REVUE DE STATISTIQUE APPLIQUÉE

I. C. LERMAN

Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification

Revue de statistique appliquée, tome 35, n° 2 (1987), p. 39-60

<http://www.numdam.org/item?id=RSA_1987__35_2_39_0>

Rev. Statistique Appliquée, 1987, **XXXV** (2), 39-60

CONSTRUCTION D'UN INDICE DE SIMILARITÉ ENTRE OBJETS DÉCRITS PAR DES VARIABLES D'UN TYPE QUELCONQUE. APPLICATION AU PROBLÈME DU CONSENSUS EN CLASSIFICATION (1)

I.C. LERMAN

IRISA, Campus de Beaulieu, 35042 Rennes Cedex

RÉSUMÉ

Le problème de la définition d'un indice d'association entre variables présente moins d'ambiguïté que celui d'un indice de similarité entre objets. Pour ce dernier problème, on présente une méthode très générale fine et féconde de normalisation — variable par variable — au niveau de l'ensemble des paires d'objets. Les variables descriptives sont supposées de types quelconques et on en distingue six : « numérique », « logique », « qualitatif nominal », « qualitatif ordinal », « préordonnance » et « graphe valué ». L'indice peut être pris en compte par l'Algorithme de la Vraisemblance du Lien, ce qui permet d'offrir une solution significative et efficace au problème du consensus en classification.

SUMMARY

The problem of the definition of an association coefficient between descriptive variables, is less ambiguous than this one of the definition of a similarity index between objects or individuals. For this last problem we developp a very general and fruitful method which is based upon standardization — variable by variable — at the level of the set of object pairs. We distinguich six types of descriptive variables : "numerical", "logical (0-1)", "nomimal qualitative", "ordinal qualitative", "preordonnance" and "weighted graph". The obtained similarity coefficient takes into account a mixing of the different types of descriptive variables. On the other hand, this coefficient is compatible with the Likelihood of the Link Algorithm. Then it becomes possible to propose an elegant and significant and efficiency solution to the general problem of consensus in classification.

Mots clés : *Classification, Normalisation de coefficients de ressemblance, Variables relationnelles « préordonnance ».*

I. Introduction

Le problème de la définition d'un indice de similarité entre éléments d'un ensemble E d'objets (n = card (E)) décrit par un ensemble V de variables (m = card (V)), est un problème difficile et délicat qu'on est loin de maîtriser. Le praticien sait bien que relativement à quelques indices dont il a l'usage courant,

(1) Nous sommes redevables à Ph. PETER (IRISA) d'avoir élaboré le logiciel correspondant à cette nouvelle famille d'indices (programme SIMOB) et d'avoir procédé à leur validation sur données réelles.

c'est tel indice qui a pu lui donner les « meilleurs résultats » dans certaines études, mais pas tel autre et que c'est l'inverse qui s'était produit dans d'autres études, sans qu'il sache vraiment pourquoi.

Notre but consiste ici à proposer une solution dans le cas le plus général où V est formé de variables de types quelconques. Cette solution repose sur une technique très générale, précise et féconde, de normalisation — variable par variable — au niveau relationnel de l'ensemble $P_2(E)$ des paires d'objets distincts de E, ou d'ailleurs, au niveau de $E \times E$.

Nous commencerons par supposer — ce qui est fréquemment le cas — que toutes les variables sont d'un même type. A cet égard, nous disposons d'une claire typologie des variables de description (cf. [LERMAN (1981)] Chap. 2) et nous distinguerons ici six types ou cas : « numérique », « logique (0-1) », « qualitatif nominal », « qualitatif ordinal », « préordonnance » et « graphe valué ». Ces différents cas de figure d'un tableau de données seront ci-dessous explicités.

Enfin, nous ne ferons que mentionner le cas que nous pouvons intégrer — mais traité par ailleurs [LERMAN-PETER (1985)] — d'une juxtaposition de tableaux de contingence.

Nous avons suffisamment insisté dans nos précédentes parutions (cf. par exemple la référence ci-dessus mentionnée) qu'en dehors de la table de contingence, un tableau de données Objets × Variables est de nature mathématique essentiellement dissymétrique : Variables et Objets ne jouent pas vis-à-vis les uns des autres le même rôle et il suffit pour s'en convaincre d'examiner le cas où les variables sont qualitatives.

Une bonne partie de notre recherche a porté sur la comparaison de variables d'un même type. D'une certaine façon — liée à ce que nous venons d'exprimer dans le dernier alinéa — ce dernier problème présente moins d'ambiguïté que celui qui nous occupe ici. En effet, dans l'évaluation de l'association entre deux variables, les différents objets de l'échantillon qui définit E ont exactement la même importance et ont a priori à intervenir de façon égale. Alors que dans notre problème — une fois normalisées les contributions des différentes variables — on ne sait pas s'il n'y a pas lieu d'accorder une plus forte pondération à certaines variables. Si on se conforme à l'opinion des promoteurs de la « Taxonomie Numérique » [SNEATH & SOKAL (1972)], c'est avec une égale importance que les variables doivent intervenir pour contribuer à la ressemblance entre deux objets. De toute façon, notre procédure de normalisation — variable par variable — permettra de façon très souple de prendre en compte une pondération des variables, posée a priori ou résultant d'une technique objective (e.g. analyse factorielle ou « importance projective »).

En prenant en compte un ensemble V de variables de description dont les types diffèrent, nous répondons au même problème que celui de J.C. GOWER [GOWER (1971)] dont le coefficient n'intègre finalement que deux types de variables : l'attribut logique et la variable quantitative. Notre propre codage des variables qualitatives permettra très naturellement une extension de la portée de ce dernier coefficient qui procède également en normalisant variable par variable. Mais, dans ce dernier type de normalisation, on rapporte la valeur de l'indice à celle maximale pouvant être atteinte. Ainsi, dans le cas où les variables sont toutes

des attributs logiques, le coefficient de Gower se réduit tout simplement à celui de Russel et Rao [RUSSEL et RAO (1940)]; c'est-à-dire, à la proportion des attributs présents chez les deux objets. Alors que la philosophie de notre indice se réfère à un concept de variance pour la normalisation. D'une certaine façon, notre indice sera une vaste généralisation de celui sous jacent à l'analyse en composantes principales normée.

Terminons cette introduction en signalant que ce texte reprend et enrichit sensiblement le paragraphe VI du chapitre 2 de [LERMAN (1981)]. En effet, on intègre les variables de types « préordonnance » et « graphe valué ». D'autre part, on remplace une normalisation globale par celle — plus fine — considérée ici, variable par variable.

II. Cas où les variables sont numériques

Désignons par $\{o_i / i \in I\}$ — où $I = \{1, 2, ..., i, ..., n\}$ — l'ensemble E des objets et par $\{v^j / j \in J\}$ — où $J = \{1, 2, ..., j, ..., m\}$ — l'ensemble V des variables qu'on suppose ici numériques et à valeurs positives. Cette restriction de positivité est en réalité tout à fait mineure puisque de façon quasi-générale, les variables numériques qui se présentent en analyse des données sont naturellement à valeurs positives.

L'appréhension de l'objet o_i se fait à partir de la suite des mesures $(x_i^1, x_i^2, ..., x_i^j, ..., x_i^m)$, où $x_i^j = v^j(o_i)$ est la mesure de la j-ème variable sur o_i.

Pour la comparaison de deux objets o_i et $o_{i'}$, pris dans E, on peut considérer deux indices de base :

$$\mathrm{COS}(o_i, o_{i'}) = \frac{\sum_j (x_i^j x_{i'}^j)}{\sqrt{\sum (x_i^j)^2 \sum_j (x_{i'}^j)^2}}, \tag{1}$$

et

$$L(o_i, o_{i'}) = \sum_j \frac{(x_i^j - x_\cdot^j)(x_{i'}^j - x_\cdot^j)}{s^2(j)}, \tag{2}$$

où

$$s^2(j) = \frac{1}{n} \sum_i (x_i^j - x_\cdot^j)^2$$

est la variance empirique de la variable v_j.

D'un point de vue formel chacun des deux indices est un produit scalaire sur des vecteurs transformés de $(x_i^j / 1 \leqslant j \leqslant m)$ et $(x_{i'}^j / 1 \leqslant j \leqslant m)$ par le choix d'une origine et d'une échelle.

Dans le premier cas où l'indice représente le cosinus de l'angle des deux vecteurs, l'origine est le point 0 et l'échelle est définie par une « normalisation » sur J. Cet indice possède la propriété intéressante d'invariance lorsqu'on remplace l'un ou l'autre des deux objets à associer par son « homothétique » où la suite des mesures des différentes variables se trouvent multipliées par le même coefficient. Cet indice met l'accent sur la ressemblance des « formes » indépendamment du facteur « taille ».

L'indice classique du coefficient de corrélation entre objets procède à un centrage sur J, avant la normalisation sur J :

$$\mathrm{COR}(o_i, o_{i'}) = \frac{\sum_j (x_i^j - x_i^o)(x_{i'}^j - x_{i'}^o)}{\sqrt{\sum_j (x_i^j - x_i^o)^2 \sum_j (x_{i'}^j - x_{i'}^o)^2}}, \qquad (3)$$

où x_i^o est la moyenne des mesures des différentes variables sur l'objet o_i.

Cet indice possède la même propriété d'invariance que celui (1). Néanmoins, l'interprétation de la moyenne x_i^o (resp. $x_{i'}^o$) n'est pas claire et ce d'autant plus que les échelles des valeurs prises par les différentes variables sont hétérogènes (du point de vue de leurs amplitudes respectives et de leurs variances respectives). D'ailleurs, à partir des expériences de classification hiérarchique, menées dans le cadre du critère de la « vraisemblance du lien maximal », sur un grand ensemble de jeux de données, c'est l'indice « cosinus » que nous retenons préférentiellement à celui de « corrélation » entre objets. En effet, ce dernier donne des résultats moins cohérents dans leur globalité.

Toutefois, dans le cas où il existe un grand degré d'hétérogénéité entre les variables (du point de vue de leurs amplitudes et variances respectives), on peut à première vue considérer l'indice (2), d'ailleurs sous-jacent à l'Analyse en Composantes Principales Normée et qui correspond au produit scalaire des vecteurs des mesures centrées réduites, variable par variable : $(\xi_i^1, ..., \xi_i^j, ..., \xi_i^m)$ et $(\xi_{i'}^1, ..., \xi_{i'}^j, ..., \xi_{i'}^m)$, où $\xi_i^j = (x_i^j - x_{\cdot}^j)/s_j$ (resp. $\xi_{i'}^j = (x_{i'}^j - x_{\cdot}^j)/s_j$). Ainsi, cet indice est obtenu après « centrage » et « normalisation » sur I.

C'est le cosinus de l'angle entre ces deux derniers vecteurs qui a permis à J.R. Massé [MASSÉ (1978)] d'obtenir les résultats les plus significatifs dans un problème de classification de composants électroniques d'origines diverses, sur lesquels ont été effectués différentes mesures électriques où d'une sous-classe de variables à une autre, l'ordre de grandeur de la variance pouvait être multiplié par 10, par 100 ou même par 1 000.

Toutefois, dans de nombreuses études où les rapports entre les variances des différentes variables étaient relativement appréciables, un indice tel que (2) a donné des résultats plus parcellaires que ceux (1) et (3). En effet, on peut se poser la question de savoir dans quelle mesure la variance empirique d'une variable descriptive doit-elle pondérer son importance pour l'évaluation de la ressemblance entre deux objets. D'autre part, pour ce dernier indice qui ne possède plus la propriété d'invariance par homothétie, la mesure d'une variable se trouve déconnectée de la mesure des autres variables sur le même objet.

Le type d'indice que nous allons proposer commence par relativiser la mesure d'une même variable par rapport à celles des autres sur un même objet et en cela, il possède (dans le cas numérique) la propriété d'invariance par homotétie. Désignons par η_{ij} cette mesure relative de la variable v^j sur l'objet o_i, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$.

Pour une même variable v^j, en ce qui concerne la comparaison de deux objets o_i et $o_{i'}$, on commence par admettre une forme multiplicative de l'indice

brut de ressemblance ($s_j(i, i') = \eta_{ij}\eta_{i'j}$). La contribution de la variable v^j pour l'évaluation de la ressemblance entre i et i', résulte alors du centrage et de la normalisation sur l'ensemble $P_2(I)$ des paires d'éléments de I, de $s_j(i, i')$.

Une telle réduction rend bien compte du caractère relatif de la ressemblance entre deux objets par rapport à l'ensemble des paires d'objets de l'ensemble où ils se situent et qui est à organiser. D'autre part, la relative hétérogénéité des différentes variables se trouve neutralisée lorsqu'on tient compte — de façon additive — de l'ensemble des variables pour l'évaluation de la ressemblance entre deux objets.

La réduction (i.e. centrage et normalisation) est proposée au niveau de $P_2(I)$, nous aurions pu également la proposer au niveau de l'ensemble des couples $I \times I$ avec l'intérêt de la comparaison des résultats. Mais nous ne voyions pas a priori en quoi la ressemblance brute d'un objet avec lui-même devrait intervenir dans le procédé de réduction. On remarquera qu'en adoptant pour η_{ij}, $(x_i^j - x_o^j)$ et en réduisant au niveau de $I \times I$, on retombe sur l'indice (2). Si on veut garder ce niveau de réduction pour situer notre indice par rapport à celui de J.R. MASSÉ, alors que ce dernier procède du schéma suivant :

Réduction sur I → Normalisation sur J,

le nôtre correspond à

Normalisation sur J → Réduction sur $I \times I$;

la réduction sur $I \times I$ ne se ramenant pas à celle sur I (la moyenne d'un produit n'est pas égale au produit des moyennes).

Dans [LERMAN (1981), chap. 2 §VI] nous avions bien proposé ce type de normalisation, mais pour un indice directement global tel que (1). Ce qu'il y a de différent ici, c'est que la normalisation doit d'abord être effectuée variable par variable, de façon que l'indice se présente comme une somme de contributions normalisées. Pour ce dernier, on considère une nouvelle normalisation de même type avant l'application de l'Algorithme de la Vraisemblance du Lien.

1. Contribution brute d'une variable à la comparaison de deux objets

Pour ne pas déconnecter la mesure d'une variable v^j des mesures des autres variables sur le même objet o_i dont il y a lieu de préserver l'entité, nous proposerons comme mesure réduite de la $j^{ème}$ variable sur o_i :

$$w^j(o_i) = \frac{v^j(o_i)}{\sqrt{\sum_j (v^j(o_i))^2}} = \frac{x_i^j}{\sqrt{\sum_j (x_i^j)^2}} = \eta_i^j \,. \tag{4}$$

Une telle mesure réduite est évidemment invariante par homothétie portée sur la suite des mesures initiales.

La contribution brute de la $j^{ème}$ variable à la comparaison de deux objets o_i et $o_{i'}$ sera de façon multiplicative posée comme suit :

$$s_j(o_i, o_{i'}) = \eta_i^j \eta_{i'}^j \,. \tag{5}$$

La somme pour $j = 1, \ldots, m$, des $s_j(o_i, o_{i'})$ est l'indice COS $(o_i, o_{i'})$ (cf. (2)) que nous ne considérons pas ici, mais dont nous nous sommes inspirés pour déterminer les contributions élémentaires (4) et (5).

2. Moyenne sur $P_2(I)$ de $s_j(o_i, o_{i'})$

Nous désignerons par M^j cette moyenne. On a :

$$M^j = \frac{2}{n(n-1)} \sum \{\eta_i^j \eta_{i'}^j / \{i, i'\} \in P_2(I)\} \tag{6}$$

$$= \frac{n}{(n-1)} \{(\mu(w^j))^2 - \frac{1}{n} \mu_2(w^j)\} , \tag{7}$$

où $\mu(w^j)$ et $\mu_2(w^j)$ sont la moyenne et le moment d'ordre 2 de w^j :

$$\mu(w^j) = \eta_{\cdot}^j = \frac{1}{n} \sum_i \eta_i^j$$

et

$$\mu_2(w^j) = \frac{1}{n} \sum_i (\eta_i^j)^2$$

Le passage de (6) à (7) repose sur l'identité

$$2 \sum \{\eta_i^j \eta_{i'}^j / \{i, i'\} \in P_2(I)\} = \left(\sum_i \eta_i^j \right)^2 - \sum_i (\eta_i^j)^2$$

3. Variance sur $P_2(I)$ de $s_j(o_i, o_{i'})$

La structure du calcul du moment absolu d'ordre 2 est la même que celle de la moyenne M^j ci-dessus. On a — en désignant par M_2^j ce moment —

$$M_2^j = \frac{n}{(n-1)} \{(\mu_2(w^j))^2 - \frac{1}{n} \mu_4(w^j)\} , \tag{8}$$

où $\mu_4(w^j) = \frac{1}{n} \sum_i (\eta_i^j)^4$ est le moment absolu d'ordre 4 de w^j.

On a, très sensiblement,

$$M^j = \{(\mu(w^j))^2 - \frac{1}{n} \mu_2(w^j)\} \tag{9}$$

et

$$M_2^j = \{(\mu_2(w^j))^2 - \frac{1}{n} \mu_4(w^j)\} \tag{10}$$

De sorte que la variance $(\sigma^j)^2$ s'écrit

$$(\sigma^j)^2 = (\mu_2(w^j))^2 - (\mu(w^j))^4 - \frac{1}{n} \{\mu_4(w^j) - 2\mu_2(w^j)(\mu(w^j))^2 + \frac{1}{n} (\mu_2(w^j))\} , \tag{11}$$

dont la partie dominante est

$$(\mu_2(w^j))^2 - (\mu(w^j))^4 . \tag{12}$$

4. L'indice

L'indice que nous proposons entre les deux objets o_i et $o_{i'}$ se met sous la forme de la somme réduite des contributions normalisées des différentes variables :

$$s(o_i, o_{i'}) = \frac{1}{\sqrt{m}} \sum_{1 \leqslant j \leqslant m} (s_j(o_i, o_{i'}) - M^j)/\sigma^j \ , \tag{13}$$

où la réduction au moyen de $1/\sqrt{m}$ se réfère à un modèle d'indépendance où les variables aléatoires associées aux v^j, $1 \leqslant j \leqslant m$, ont une variance unité.

De toute façon, cette réduction n'intervient plus après la réduction globale des similarités où on substitue à la table

$$\{S(o_i, o_{i'})/\{i, i'\} \in P_2(I)\} \ , \tag{14}$$

celle

$$\{T(o_i, o_{i'})/\{i, i'\} \in P_2(I)\} \ , \tag{15}$$

avec

$$T(o_i, o_{i'}) = (S(o_i, o_{i'}) - \overline{S})/\sqrt{\text{var}(S)} \ , \tag{16}$$

où $\overline{S}$ et var (S) sont respectivement la moyenne et la variance de la table (14).

La table qui est directement l'argument de l'algorithme de la vraisemblance du lien, se met sous la forme

$$\{P(i, i')/\{i, i'\} \in P_2(I)\} \ , \tag{17}$$

où

$$P(i, i') = \Phi(T(o_i, o_{i'})) \ ,$$

où Φ est la fonction de répartition de la loi N (0, 1), normale centrée-réduite.

5. Prise en compte d'une pondération des variables

Nous avons déjà signalé que la prise en compte a priori d'une pondération des variables est tout à fait problématique et discutable, même si cette pondération est basée sur une méthode objective [LERMAN (1970b), SNEATH & SOKAL (1972)]. Néanmoins, il faut laisser la porte ouverte aux possibilités expérimentales de l'analyse classificatoire des données.

Si $\alpha_1, \alpha_2, ..., \alpha_j, ..., \alpha_m$ est une suite de coefficients positifs de somme unité, définissant les importances respectives qu'on veut donner aux différentes variables pour l'évaluation de la similarité entre deux objets. La forme additive de l'indice S (cf. (13)) permet de les intégrer au moyen de l'expression suivante :

$$S'(i, i'; \alpha_1, ..., \alpha_m) = \sqrt{m} \sum_{1 \leqslant j \leqslant m} \alpha_j (s_j(i, i') - M^j)/\sigma^j \ , \tag{18}$$

où nous avons noté i pour o_i.

Certains spécialistes de l'analyse factorielle conseillent la classification des objets décrits par les quelques premiers facteurs. Dans cette démarche, c'est une pondération implicite des variables qui est — par rapport au modèle factoriel — prise en compte. Cette pondération peut être plus explicitement mise en évidence dans notre indice si on lie α_j au coefficient défini par la somme des carrés des coefficients de corrélation entre v^j et les quelques facteurs retenus, $1 \leqslant j \leqslant m$.

On peut envisager d'autres pondérations objectives. Nous avons pu définir — de façon intrinsèque au tableau des données — « l'importance projective » d'une

même variable au moyen de la variance de ses indices d'association avec les autres variables [LERMAN (1981), Chap. 3]. Dans ces conditions, on peut lier α_i à « l'importance projective » de la variable v^j, $1 \leqslant j \leqslant m$.

Nous nous sommes ci-dessus posé la question de savoir dans quelle mesure la variance d'une variable doit discriminer la ressemblance entre deux objets. Dans notre démarche ci-dessus, nous avons apporté une forme de neutralisation de cette variance, mais au niveau de l'ensemble des paires d'objets. On peut — compte tenu de la forme de l'indice — chercher à réintroduire l'importance relative des différentes variances s_j^2 (s_j^2 = var. (v^j)), $1 \leqslant j \leqslant m$, à partir de coefficients : α_1, α_2 ,..., α_j ,..., α_m, en liant α_j à s_j^2 d'une façon qui reste à préciser.

Ce problème de pondération se pose dans les mêmes termes quel que soit le type de variable. De sorte que dans la suite nous n'évoquerons plus ce problème. Toutefois, en ce qui concerne les méthodes « objectives » de pondération, certaines notions de variance et d'analyse factorielle deviennent — dans le cas où les variables sont qualitatives ou relationnelles — très délicates à circonscrir.

III. Cas où les variables sont des attributs de description (variables logiques (0, 1))

Nous allons suivre les mêmes étapes que dans le cas où les variables sont numériques.

Si on ignore la normalisation — variable par variable — on peut a priori se référer à trois indices d'inspirations voisines. Le premier a la même expression formelle que le cosinus (cf. formule (2) §II) et correspond à l'indice d'Ochiai [OCHIAI (1957)]. Il peut se mettre sous la forme :

$$s_0(i, i') = \frac{\sum_j \xi_i^j \xi_{i'}^j}{\left(\sum_j \xi_i^j\right)\left(\sum_j \xi_{i'}^j\right)}, \qquad (1)$$

où ξ_i^j (resp. $\xi_{i'}^j$) = 1 ou 0 selon que l'attribut j est présent ou absent chez l'objet o_i (resp. $o_{i'}$).

Les deux autres indices résultent directement — par transposition des rôles des lignes et des colonnes — de ceux conçus pour la comparaison des attributs de description [LERMAN (1981), Chap. 2]. A l'exception de la structure parfaitement symétrique d'un tableau de contingence et de ses dérivés, on peut davantage admettre — que dans le cas des autres types de tableaux de données — cette transposition des rôles.

Ces deux indices peuvent respectivement être mis sous la forme :

$$s_1(i, i') = \frac{\sum_j (\xi_i^j - p_i^\cdot)(\xi_{i'}^j - p_{i'}^\cdot)}{m \sqrt{p_i^\cdot (1 - p_i^\cdot)\, p_{i'}^\cdot (1 - p_{i'}^\cdot)}} \qquad (2)$$

où $p_i^\cdot$ (resp. $p_{i'}^\cdot$) est la proportion d'attributs présents chez l'objet i (resp. i') et

$$s_2(i, i') = \frac{\sum_j (\xi_i^j - p_i^\cdot)(\xi_{i'}^j - p_{i'}^\cdot)}{m \sqrt{p_i^\cdot\, p_{i'}^\cdot}} \qquad (3)$$

Lorsque les indices (2) et (3) sont conçus dans la situation transposée de la comparaison d'attributs de description, c'est une même méthode (cf. référence ci-dessus mentionnée) qui nous permet de les obtenir, respectivement par rapport à deux modèles aléatoires de l'hypothèse d'absence de liaison. L'indice (2) correspond à celui de K. PEARSON.

1. Contribution brute d'un attribut a^j à la comparaison de deux objets o_i et $o_{i'}$.

Nous allons considérer a priori trois formes de cette contribution qui, respectivement, correspondent à chacun des indices (1), (2) et (3) :

$$s_{0j}(i, i') = \frac{\xi_i^j \xi_{i'}^j}{\sqrt{p_i p_{i'}}}, \tag{4}$$

et

$$s_{1j}(i, i') = \frac{(\xi_i^j - p_i)(\xi_{i'}^j - p_{i'})}{\sqrt{p_i(1 - p_i)\, p_{i'}(1 - p_{i'})}}, \tag{5}$$

$$s_{2j}(i, i') = \frac{(\xi_i^j - p_i)(\xi_{i'}^j - p_{i'})}{\sqrt{p_i\, p_i}}. \tag{6}$$

En posant

$$\eta_{0i}^j = \frac{\xi_i^j}{\sqrt{p_i}}, \tag{4'}$$

$$\eta_{1i}^j = \frac{(\xi_i^j - p_i)}{\sqrt{p_i(1 - p_i)}} \tag{5'}$$

et

$$\eta_{2i}^j = \frac{(\xi_i^j - p_i)}{\sqrt{p_i}}, \tag{6'}$$

on a

$$s_{0j}(i, i') = \eta_{0i}^j \eta_{0i'}^j, \tag{4''}$$

$$s_{1j}(i, i') = \eta_{1i}^j \eta_{1i'}^j, \tag{5''}$$

et

$$s_{2j}(i, i') = \eta_{2i}^j \eta_{2i'}^j. \tag{6''}$$

2. Moyenne et Variance sur $P_2(I)$ de $s_j(i, i')$; proposition de l'indice de similarité

($\alpha = 0$, 1 ou 2 conformément aux expressions (4''), (5'') et (6'')).

Les calculs sont ceux des paragraphes II.2 et 3 ci-dessus. On remplacera selon les cas η_i^j par η_{0i}^j, η_{1i}^j ou η_{2i}^j. En désignant par M_α^j et $(\sigma_\alpha^j)^2$ la moyenne et la variance — sur $P_2(I)$ — de $s_{\alpha j}(i, i')$, on a l'expression de l'indice qui correspond à celui (13) du paragraphe II ci-dessus :

$$S_\alpha(o_i, o_{i'}) = \frac{1}{\sqrt{m}} \sum_{1 \leqslant j \leqslant m} (s_{\alpha j}(o_i, o_{i'}) - M_\alpha^j)/\sigma_{\alpha'}^j \qquad (7)$$

lequel pourra prendre trois formes selon que $\alpha = 0$, 1 ou 2.

Les dernières étapes avant l'application de l'algorithme de la vraisemblance du lien correspondent à la réduction globale — sur $P_2(I)$ — des similarités S et à la transformation de l'échelle de mesure des similarités en une échelle de probabilité ou de fréquence mathématique. Les expressions formulées sont celles (14) à (17) du paragraphe II où il y a lieu de substituer S_α à S, T_α à T et P_α à P, avec $\alpha = 0$, 1 ou 2 correspondants aux trois formes de l'indice.

IV. Cas où les variables sont qualitatives nominales

Désignons par C l'ensemble des variables qui sont ici des caractères descriptifs où l'ensemble des modalités d'un même caractère n'est muni d'aucune structure. On désigne ici — et dans la suite — par Q le cardinal de C. J_q indiquera l'ensemble des codes des modalités de la variable qualitative c_q ($c_q \in C$, $1 \leqslant q \leqslant Q$). De façon plus précise, on posera $J_q = \{j_q / 1 \leqslant j_q \leqslant m_q\}$ où m_q est le nombre de modalités de la variable c_q, $1 \leqslant q \leqslant Q$.

On se ramène à un codage en « 0-1 » du type « absence-présence » en associant à chacune des modalités d'un même caractère un attribut de description que nous appelons « attribut-modalité » et dont la valeur est 0 ou 1 selon que la modalité en question n'est pas ou est possédée. On a ainsi ce que les factorialistes appellent un « codage disjonctif complet ». Ainsi, le codage de la réponse d'un individu ou objet au caractère c_q à m_q modalités est un vecteur logique à m_q composantes dont la $j_q^{ème}$ est égale à 1 et les autres à 0, si et seulement si l'individu ou objet possède la $j_q^{ème}$ modalité du caractère c_q.

De la sorte, la représentation d'un objet — par rapport à la suite $\{c_q / 1 \leqslant q \leqslant Q\}$ des variables — se fait au moyen d'un vecteur logique à $m = \Sigma\{m_q / 1 \leqslant q \leqslant Q\}$ composantes dont exactement Q sont égales à 1 et où la $q^{ème}$ composante égale 1 se situe entre la $(m_1 + m_2 + ... + m_{(q-1)})$ position et celle $(m_1 + m_2 + ... + m_q)$. Cependant, on aura soin d'effectuer tout calcul à partir du codage initial qui est beaucoup plus économique en place mémoire.

1. Contribution brute d'une variable c_q à la comparaison de deux objets o_i et $o_{i'}$

Nous sommes dans une situation où le nombre de composantes égales à 1 dans la description de chaque objet est constante et égale à Q. Il est dans ces conditions important de remarquer qu'il n'y a pas lieu de réduire la contribution de la « mesure » d'une même variable c_q par rapport à l'ensemble de toutes les variables.

De façon tout à fait naturelle, on posera

$$s_c(i, i') = \begin{cases} 1 \text{ si } o_i \text{ et } o_{i'} \text{ possèdent la même modalité de c} \\ 0 \text{ si } o_i \text{ et } o_{i'} \text{ ne possèdent pas la même modalité de c.} \end{cases}$$

Dans ces conditions, en désignant par I_1^c, I_2^c, ..., I_{mq}^c la partition de I définie par la variable qualitative c_q, on a :

$$s_c(i, i') = 1 \Leftrightarrow \{i, i'\} \in \sum_{1 \leqslant jq \leqslant mq} P_2(I_{jq}), \tag{1}$$

où, rappelons-le, $P_2(I_{jq})$ est l'ensemble des paires ou parties à deux éléments de I_{jq}.

Nous allons à présent calculer la *moyenne* et la *variance* empiriques de $s_c(i, i')$ sur l'ensemble $P_2(I)$. $s_c(i, i')$ définit la contribution brute de c à la comparaison de o_i et $o_{i'}$. On désignera par n_j^c le cardinal de I_j^c, $1 \leqslant j \leqslant m_q$.

2. Moyenne et variance de $s_c(i, i')$. Proposition de l'indice de similarité

L'indice q étant fixé, on l'omettra dans les calculs qui suivent.

On a, en vertu de (1)

$$\sum \{s_c(i, i')/\{i, i'\} \in P_2(I)\} = \sum_{1 \leqslant j \leqslant mq} (n_j^c (n_j^c - 1)/2), \tag{2}$$

de sorte que la moyenne de $s_c(i, i')$ — sur $P_2(I)$ — se met sous la forme :

$$M^c = \sum_{1 \leqslant j \leqslant mq} (n_j^c (n^c{}_j - 1)/n (n - 1)), \tag{3}$$

On a

$$s_c^2(i, i') = s_c(i, i'),$$

de sorte que le moment d'ordre 2 de $s_c(i, i')$ s'écrit

$$M_2^c = \sum_{1 \leqslant j \leqslant mq} (n_j^c (n^c{}_j - 1)/n (n - 1)). \tag{4}$$

Finalement,

$$(\sigma^2)^c = \text{var.}\,(s_c(i, i')) = \sum_{1 \leqslant j \leqslant mc} (n_j^c (n_j^c - 1)/n (n - 1)) - \sum_{1 \leqslant j \leqslant mc} (n_j^c (n_j^c - 1)/n (n - 1))^2. \tag{5}$$

Dans ces conditions, l'indice de similarité entre les deux objets o_i et $o_{i'}$, tenant également compte de l'ensemble des variables, se met sous la forme :

$$S(o_i, o_{i'}) = \frac{1}{\sqrt{Q}} \sum_{1 \leqslant q \leqslant Q} \frac{s_{cq}(i, i') - M^{cq}}{\sqrt{\sigma^c q}}, \tag{6}$$

avec des notations ci-dessus explicitées.

On reprendra ici le sens de l'expression du dernier alinéa du paragraphe III ci-dessus.

V. Cas où les variables sont qualitatives ordinales

Les notations sont exactement les mêmes qu'au paragraphe IV ci-dessus. La différence est que l'ensemble J_q des modalités d'une même variable c_q, se trouve muni d'un ordre total pour lequel le rang de la $j_q^{ème}$ modalité est j_q, $1 \leqslant j_q \leqslant m_q$. On posera $c_q(o_i) = (j_{q-1})$ si l'objet o_i possède la $j_q^{ème}$ modalité de la variable c_q.

Nous avons été conduits dans [LERMAN (1981), Chap. 2, §IV.3] à prendre comme contribution brute d'une variable qualitative ordinale c_q à la comparaison de deux objets o_i et $o_{i'}$ l'expression

$$s_{c_q}(i, i') = (m_{c_q} - 1) - |c_q(i) - c_q(i')| \tag{1}$$

Cette expression fait suite à un codage de la réponse à la variable c_q au moyen d'un vecteur logique à $2(m_q - 1)$ composantes où l'attribut défini par la $h^{ème}$ s'exprime de la façon suivante :

« code initial j_q strictement inférieur à $(m_q - h + 1)$ » pour $1 \leqslant h \leqslant m_{q'}$
« code initial j_q supérieur ou égal à $(2m_q - h)$ » pour $h \geqslant m_q$.

De la sorte, il y a exactement $(m_q - 1)$ composantes égales à 1 qui sont réparties aux extrémités du vecteur logique codant la réponse de l'objet o_i et où $c_q(o_i)$ est exactement le nombre de composantes égales à 1 qui se trouvent à l'extrémité droite du vecteur.

Ainsi, relativement à l'ensemble C des variables, le vecteur logique de description d'un même objet comporte un nombre de composantes égales à 1, indépendant de l'objet et égal $\sum \{(m_q - 1)/1 \leqslant q \leqslant Q\}$. Dans ces conditions — comme dans le cas nominal — il n'y a pas lieu de réduire la contribution de la « mesure » d'une même variable c_q par rapport à l'ensemble de toutes les variables.

1. Moyenne et variance de $s_c(i, i')$. Proposition de l'indice de similarité

Le calcul repose sur la décomposition de $P_2(I)$ conformément à la partition $\{I_j^c / 1 \leqslant j \leqslant m\}$. On a

$$P_2(I) = \sum_{1 \leqslant j \leqslant m} P_2(I_j^c) + \sum_{1 \leqslant j < h \leqslant m} I_j^c * I_{h'}^c \tag{2}$$

où $I_j^c * I_h^c$ désigne l'ensemble des paires $\{i, i'\}$ où $i \in I_j^c$ et $i' \in I_h^c$.

Le calcul de la moyenne de $s_c(i, i')$ sur $P_2(I)$ repose sur celui de la somme de $|c(i) - c(i')|$. Dans cette dernière :

$$\sum \{|c(i) - c(i')| / \{i, i'\} \in P_2(I)\} , \tag{3}$$

la contribution de $\{i, i'\}$ est nulle si $\{i, i'\} \in P_2(I_j^c)$, $1 \leqslant j \leqslant m$. Dans ces conditions, la somme (3) se réduit à

$$\sum_{1 \leqslant j < h \leqslant m} \; \sum_{\{i,i'\} \in I_j^c * I_h^c} (|c(i) - c(i')|) = \sum_{1 \leqslant j < h \leqslant m} n_j^c n_h^c (h - j) . \tag{4}$$

D'où la moyenne de $s_c(i, i')$:

$$M^c = (m - 1) - \sum_{1 \leqslant j < h \leqslant m} \frac{2 n_j^c n_h^c}{n(n - 1)} \times (h - j) \tag{5}$$

et le moment absolu d'ordre 2 de $|c(i) - c(i')|$:

$$\sum_{1 \leqslant j < h \leqslant m} \frac{2n_j^c n_h^c}{n(n-1)} \times (h - j) \tag{6}$$

D'où la variance $(\sigma^2)^c$ de $s_c(i, i')$:

$$V^c = \sum_{1 \leqslant j < h \leqslant m} \frac{2n_j^c n_h^c}{n(n-1)} \times (h - j)^2 - \left(\sum_{1 \leqslant j < h \leqslant m} \frac{2n_j^c n_h^c}{n(n-1)} \times (h - j) \right)^2 . \tag{7}$$

Comme dans le cas qualitatif nominal, l'indice de similarité entre deux objets o_i et $o_{i'}$, tenant également compte de l'ensemble des variables, se met sous la forme :

$$S(o_i, o_{i'}) = \frac{1}{\sqrt{Q}} \sum_{1 \leqslant q \leqslant Q} \frac{s_{c_q}(i, i') - M^{c_q}}{\sqrt{V^{c_q}}} . \tag{8}$$

On continuera à reprendre ici le sens de l'expression du dernier alinéa du paragraphe III.

VI. Cas où les variables sont des préordonnances ou des graphes values

Les notations sont les mêmes qu'aux paragraphes IV et V ci-dessus.

VI.1. Structure de similarité sur J_q (q fixé)

Tout en restant extrêmement générale, la structure descriptive la plus riche et la moins arbitraire d'une variable quantitative est fournie par une préordonnance totale sur l'ensemble de ses modalités (cf. [PETER (1987)] et aussi dans un tout autre contexte [CHAH (1984)]).

Dans ces conditions — relativement à la variable c_q — on introduit l'ensemble suivant des couples de modalités :

$$H_q = \{(j_q, h_q) / 1 \leqslant j_q \leqslant h_q \leqslant m_q\} , \tag{1}$$

sur lequel se trouve défini — par l'expert — un préordre total ω_q (i.e. préordonnance sur J_q) pour lequel un couple (j_q, h_q) est d'autant plus grand — d'un point de vue ordinal — que la modalité j_q ressemble à celle h_q ; ainsi, la dernière classe de ce préordre comporte m_q termes de la forme (j_q, j_q), $1 \leqslant j_q \leqslant m_q$.

Exemple

Dans le cas d'un problème d'organisation d'un important corpus de petites annonces immobilières [PETER (1987)], on considère la variable « objet de la transaction » dont la suite des modalités — respectivement codées 1, 2, 3, 4, 5, 6, 7, 8, 9 — est : « maison », « pavillon », « appartement », « habitation », « studio », « chambre », « local », « garage », « terrain ». On peut proposer la préordonnance :

$$15 \sim 16 \sim 17 \sim 18 \sim 19 \sim 25 \sim 26 \sim 27 \sim 28 \sim 29 \sim 36 \sim 37 \sim 38 \sim 39 \sim 46 \sim 47 \sim 48 \sim 49 \sim 57 \sim 58 \sim 59 \sim 67 \sim 68 \sim 69 \sim 78 \sim 79 \sim 89 < 13 \sim 23 \sim 35 \sim 45 < 14 \sim 24 \sim 34 \sim 56 < 12 < 11 \sim 22 \sim 33 \sim 44 \sim 55 \sim 66 \sim 77 \sim 88 \sim 99,$$

où ij avec $i \leqslant j$ indique le couple (i, j).

Dans l'introduction même de H_q, nous admettons le caractère symétrique de la notion de similarité. Sinon, comme cela pourrait se présenter dans un problème d'affectation, il suffit de définir le préordre total sur l'ensemble $K_q = J_q \times J_q$ et les mêmes considérations ci-dessous — conceptuelles et de calcul — restent valables.

A chaque élément de l'ensemble préordonné (H_q dans notre cas) on associe un « rang ». Pour définir précisément la fonction ordinale « rang » désignons par $(\ell_1, \ell_2, ..., \ell_k)$ la suite des cardinaux de la suite ordonnée des classes du préordre total. Le rang d'un élément appartenant à la $j^{ème}$ classe, $1 \leqslant j \leqslant k$, est posé égal à

$$\sum_{1 \leqslant i \leqslant (j-1)} \ell_i + (\ell_j + 1)/2$$

Ainsi, relativement à l'exemple ci-dessus, le rang de l'élement 24 est égal à $27 + 4 + 2 = 33$. De la sorte, la somme de tous les rangs est — comme dans le cas totalement et strictement ordinal — égal à $L(L + 1)/2$, où $L = \ell_1 + \ell_2 + ... + \ell_k$. La structure descriptive sera donc basée sur le tableau des rangs ainsi calculés :

$$\{r_{j_q h_q}/(j_q, h_q) \in H_q\} . \tag{2}$$

Une autre forme, plus riche mais moins générale de la relation de similarité sur J_q, pour une fine description de E, est fournie au moyen d'une table numérique indexée par $K_q = J_q \times J_q$, où le nombre qui se trouve à l'intersection de la ligne j_q et de la colonne h_q est sensé « mesurer » le degré de ressemblance entre les deux modalités j_q et h_q. Cette table de nombres qu'on peut admettre — sans que cela soit nécessaire pour les calculs — symétrique, est supposée donnée par l'« expert ». Nous l'écrivons sous la forme

$$\{p_{j_q h_q}/(j_q, h_q) \in J_q \times J_q\} ,$$

ou, plus simplement, en tenant compte de la symétrie,

$$\{p_{j_q h_q}/(j_q, h_q) \in H_q\} . \tag{3}$$

En réalité, la nature des calculs sera exactement la même qu'on travaille avec la table (2) des rangs, ou avec celle (3) des coefficients numériques, de sorte qu'on désignera par

$$\{s_{j_q h_q}/(j_q, h_q) \in H_q\} , \tag{4}$$

l'une ou l'autre des deux tables.

Si l'objet o_i (resp. $o_{i'}$) possède la modalité j_q (resp. h_q), $s_{j_q h_q}$ définira la contribution brute de la $q^{ème}$ variable à la ressemblance entre o_i et $o_{i'}$.

Nous commencerons par déterminer la contribution réduite d'une même variable c_q (dont l'ensemble des modalités est codé par J_q) à la similarité entre deux objets, puis — de façon égale et parallèle — nous intégrerons l'ensemble des variables.

VI.2. Contribution de J_q à la ressemblance entre deux objets

L'indice q restant fixé dans ce paragraphe, nous l'omettrons pour des raisons de simplicité d'écriture.

L'élaboration de l'indice obéit dans notre approche à un principe statistique général de construction, où à partir d'un premier indice, localement défini, une

normalisation est effectuée à partir de la distribution empirique de cet indice sur l'ensemble $P_2(E)$ des paires d'objets (ou parties à deux éléments) de E.

x et y désignant les deux objets à comparer, si $c(x) = j_0$ et $c(y) = h_0$, l'indice sera localement défini par le nombre $s_{j_0h_0}$ de la table (4). Il y a lieu par conséquent de préciser la distribution de $\{s_{jh}/(j, h) \in H\}$ sur $P_2(E)$. Cette distribution s'obtient très aisément à partir de la décomposition de $P_2(E)$ conformément à la partition de E déterminée par la variable qualitative c.

Plus directement, en désignant par n (j) le cardinal de la classe d'objets E_j, possédant la $j^{ème}$ modalité du caractère c dont nous désignons par m le nombre de modalités,

$$\text{card } P_2(E) = \sum_{1 \leqslant j \leqslant m} n(j)\,(n(j) - 1)/2 + \sum_{1 \leqslant j < h \leqslant m} n(j)\,n(h)) \,. \qquad (5)$$
$$= n(n - 1)/2.$$

Désignons respectivement par

$$\rho_j = n(j)\,(n(j) - 1)/n(n - 1) \quad \text{et} \quad \sigma_{jh} = 2n(j)\,n(h)/n(n - 1) \,, \qquad (6)$$

la proportion de paires d'objets $\{x', y'\}$ dont les deux composantes sont dans la classe E_j et celle pour lesquelles x' est dans la classe E_j et y' dans celle $E_{h'}$, $1 \leqslant j \leqslant m$ et $1 \leqslant j < h \leqslant m$.

Désignons encore par

$$\rho = \sum_{1 \leqslant j \leqslant m} \rho_j \quad \text{et} \quad \sigma = \sum_{1 \leqslant j < h \leqslant m} \sigma_{jh'} \qquad (7)$$

qui sont respectivement, la proportion de paires réunies et séparées par la partition $\{E_j/1 \leqslant j \leqslant m\}$.

Toutes les paires d'objets appartenant à E_j (resp. $E_j * E_h$) ont la même valeur s_{jj} (resp. s_{jh}) de l'indice local.

La distribution de la table des similarités $\{s_{jh}/(j, h) \in H\}$ sur $P_2(E)$ est donc définie par

$$\{(s_{kk}, \rho_k), (s_{jh}, \rho_{jh})/k \in J \quad \text{et} \quad (j, h) \in J^{[2]}\} \,, \qquad (8)$$

où nous avons noté $J^{[2]} = \{(j, h)/1 \leqslant j < h \leqslant m\}$.

Calcul de la moyenne et de la variance de la distribution (8)

Si μ et $\mathcal{M}_2$ désignent la moyenne et le moment absolu d'ordre 2, on a :

$$\mu = \sum_{1 \leqslant k \leqslant m} \sigma_k s_{kk} + \sum_{1 \leqslant j < h \leqslant m} \sigma_{jh} s_{jh} \,, \qquad (9)$$

$$\mathcal{M}_2 = \sum_{1 \leqslant k \leqslant m} \rho_k s_{kk}^2 + \sum_{1 \leqslant j < h \leqslant m} \sigma_{jh} s_{jh}^2 \,. \qquad (10)$$

On suppose — ce qui est naturel — que s_{kk} est le même pour tout $k = 1, 2, ..., m$. Notons s cette valeur commune.

$$\mu = \rho s + \sum_{1 \leqslant j < h \leqslant m} \sigma_{jh} s_{jh} \,, \qquad (9')$$

$$\mathcal{M}_2 = s^2 \sum_{1 \leqslant j < h \leqslant m} \sigma_{jh} s_{jh}^2 \,. \qquad (10')$$

Le calcul informatique de la variance que nous notons ici $\mathscr{V}$, utilise

$$\mathscr{V} = \mathscr{M}_2 - \mu^2 \ . \tag{11}$$

Plus directement,

$$\mathscr{V} = \rho\,(\sigma s - \sum_{j<h} \sigma_{jh} s_{jh})^2 + \sum_{1 \leqslant j < h \leqslant m} \sigma_{jh}(s_{jh} - \rho s - \sum_{\ell < k} \sigma_k s_{\ell k})^2 \ , \tag{12}$$

qui se met sous la forme

$$\sum_{1 \leqslant j < h \leqslant m} \sigma_{jh} \left(\sum_{1 \leqslant \ell < k \leqslant m} \sigma_{\ell k}(s_{jh} - s_{\ell k}) \right)^2 \ , \tag{13}$$

et ayant au préalable noté σ_{kk} pour ρ_k, $1 \leqslant k \leqslant m$.

Le numérateur de l'indice d'association s'écrit :

$$s_{j_o h_o} - \rho s - \sum_{1 \leqslant j < h \leqslant m} \sigma_{jh} s_{jh} \ , \tag{14}$$

qu'on peut mettre sous la forme

$$\sum_{1 \leqslant \ell < k \leqslant m} \sigma_{\ell k}(s_{j_o h_o} - s_{\ell k}) \ . \tag{15}$$

L'indice d'association entre les deux objets x et y, relativement à la variable en question est égal à

$$S(x, y) = \frac{\sum_{1 \leqslant \ell < k \leqslant m} \sigma_{\ell k}(s_{j_o h_o} - s_{\ell k})}{\left\{ \sum_{j \leqslant k} \sigma_{jh} \left(\sum_{\ell \leqslant k} \sigma_{\ell k}(s_{jh} - s_{\ell k}) \right)^2 \right\}^{1/2}} \ . \tag{16}$$

Nous allons à présent établir une propriété d'invariance de cet indice lorsque le nombre de modalités de la variable qualitative est deux.

Dans ce cas là, on a

$H = \{(1,1), (1,2), (2,2)\}$ et on posera

$s_{12} = p$, $s_{11} = s_{22} = q$, où $p < q$.

Dans ces conditions, la moyenne et la variance de la distribution des indices locaux sont respectivement égaux à

$$\begin{aligned} \mu &= p\sigma + q\rho \\ \mathscr{V} &= p^2\sigma + q^2\rho - (p\sigma + q\rho)^2 \end{aligned} \tag{17}$$

Si les deux objets à comparer possèdent deux modalités distinctes, la valeur de l'indice S s'écrit :

$$\frac{p - (p\sigma + q\rho)}{\sqrt{p^2\sigma + q^2\rho - (p\sigma + q\rho)^2}} \tag{18}$$

qui — après calcul — se réduit à

$$-\sqrt{\rho/\sigma} \tag{19}$$

Si les deux objets à comparer possèdent la même modalité, on remplacera le numérateur de (18) par [q − (p + q)] et l'indice S se réduit à

$$+ \sqrt{\sigma/\rho} \tag{20}$$

D'où l'énoncé du résultat :

Propriété

Si le nombre de modalités de la variable qualitative se réduit à deux, l'indice globalement réduit ne dépend plus que de la répartition des deux modalités sur l'ensemble des objets.

De façon précise, si n (1) (resp. n (2)) est le nombre d'objets possédant la modalité 1 (resp. 2), deux objets x et y possédant respectivement les modalités 1 et 2, ont pour valeur de l'indice S :

$$S(x, y) = - \sqrt{\frac{1}{2}\left(\frac{(n(1) - 1)}{n(2)} + \frac{(n(2) - 1)}{n(1)}\right)}. \tag{19'}$$

Si par contre les deux objets ont la même modalité :

$$S(x, y) = + \sqrt{2\left\{\frac{n(1)\, n(2)}{n(1)\,(n(1) - 1) + n(2)\,(n(2) - 1)}\right\}}. \tag{20'}$$

Cette propriété qui peut surprendre est en fait heureuse et naturelle : la perception des ressemblances mutuelles entre objets pris dans un même ensemble, face à une simple dichotomie, n'a plus à dépendre d'une « quantification » de cette dichotomie.

Considérons deux objets face à plusieurs variables dichotomiques. Les formules (19′) et (20′) montrent que la contribution d'une même variable à la ressemblance des deux objets qui en possèdent la même modalité (resp. qui n'en possèdent pas la même modalité) est d'autant plus élevée (resp. faible) que les deux modalités se trouvent plus également réparties.

La propriété d'invariance ci-dessus — par rapport à la table (4) des similarités locales — n'est plus valable si la variable a plus de deux modalités.

Pour la plupart des applications, une information très générale de type « préordonnance » sur H est suffisamment fine pour une excellente reconnaissance des classes de proximité sur l'ensemble des objets.

La prise en compte de variables « préordonnance » permet d'enrichir sensiblement la structure descriptive des variables qualitatives et ce, à partir de la perception a priori de la ressemblance entre modalités d'une même variable. Ainsi, même dans le cas le plus pauvre d'une variable logique où a (resp. $\overline{a}$) désigne la présence (resp. absence) de l'attribut, on peut définir la préordonnance : $a\overline{a} < \overline{a}\overline{a} < aa$, où la présence commune est plus indicative de la ressemblance que l'absence commune.

VI.3. Indice d'association dans le cas de plusieurs variables

Désignons par $S_q(x, y)$ la contribution de la variable c_q à la comparaison des deux objets x et y. $S_q(x, y)$ est donné par la formule (16) ci-dessus relativement à la table (4).

Pour un même q, la moyenne et la variance de S_q sur l'ensemble $P_2(E)$ sont respectivement égales à 0 et à 1.

Pour la définition de l'indice d'association, on tiendra également compte des différentes variables en proposant

$$S(x, y) = \frac{1}{\sqrt{Q}} \sum_{1 \leq q \leq Q} S_q(x, y). \tag{21}$$

Rappelons que la réduction au moyen de $1/\sqrt{Q}$ se réfère à un modèle d'indépendance où les v.a. associés aux c_q, $1 \leqslant q \leqslant Q$, ont une variance unité.

Encore une fois, on reprendra ici le dernier alinéa du paragraphe III.

VI.4. Une solution efficace et simple au problème du consensus entre arbres de classifications

Ces dernières années, on a vu se développer un intérêt tout particulier pour le problème de la recherche d'un consensus entre arbres de classifications sur le même ensemble [ROHLF (1982)], [DIDAY (1982)], [BARTHELEMY, LECLERC, MONJARDET (1984)], ... Etant donné un ensemble fini d'arbres de classifications ou hiérarchies indicées H_1, H_2, ..., H_Q, sur le même ensemble fini E d'objets, il s'agit de résumer « au mieux » cet ensemble de hiérarchies au moyen d'un arbre unique de classifications par rapport auquel on pourra situer les différents arbres.

Le problème concret peut se présenter dans le cas d'un tableau de données $E \times V$ (E : ensemble des objets et V : ensemble des variables descriptives), où les variables sont mesurées à différentes dates. Pour chaque date, le tableau des mesures (indexé par $E \times V$) conduit — par une méthode de classification fixée — à un arbre de classification sur E. Pour une période relativement stable, on peut vouloir — sur la seule base des arbres obtenus — une organisation classificatoire hiérarchique « moyenne » de l'ensemble E des objets. Mais alors, pour une telle situation, on peut poser le problème de savoir s'il ne vaut pas mieux obtenir cette classification hiérarchique « moyenne » à partir d'un indice tel que par exemple celui du paragraphe VII ci-dessous, conçu à partir d'un tableau de données résultant d'une juxtaposition « horizontale » des tableaux relatifs à la période en question.

Une autre situation concrète est celle où on dispose de Q ensembles de variables V_1, V_2, ..., V_Q, mesurées sur le même ensemble E des objets et où, relativement à un même tableau de données $E \times V_q$, on suppose construit un arbre de classification H_q sur E. On se pose alors le problème de résumer $\{H_q / 1 \leqslant q \leqslant Q\}$ au moyen d'un seul arbre H de classification. Mais alors, ici encore, le même type de question que ci-dessus, se pose relativement à l'usage d'un indice tel que celui du paragraphe VII.

Il reste quand même un problème incontournable et que les méthodes classiques ne savent pas résoudre. C'est celui où chaque variable est un caractère hiérarchisé tel une question d'un questionnaire, présentant modalités et sous-modalités de réponses.

Un des initiateurs du problème de la recherche d'un consensus d'une suite $\{P_q / 1 \leqslant q \leqslant Q\}$ de partitions au moyen d'une seule partition qu'il appelle « centrale » est S. Régnier [REGNIER (1965)]. Mais le problème qui nous occupe ici est de fournir ce consensus sous la forme sensiblement plus riche d'un arbre hiérarchique des classifications que — dans notre méthode — on condense aux niveaux où apparaît un nœud significatif [LERMAN (1970a), (1973), (1981), (1983a)].

Nous allons ici directement mentionner le cas où la donnée est une famille $\{H_q / 1 \leqslant q \leqslant Q\}$ de chaînes de partitions ou arbres de classifications. En effet, le cas où la donnée est une suite de partitions $\{P_q / 1 \leqslant q \leqslant Q\}$ sera particulier puisque la donnée d'une partition P est équivalente à celle d'un arbre à trois niveaux définissant la suite des partitions P_0, P et P_1, où P_0 et P_1 sont respectivement la partition discrète à n classes et celle grossière à une seule classe.

Nous résolvons bien ce dernier problème au niveau du paragraphe IV ci-dessus. La solution suggérée ici correspondra à un codage en termes de préordonnance. Pour ce type de codage, une solution a déjà été proposée au paragraphe VI.2. qui précède, mais elle concerne le cas où les partitions sont à deux classes.

La solution que nous proposons ici dans le cas le plus général où la donnée est une suite $\{H_q / 1 \leqslant q \leqslant Q\}$ d'arbres des classifications ou hiérarchies indicées sur l'ensemble E des objets, repose sur une équivalence que nous avons établie dans [LERMAN (1970a)] : un arbre de classification sur E est représentable par une préordonnance sur E d'un type particulier que nous appelons « ultramétrique ».

La préordonnance $\omega(E)$ est ultramétrique si et seulement si, quel que soit l'élément $\{x, y, z\}$ de l'ensemble $P_3(E)$ des parties à trois éléments de E, chacune des trois paires $\{x, y\}$, $\{x, z\}$ et $\{y, z\}$ est à droite de celle des deux autres paires la plus à gauche. Plus précisément, si on appelle ρ une fonction ordinale compatible avec le préordre, qu'on peut directement prendre définie comme il est exprimé au paragraphe VI.1., la condition s'écrit

$$(\forall \{x, y, z\} \in P_3(E)),\ \rho(x, y) \geqslant \min(\rho(x, z),\ \rho(y, z)).$$

On se ramène ainsi au cas d'un ensemble d'objets décrits par des variables « préordonnance », mais avec des accents très particuliers qui sont développés dans [LERMAN-PETER (1985)].

VII. Cas où les variables sont de type divers

Dans les précédents paragraphes, les variables de description de l'ensemble des objets sont toutes d'un seul type et nous avons pris en considération cinq cas différents : le « numérique » (n), le « logique » (l), le « qualitatif nominal » (qn), le « qualitatif ordinal » (qo) et le « qualitatif préordonnance » (pr).

Nous supposons ici que dans le cadre d'un même tableau de données Objets × Variables, on rencontre des variables de types différents. Ainsi, l'ensemble V des variables est supposé pouvoir être décomposé en la somme ensembliste :

$$V = V_n + V_l + V_{qn} + V_{qo} + V_{pr}, \quad (1)$$

où, respectivement, V_n, V_l, V_{qn}, V_{qo} et V_{pr} sont les ensembles de variables numériques, logiques (i.e. attributs), qualitatives nominales, qualitatives ordinales et préordonnances.

L'intérêt de notre indice est de procéder de façon additive, variable par variable, par contributions normalisées. Toutefois, dans les cas numérique et logique où les variables sont — en se plaçant d'un point de vue « géométrique » — de caractère unidimensionnel et « orienté » il y a lieu au préalable de connecter la mesure d'une variable sur un objet à celle des autres variables de même type.

VIII. Cas d'un seul tableau puis d'une juxtaposition de tableaux de contingence

Le cas d'un tableau de contingence I × J est d'un point de vue mathématique assez particulier puisqu'il s'agit d'un tableau parfaitement symétrique où les rôles de I et de J peuvent être interchangés.

Jusqu'à présent, dans notre approche corrélative basée sur la vraisemblance du lien, nous avons été conduits — pour le problème de la classification de I à interpréter, moyennant une représentation géométrique adéquate du tableau des données, chaque i et I comme une variable et chaque j de J, comme un point-objet. Dans ces conditions, l'indice d'association entre i et i′ de I a la nature d'une corrélation dont — avec B. TALLUR — nous avons donné une expression et une interprétation géométrique [LERMAN-TALLUR (1980)], puis une expression et une interprétation ensembliste et statistique [LERMAN (1983b)].

Dans [LERMAN-PETER (1985)], nous proposons une approche — conforme au point de vue développé dans cet article — en interprétant I comme l'ensemble des points objets et J, comme l'ensemble des variables. Un indice de type « cosinus » pour cette interprétation apparaît comme un indice de type « corrélation » pour celle qui a précédé (cf. référence ci-dessus).

C'est à partir des contributions à l'indice cosinus des différents j de J que l'indice est construit. Le plus fécond s'est avéré celui où l'origine est placé au centre de gravité du nuage N (I) (le même que celui de l'analyse des correspondances) et où la normalisation s'effectue au niveau de I × I, I étant muni de son système de poids défini au niveau de N (I).

Il y a une très grande souplesse pour étendre ce type d'indice au cas d'une juxtaposition « horizontale » de tableaux de contingence de la forme :

$$I \times (J^{(1)} \cup J^{(2)} \cup \ldots \cup J^{(\ell)} \cup \ldots \cup J^{(L)})$$

où I (resp. $J^{(\ell)}$, $1 \leqslant \ell \leqslant L$) se trouve défini par l'ensemble des modalités d'une variable partition (i.e. qualitative nominale).

IX. Conclusion

La plupart des indices exprimés ici ont été testés et développés dans le cadre d'un élégant programme SIMOB [LERMAN-PETER (1985)]. Ces indices ont montré toute leur pertinence par rapport à ceux que nous utilisions précédemment et qui nous ont d'ailleurs conduit à notre actuelle démarche.

Ainsi, avec la présente étude, nous enlevons une grande indétermination quant au choix de l'indice de ressemblance entre objets, compatible avec l'algorithme de la vraisemblance du lien. De la sorte, la méthode de classification hiérarchique basée sur la vraisemblance des liens atteint un très grand niveau d'achèvement, puisqu'elle permet, quel que soit la structure du tableau des données et avec une très grande fidélité dans la représentation mathématique, de classificer l'ensemble des variables [LERMAN (1981)] ainsi que l'ensemble des objets. Une vue générale de l'étude du cas spécifique mais important, de la juxtaposition de tables de contingence, est fournie dans [LERMAN (1984), LERMAN-PETER (1985)].

Bibliographie

J.P. BARTHÉLEMY, B. LECLERC et B. MONJARDET (1984). — « Quelques aspects du consensus en classification » in *Data Analysis and Informatics,* North Holland.

J.P. BENZECRI et Collaborateurs (1973) — « L'analyse des Données, Tome I : *La Taxinomie, IB n° 5* », Dunod, Paris.

S. CHAH (1984). — « Agrégation des préordonnances », *Etude F-063,* Centre scientifique IBM de Paris.

E. DIDAY (1982). — « Croisements, ordres et ultramétriques : applications à la recherche de consensus en classification automatique », *Rap. de rech. n° 144, I.N.R.I.A.*

J.C. GOWER (1971). — "A general coefficient of similarity and some of its properties", *Biometrics,* 27, pp. 857-872.

I.C. LERMAN (1970a). — *Les bases de la classification automatique,* Gauthier-Villars, « Collection Programmation », Paris.

I.C. LERMAN (1970b). — « Sur l'analyse des données préalable à une classification automatique. Proposition d'un nouvel indice de similarité », *Rev. Math. et Sc. Hum.,* n° 32, également paru dans *Mathematics in the Archaeological and Historical Sciences,* Eddiburgh University Press, (1971).

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.

I.C. LERMAN (1981). — *Classification et analyse ordinale des données,* Dunod, Paris.

I.C. LERMAN (1983a). — « Sur la signification des classes issues d'une classification automatique », in *Numerical Taxonomy,* Springer.

I.C. LERMAN (1983b). — « Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence », *Rev. Math. et Sc. Hum.,* 21^{e} année, n° 83, p. 5 à 30.

I.C. LERMAN (1984). — « Analyse classificatoire d'une correspondance multiple, typologie et régression », in *Data Analysis and Informatics,* North Holland.

I.C. LERMAN et B. TALLUR (1980). — « Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence », *Rev. de Stat. Appl.,* n° 28, 33, pp. 5-28, Paris.

I.C. LERMAN et PETER Ph. (1985). — « Elaboration et Logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème du concensus en classification ». *Publ. Int. n° 262,* IRISA-Rennes, 72 p.

J.R. MASSE (1978). — « Classes de tableaux équivalents en analyse descriptive des données. Application à l'étude de mesures statiques sur circuits intégrés logiques », Thèse de 3^{e} cycle, Univ. de Rennes I.

A. OCHIAI (1957). — "Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions", *Bull. Jap. Soc. Sci. Fish.,* T. 22, pp. 526-530.

P. PETER (1987). — « Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations assistées par ordinateur », Thèse de l'Université de Rennes I, Informatique, mars 1987.

S. REGNIER (1965). — « Sur quelques aspects mathématiques des problèmes de classification automatique », *I.C.C. Bulletin,* Vol. 4, pp. 175-191.

F.J. ROHLF (1982). — "Consensus indices for comparing classifications", *Math. Biosc.,* 59, pp. 131-144.

P.H.A. SNEATH and R. SOKAL (1972). — *Numerical Taxonomy,* Freeman, Sans Francisco and London.