum\{\phi(x,y)/(x,y)\in X\times X\},$$

$$p_2 = \frac{1}{n^2}\sum\{\phi^2(x,y)/(x,y)\in X\times X\}, \qquad (8)$$

$$t = \frac{1}{n^3}\sum\{\phi(x,y)\phi(x,z)/(x,y,z)\in X\times X\times X\},$$

d'autre part, respectivement, q_1,q_2 et u qui sont à ψ, ce que p_1,p_2 et t sont à ϕ.

Dans la référence ci-dessus mentionnée, nous établissons le résultat suivant :

Lemme. $$\mathrm{var}(S) = \frac{4\,n^6}{(n-1)(n-2)^2}(t-p_1^2)(u-q_1^2) + \frac{2n(n-1)^2}{(n-3)}\{[(\lambda_2-\lambda_1^2) - \frac{2\,n^2}{(n-1)(n-2)}(t-p_1^2)] \times [(\mu_2-\mu_1^2) - \frac{2\,n^2}{(n-1)(n-2)}(u-q_1^2)]\}. \qquad (9)$$

En notant

$$w = \frac{1}{n^2} \sum\{\phi(x,y)\psi(x,y)/(x,y)\in X\times X\}, \quad (10)$$

on a le résultat :

Théorème 2. La forme limite de l'indice centré et réduit $Q(\phi,\psi)$ est :

$$Q(\varphi,\psi) - \frac{\frac{\sqrt{n}}{2}(w-p_1q_1)}{\sqrt{(t-p_1^2)(u-q_1^2)+\frac{1}{2n}\left[(p_2-p_1^2)-2(t-p_1^2)\right]\left[(q_2-q_1^2)-2(u-q_1^2)\right]}} \quad (11)$$

En général $(t-p_1^2)(u-q_1^2)$ est différent de zéro et positif. Dans ce cas, l'expression de $Q(\phi,\psi)$ devient pour n "assez grand" :

$$Q(\phi,\psi) - \frac{\frac{n}{\sqrt{2}}(w-p_1-q_1)}{\sqrt{(t-p_1^2)(u-q_1^2)}}. \quad (12)$$

On remarque avec beaucoup d'intérêt que -comme dans le cas de l'obtention du coefficient de K. Pearson- $Q(\phi,\psi)$ est au facteur $\sqrt{n}$ près, un coefficient pur dont la limite ne dépend pas de n, en supposant que w, p_1, p_2 et t (resp.q_1, q_2 et u) tendent vers des limites finies.

Donnons à présent la forme limite de l'expression de $Q(\pi,\chi)$, en cas de comparaison de deux relations partition, dans ce cas ϕ et ψ sont symétriques et à valeurs 0 ou 1 ; $\phi(x,y)=1$ $[\text{resp.}\psi(x,y)=1]$ si x et y sont réunis par la partition π(resp.χ) et $\phi(x,y)=0$ $|\text{resp.}\psi(x,y)=0|$ si x et y sont séparés par la partition π(resp.χ).

En se rappelant des notations du début du paragraphe II, posons :

$\pi_i = \frac{a_i}{n}$, $\chi_j = \frac{b_j}{n}$ et $\gamma_{ij} = \frac{c_{ij}}{n}$ pour tout (i,j), $1\leq i\leq I$, $1\leq j\leq J$.

On a alors

$$p_1=p_2=p=\sum_i \pi_i^2,\quad t=\sum_i \pi_i^3,$$

$$q_1=q_2=q=\sum_j \chi_j^2,\quad u=\sum_j \chi_j^3 \quad (13)$$

$$\text{et } w=\sum_{(i,j)} \gamma_{ij}^2.$$

Corollaire. La forme limite de l'indice centré réduit $Q(\pi,x)$ est :

$$Q(\pi,x) - \frac{\frac{\sqrt{n}}{2}(w-pq)}{\sqrt{(t-p^2)(u-q^2)+\frac{1}{2n}\left[(p-p^2)-2(t-p^2)\right]\left[(q-q^2)-2(u-q^2)\right]}}. \quad (14)$$

$(t-p^2)$ $[\text{resp.}(u-q^2)]$ est en général strictement positif et exceptionnellement nul ; la nullité ayant lieu si tous les π_i (resp.x_i) sont égaux entre eux ; c'est-à-dire à (1/I) [resp.(1/J)]. Dans ce cas, l'ordre de grandeur du dénominateur change et on a l'impression d'une rupture dans le comportement de l'indice $Q(\pi,\chi)$. Mais, il ne faut pas oublier que l'ordre de grandeur du numérateur change également ; l'indice brut chutant brutalement lorsque l'une des partitions est en classes de même cardinal, par rapport à la situation où chacune des deux partitions est en classes de cardinaux respectifs très différents.

Revenons au cas plus général de $Q(\phi,\psi)$. Nous avons émis la possibilité de définir un coefficient de la forme :

$$R(\phi,\psi)=\frac{Q(\phi,\psi)}{\sqrt{Q(\phi,\phi)Q(\psi,\psi)}} \qquad (15)$$

Si n n'est pas assez grand pour que le second terme sous le signe radical ($\sqrt{\ }$) de (14) garde une certaine influence, on a alors un indice tout à fait nouveau. Sinon, l'indice qu'on obtient correspond à un coefficient de corrélation entre les deux pondérations ϕ et ψ, équivalent formellement à celui de K. Pearson.

IV. NORMALISATION PAR LE MAXIMUM

L'objectif est la définition d'un coefficient conforme à la philosophie de $\tau(\alpha,\beta)$ [cf. expression (6) §I]. Relativement à la comparaison de deux variables relationnelles ϕ et ψ, correspondantes à deux codages de $O^{[2]}$, le problème de la maximisation de $\sum\{\phi(x,y)\psi(x,y)/(x,y)\in X^{[2]}\}$ sur l'ensemble $\{\sum\{\phi[\sigma(x),\sigma(y)]\psi(x,y)/(x,y)\in X^{[2]}\}/\sigma\in G_n\}$ -où G_n est l'ensemble des n! permutations sur X- est reconnu comme très difficile, quels que soient les deux types de structure à comparer [Hubert (1983), Hubert & Arabie(1985), F. Marcotorchino(1984)]. Nous avons pu proposer une solution exacte au problème posé et cela dans deux situations. La première -qui nous concerne ici- est celle de la comparaison de deux partitions π et χ [Lerman & Peter (1986)]. La deuxième est celle de la comparaison de deux préordres totaux [Lerman (1987a)]. C'est la première des deux situations qui a été la plus difficile à réduire. Nous allons donner un bref aperçu de la solution, le lecteur intéressé se reportera au rapport détaillé que nous venons de mentionner.

Conformément aux notations introduites (cf.§I), il s'agit de résoudre, en nombres entiers :

$$\mathrm{Max}\sum\{c_{ij}^2/1\leq i\leq I,\ 1\leq j\leq J\} \qquad (1)$$

sous les contraintes

$$\left.\begin{array}{l} \displaystyle\sum_{1\leq j\leq J} c_{ij}=a_i \text{ pour tout } i,\ 1\leq i\leq I, \\ \displaystyle\sum_{1\leq i\leq I} c_{ij}=b_j \text{ pour tout } j,\ 1\leq j\leq J. \end{array}\right. \qquad (2)$$

La première idée très importante est de remplacer la notion de 'formule mathématique' jusque là utilisée par celle beaucoup plus générale d''algorithme récursif'. La deuxième idée liée d'ailleurs à la première, consiste à travailler au niveau du tableau de contingence (à I lignes et J colonnes) dont on commencera par remplir les marges qu'il s'agit de répartir au mieux à l'intérieur de la table. De la sorte on aura en plus défini une configuration optimale de la table de contingence.

Sans risque d'ambiguité, notons π (resp.x) la fonction indicatrice de $R(\pi)$ [resp. $R(\chi)$] dans $P=O^{\{2\}}$ (cf. notations du paragraphe I). Nous démontrons (cf. référence mentionnée ci-dessus) que parmi les bornes classiques majorant (1) au moyen d'une expression mathématique symétrique en π et χ, la meilleure est fournie par une application de l'inégalité de Shwartz conçue dans un cadre logique :

$$\sum_{p\in P} \left[\pi(p)-\mu\right]\left[\chi(p)-\nu\right] \leqslant \sqrt{\{(\sum_{p\in P} \left[\pi(p)-\mu\right]^2)(\sum_{p\in P} \left[x(p)-\nu\right]^2)\}} \qquad (3)$$

où

$$\mu = \sum_i a_i(a_i-1)/n(n-1) \text{ et } \nu = \sum_j b_j(b_j-1)/n(n-1). \qquad (4)$$

Toutefois, comme nous le démontrons, cette meilleure borne 'analytique' qui en résulte pour $\sum\{c_{ij}/1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}$ reste trop grande devant celle -dissymétrique- définie par $\min(\sum_i a_i^2, \sum_j b_j^2)$. Cette dernière devient la borne exacte dès lors que l'un des partages $(a_1,\ldots,a_I)$ ou $(b_1,\ldots,b_J)$ est plus fin que l'autre.

La solution récursive repose sur le fait que face à une configuration optimale du tableau, la suppression d'une ligne (resp. colonne) et le réaménagement en conséquence des marges, conduit également à une configuration optimale.

Relativement au tableau vide à l'intérieur mais ayant ses marges remplies, on sera conduit -à chaque pas- à installer le contenu d'une marge ligne que nous notons α_i dans une colonne j de marge $\beta_j \geqslant \alpha$ ou bien le contenu d'une marge colonne β_j dans une ligne i de marge $\alpha_i \leqslant \beta_j$. On dira qu'on résoud le couple (i,j). Une telle résolution -que suppose la configuration optimale du tableau- diminue la dimension du problème en diminuant d'une unité, voire même de deux (si les deux arguments du couple résolu sont identiques) le cardinal défini par (nombre de lignes + nombre de colonnes).

C'est assez rapidement que nous montrons que le plus grand entier $c_{i_0j_0}$ de la configuration optimale T_0 du tableau T, correspond nécessairement à la résolution du couple (i_0,j_0). Nous démontrons d'autre part que si une même part marginale se retrouve en ligne et en colonne : $a_{i_1}=b_{j_1}$, la configuration optimale T_0 comprend nécessairement la résolution de (i_1,j_1).

Une analyse expérimentale poussée nous conduit à définir sur l'ensemble des couples $\{(a_i,b_j)/1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}$ une relation de préordre partiel résultant de l'intersection de deux préordres totaux ω_d et ω_s, où le premier est conforme à la différence $|a_i-b_j|$ décroissante et où le second est conforme à la somme (a_i+b_j) croissante. Plus précisément,

$$\{\forall|(i,j),(i',j')|\},\ (a_{i'},b_{j'}) \leqslant (a_i,b_j)(\text{pour } \omega_d) \Leftrightarrow |a_i-b_j| \leqslant |a_{i'}-b_{j'}| \qquad (5)$$

et

$$\{\forall|(i,j),(i',j')|\},\ (a_{i'},b_{j'}) \leqslant (a_i,b_j)(\text{pour } \omega_s) \Leftrightarrow (a_i+b_j) \geqslant (a_{i'}+b_{j'}). \qquad (6)$$

L'algorithme que nous proposons repose sur la simple propriété suivante : "Une configuration optimale du tableau T peut être obtenue en commençant par la résolution d'un couple (a_i,b_j) extrémal par rapport à $\omega=\omega_d \cap \omega_s$".

Le résultat majeur que nous avons pu établir [Lerman & Peter(1986)] est la démonstration complète de cette propriété en cas où il y a un seul couple extrémal. D'autre part, dans différentes situations mathématiques où il existe plus d'un seul couple extrémal, nous avons pu nous rendre compte que la solution optimale passe par la résolution d'un couple extrémal. Un contre exemple à cette propriété serait tout à fait invraisemblable ; il exprimerait qu'aucun des couples résolus au niveau d'une configuration optimale ne correspond à un couple extrémal alors que d'une part, la résolution d'un couple extrémal correspond à "bien vider" une marge pour remplir au mieux l'intérieur du tableau et qu'on sait que le plus grand entier correspond à la résolution d'un couple, que celui qui suit correspond à la résolution d'un couple du tableau résultant de la supression de la marge supportant ce plus grand entier et réaménagement de l'autre marge et ainsi de suite...

La procédure de recherche récursive -qui utilise le 'back-tracking'- suppose la détermination -après chaque résolution- de l'ensemble des couples extrémaux. La propriété importante suivante limite considérablement l'empilement : Si (a_{i_o},b_{j_o}) est le couple extrémal qui réalise la plus petite valeur de $|a_i-b_j|$, le déchargement de a_i dans j (si $a_i \leq b_j$) [resp. de b_j dans i (si $a_i > b_j$)] préserve le caractère extrémal de tout autre couple (a_{i_1},b_{j_1}) avec $i_1 \neq i_o$ et $j_1 \neq j_o$.

Ainsi pour obtenir le coefficient $\tau(\pi,\chi)$ qui serait défini par une formule analogue à celle (6) du paragraphe II, on utilisera non une formule 'mathématique', mais l'algorithme dont nous venons de donner un rapide aperçu.

V. ASPECTS STATISTIQUES

Nous avons bien souligné la non pertinence des tests d'indépendance statistique entre variables qualitatives en analyse des données. Ce qui est plus en jeu est l'organisation mutuelle d'une famille de variables selon leurs liaisons respectives. Notre méthode de classification hiérarchique basée sur la vraisemblance du lien maximal [Lerman(1970a)(1981)] fournit une solution à ce problème en profitant et en développant les acquis de la statistique non paramétrique. Pour rester dans le cadre de ce papier, considérons que la donnée est une famille $\{\pi_l / 1 \leq l \leq L\}$ de partitions et supposons établie la matrice des indices d'association $\{Q(\pi_l,\pi_m)/1 \leq l < m \leq L\}$ ou bien celle $\{R(\pi_l,\pi_m)/1 \leq l < m \leq L\}$ Désignons par $\{S(\pi_l,\pi_m)/1 \leq l < m \leq L\}$ l'une ou l'autre de ces deux tables. Le passage de la table de ces indices à celle $\{P(\pi_l,\pi_m)/1 \leq l < m \leq L\}$ qui se réfère à une échelle de probabilité -de vraisemblance du lien- se fait au moyen de la formule

$$P(\pi_l,\pi_m) = \Phi\left[\frac{S(\pi_l,\pi_m) - \mathrm{moy}_e(S)}{\sqrt{\mathrm{var}_e(S)}}\right], \quad 1 \leq l < m \leq L \qquad (1)$$

où Φ est la f.r. de la loi normale centrée et réduite N(0,1) et où $\mathrm{moy}_e(S)$ et $\mathrm{var}_e(S)$ sont la moyenne et la variance de la table des valeurs $\{S(\pi_l,\pi_m)/1 \leq l < m \leq L\}$.

Or la référence (1) à la loi normale -bien qu'elle soit algorithmiquement toujours possible- se justifie d'autant plus que la loi de $Q(\pi_l^*,\pi_m^*)$ a tendance à être normale. Toutefois, P.W. Mielke [Mielke(1979)] met en évidence une tendance asymptotique non normale dans le cas où l'une des partitions est en classes de même cardinal. Cependant, les partitions en classes de même effectif correspondent à une pure construction de l'homme et ne se trouvent pas naturellement dans les données.

REFERENCES

[1] Daniels, H.E. (1944). 'The relation between measures of correlation in the universe of sample permutations', Biometrika, vol. 33, 129-135.

[2] Fowlkes, E.B. and Mallows C.L. (1983). 'A method for comparing two hierarchical clusterings', J.A.S.A. 78 553-569.

[3] Goodman L.A. and Kruskal W.H. (1954), 'Measures of association for cross classifications', J.A.S.A. 49, 732-764.

[4] Hamann V. (1961), 'Merkmalbestand und Verwandtschaftsbeziehungen der Farinosae. Ein Beitragzum System der Monokotyledonen', Willdenowia, 2, 639-768.

[5] Hubert L.J. (1983), 'Inference procedures for the evaluation and comparison of proximity matrices' in 'Numerical Taxonomy', NATO ASI Series, Ed. J. Felsenstein, Springer Verlag.

[6] Hubert L.J. and Arabie Ph. (1985), 'Comparing partitions', Journal of Classification, 2,2-3, 193-218.

[7] Jaccard P. (1908), 'Nouvelles recherches sur la distribution florale', Bull. Soc. Vaud. Sci Nat., t.44, 223-270.

[8] Kendall M.G. (1970), 'Rank correlation methods', Charles Griffin, fourth edition (first edition in 1948).

[9] Lecalve G. (1976), 'Un indice de similarité pour des variables de types quelconques', Stat. et Anal. des Données, 01-02, 39-47.

[10] Lerman I.C. (1970a), 'Sur l'analyse des données préalable à une classification automatique. Proposition d'une nouvelle mesure de similarité', Rev. Math. & Sc. Hum. 8è année n°32.

[11] Lerman I.C. (1970b), 'Les bases de la classification automatique', Gauthier-Villars, Paris.

[12] Lerman I.C. (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.

[13] Lerman I.C. (1976), 'Formal analysis of a general notion of proximity between variables', Congrès Européen des Statisticiens, Grenoble, published by North Holland in 1977.

[14] Lerman I.C. (1981), 'Classification et analyse ordinale des données', Dunod, Paris.

[15] Lerman I.C. (1983a), 'Indices d'association partielle entre variables qualitatives nominales', R.A.I.R.O., série R.O., vol.17, n°3, 213-259.

[16] Lerman I.C. (1983b), 'Indices d'association partielle entre variables qualitatives ordinales', Publ. I.S.U.P., XXVIII, fasc 1,2, 7-46.

[17] Lerman I.C. (1984), 'Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées', Publ. ISUP, XXIX, fasc. 3-4, 27-57.

[18] Lerman I.C. (1987a), 'Maximisation de l'association entre deux variables qualitatives ordinales', Publ. Int. n°341, IRISA, Rennes.

[19] Lerman I.C. (1987b), 'Comparing relational variables according to Likelihood of the links classification method', in Recent Developments in Clustering and Data Analysis, Japanese-French Scientific Seminar, 24-26 March 1987, Tokyo, to be published by Academic Press.

[20] Lerman I.C. (1987c), 'Analyse formelle de coefficients statistiques d'association entre variables relationnelles', Publ. Int. à paraître, IRISA, Rennes.

[21] Lerman I.C. et Peter Ph. (1986), 'Structure maximale pour la somme des carrés d'une contingence aux marges fixées. Une solution algorithmique programmée', Publ. Int. n°318, IRISA, Oct. 90 pages, Rennes.

[22] Mantel N. (1967), 'The detection of disease clustering and a generalized regression 'approach', Cancer Research, vol.27, n°2, 209-220.

[23] Marcotorchino F. (1984), 'Utilisation des comparaisons par paires en statistique des contingences (Partie I)', Etude F-069, Centre Scientifique IBM-France.

[24] Mielke P.W. (1979), 'On asymptotic non normality of null distributions of MRPP Statistics', Communications in Statistics, Thoery and Methods, A8(15), 1541-1550.

[25] Ochiai A. (1957), 'Zoogeographic studies on the soleoid fishes found in Japan and its neighbowring regions', Bull. Jap. Soc. Sci. Fish, T22, 526-530.

[26] Pearson K. (1926), 'On the coefficient of racial likeness', Biometrika t.18, 105-117.

[27] Rand W.M. (1971), 'Objective criteria for the evaluation of clustering methods', J.A.S.A. 66, 846-850.

[28] Sokal R.R. and Michener C.D. (1958), 'A statistical method for evaluating systematic relationships', Univ Kansas Sci. Bull. 38, 1409-1438.

[29] Yule G.U. (1911), 'An introduction of the theory of statistics', Charles Griffin and Co. Ltd, London.

[30] Yule G.U. (1912), 'On the methods of measuring the association between two attributes', J. Roy. Statist. Soc. 75, 579-652.

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