Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens).

- DEUXIÈME PARTIE -

I.C. LERMAN* PH. PETER† et H. LEREDDE‡

Table des matières

* IRISA, UNIVERSITÉ DE RENNES 1, campus de Beaulieu, avenue du Général Leclerc, 35042 Rennes cedex, tél. (33)99.84.71.00, E-mail : lerman@irisa.fr, fax (33)99.38.38.32

† IRESTE, UNIVERSITÉ DE NANTES, La Chantrerie, CP 3003, 44087 Nantes cedex 03, tél. (33)40.68.30.00, E-mail : ppeter@ireste.fr

‡ UNIVERSITÉ PARIS NORD, avenue Jean Baptiste Clément, 93430 Villetaneuse, tél. (33)1.49.40.30.00

Important

La première partie de cet article est parue dans le numéro de décembre 1993.

Les numéros des paragraphes et des figures recommencent néanmoins à 1.

Les références bibliographiques présentées à la fin du texte concernent *les deux parties.*

1 Les différentes structures de données traitées dans CHAVL - Indices calculés - Extensions

1.1 Introduction

Référons-nous à la structure générale d'un tableau descriptif de données (voir à la fin du paragraphe 2 de la première partie , dans laquelle s'inscrit la figure 2). Rappelons que d'un point de vue logique nous distinguons de façon fondamentale la description d'un ensemble O d'objets élémentaires (ou individus) de celle d'un ensemble C de catégories (on dit encore classes, concepts ou même modalités). Dans le cadre de CHAVL, on ne suppose pas l'existence de données manquantes; toutefois, il s'agit d'une situation qui peut être traitée avec un minimum de perte d'information dans le contexte de la méthodologie proposée, nous y reviendrons.

D'autre part, nous supposons qu'on se trouve dans le cas (ou qu'on s'y ramène) où toutes les variables descriptives composant A, sont d'un même type quant à la nature de l'échelle sous-jacente (e.g. toutes les variables sont quantitatives ou bien toutes les variables sont qualitatives ordinales...).

Interprétant une variable descriptive d'un ensemble O d'objets, comme une relation sur ce dernier, le type général de la variable est défini par l'arité de la relation. Cependant, pour une même arité, il y a différents sous types. Ainsi, les variables qualitatives nominales, et ordinales définissent respectivement des relations binaires sur l'ensemble décrit. Mais, la relation sur O définie par une variable qualitative nominale n'est pas de même nature que celle, définie par une variable qualitative ordinale; nous y reviendrons.

Nous allons commencer par considérer le cas où la description concerne un ensemble O d'objets élémentaires, puis celui, concernant un ensemble C. Pour chacun des deux cas, il peut s'agir de la classification de l'ensemble A des variables (on dit encore attributs) de description ou de l'ensemble E des entités descriptives [$E = O$ ou C (voir première partie, figure 3)].

1.2 Cas où la description concerne un ensemble O d'objets élémentaires

1.2.1 Structures de données prises en compte dans CHAVL pour la classification de A

Cette prise en compte se fait au niveau de l'étape ASVAR (ASsociation entre VARiables). Le type d'un tableau de données est celui, supposé commun (cf. ci-dessus), des différentes variables descriptives composant A. Ce dernier type peut-être:

* Quantitatif :
Une même variable, faisant partie de A, qu'on notera v, est initialement définie par une application de l'ensemble O des objets dans un sous-ensemble de l'ensemble $\mathbb{R}$ des nombres réels :

$$\begin{aligned} v \quad &: \quad O \longrightarrow \mathbb{R} \\ &\quad\ x \longrightarrow v(x) \end{aligned} \tag{72}$$

attachant à chaque objet x, une valeur numérique $v(x)$. La variable est ici considérée comme définissant sur O une relation unaire valuée ; qu'on représente donc par une valuation sur O.

* Logique de présence-absence :
On notera ici a l'attribut logique de présence-absence qu'on dit également booléen. C'est une application de O dans l'ensemble $\{0,1\}$ des deux codes logiques où 0 indique l'absence et où 1 indique la présence :

$$\begin{aligned} a \quad &: \quad O \longrightarrow \{0,1\} \\ &\quad\ x \longrightarrow a(x) \end{aligned} \tag{73}$$

a induit sur O une relation unaire qui est représentée par le sous-ensemble

$$O(a) \ = \ a^{-1}(1) \tag{74}$$

des objets de O où l'attribut est présent (on dit encore "à vrai").

* Qualitatif nominal :
Désignons par π la variable qualitative nominale. Il s'agit formellement d'une application de O dans un ensemble de codes qui définit l'échelle nominale. Notons par

$$E_{nom} = \{1, 2, \ldots, i, \ldots, h\} \tag{75}$$

une telle échelle, où on ne suppose aucune structure sur l'ensemble de ses éléments ou valeurs :

$$\begin{aligned} \pi \quad &: \quad O \longrightarrow E_{nom} \\ &\quad\ x \longrightarrow \pi(x) \end{aligned} \tag{76}$$

La variable π induit une partition sur l'ensemble O des objets que nous notons :

$$\pi(O) \ = \ \{E_i / 1 \leq i \leq h\}$$

où

$$E_i \ = \ \pi^{-1}(i) \tag{77}$$

est le sous-ensemble des objets pour lesquels la valeur de π est i, $1 \leq i \leq h$.

Une telle variable qui définit une relation d'équivalence sur O est représentée au niveau de l'ensemble que nous notons $O^{(2)}$ des parties à 2 éléments de O. Nous la représentons en représentant la partition $\pi(O)$ par l'ensemble des paires qu'elle réunit. Très précisément, il s'agit de

$$R(\pi) = \Sigma\{E_i^{(2)}/1 \leq i \leq h\} \tag{78}$$

où la somme est ensembliste. En désignant par m_i le cardinal de E_i, on a

$$n = card(O) = \sum_{1 \leq i \leq h} m_i$$

et

$$card[R(\pi)] = \sum_{1 \leq i \leq h} [m_i(m_i - 1)/2] \tag{79}$$

Signalons pour terminer qu'il est tout à fait équivalent pour le but poursuivi de représenter π par le sous-ensemble des paires $S(\pi)$ que la partition $\pi(O)$ sépare [Lerman 1981].

* Qualitatif ordinal :

ω désigne la variable qualitative ordinale. Le formalisme général est identique à ci-dessus. Mais ici, l'échelle des valeurs que nous notons

$$E_{ord} = \{1, 2, \ldots, i, \ldots, h\} \tag{80}$$

est munie d'une structure d'ordre total ; en d'autres termes, on suppose qu'on a, pour les modalités codées de ω

$$1 < 2 < \ldots < i < \ldots < h \tag{81}$$

La variable ω définie par l'application :

$$\begin{aligned} \omega \quad : \quad & 0 \longrightarrow E_{ord} \\ & x \longrightarrow \omega(x) \end{aligned} \tag{82}$$

induit sur O un préordre total dont la suite ordonnée des classes peut être notée

$$\{E_i/1 \leq i \leq h\}$$

où

$$E_i = \omega^{-1}(i),\ 1 \leq i \leq h$$

et où donc,

$$E_1 < E_2 < \ldots < E_i < \ldots < E_h \tag{83}$$

Nous représentons ce préordre total de façon ensembliste au niveau de $O \times O$, au moyen de

$$R(\omega) = \Sigma\{E_i \times E_{i'}/1 \leq i < i' \leq h\} \tag{84}$$

où la somme est ensembliste [Lerman 1973, voir dans 1981,1983]. D'autres représentations sont considérées dans [Giakoumakis et Monjardet 1987].

En désignant par m_i le cardinal de E_i, $1 \leq i \leq h$; on a :

$$card[R(\omega)] = \Sigma\{m_i\ m_{i'}/\ 1 \leq i < i' \leq h\} \tag{85}$$

1.2.2 Extensions

L'extension d'importance -non prise en compte dans CHAVL- concerne la structure "préordonnance" des échelles de description sous-jacentes aux variables qualitatives. Si

$$E_{pre} = \{1, 2, \ldots, i, \ldots, h\} \tag{86}$$

est l'ensemble des valeurs codées d'une variable qualitative w, une préordonnance sur E_{pre} est un préordre total sur tout ou partie de l'ensemble $E_{pre} \times E_{pre}$. Généralement, il s'agit soit de l'ensemble de tous les couples, soit de l'ensemble

$$H = \{(i, i')/1 \leq i \leq i' \leq h\} \tag{87}$$

Ce préordre total traduit de façon ordinale les ressemblances perçues entre les modalités de w, dite variable préordonnance. Le codage en termes de préordonnance d'une variable qualitative permet de mieux tenir compte de la connaissance de l'expert des données. Ce qui conduit à une organisation classificatoire plus cohérente dans ses nuances, aussi bien de l'ensemble des variables descriptives que de l'ensemble des entités décrites. Signalons ici à titre d'illustration le cas de données résultant d'une enquête 1989 de l'association AGORAmétrie, mises à notre disposition par J.P. Pagès. Une partie du questionnaire comporte 19 variables ayant la même expression ; mais dont chacune concerne un homme politique (e.g. Delors, Léotard, ...) Une même question qui définit une variable qualitative se présente sous la forme :

"Souhaitez-vous voir jouer un rôle important dans l'avenir, à cet homme politique?

Réponse	Code
Oui	1
Non	2
Sans réponse	3

Deux traitements classificatoires par AVL de l'ensemble des 19 variables ont été considérés dans [Ouali-Allah 1991]. Le premier correspond à interpréter chacune des variables qualitatives comme nominale. Pour le second traitement, chacune des variables est interprétée comme qualitative préordonnance. Dans ce dernier cas, le préordre total traduisant les ressemblances entre modalités est établi sur H [cf. (87)], où $h = 3$, de la manière suivante :

$$12 < 23 < 13 < 11 \sim 22 \sim 33$$

La prise en compte d'une telle structure ordinale des similarités entre modalités a permis la mise en évidence de tendances et sous tendances de comportement correspondantes à une facette de l'interprétation nette et claire.

Nous avons déjà mentionné ci-dessus que le codage en termes de préordonnances des variables qualitatives permet, en s'aidant si nécessaire de l'appréhension du spécialiste des données, d'enrichir la structure des échelles sous jacentes. Il en résulte par ailleurs un traitement homogène de l'ensemble des variables qualitatives. C'est précisément ce qui est fait dans le programme AVARE (Association entre VAriables RElationnelles) mis au point par M. Ouali-Allah et que nous mentionnerons à nouveau ci-dessous.

Dans notre cadre ici, où rappelons le, la description concerne un ensemble O d'objets élémentaires, la structure la plus générale de la donnée pouvant être traitée par AVL est fournie par ce que l'on appelle, un système relationnel de Tarski (1954). Ce dernier se met sous la forme

$$T = < O ; R_1, R_2, \ldots, R_j, \ldots, R_p > \tag{88}$$

où $R_1, R_2, \ldots, R_j, \ldots, R_{p-1}$ et R_p sont p relations définies sur l'ensemble O des objets élémentaires. Dans notre cas, la relation R_j se trouve définie par le j-ème attribut (on dit encore variable) de description a^j , $1 \leq j \leq p$. Dans ces conditions, on suppose ou on se ramène au cas où les différentes relations sont d'un même type combinatoire et donc, de même arité. Ainsi, par exemple, dans les cas classiques où les a^j sont des variables qualitatives nominales ou, respectivement, ordinales, les relations R_j sont des partitions, ou, respectivement, des préordres totaux sur O, $1 \leq j \leq p$. Dans ce cas, l'arité q commune des relations est égale à 2. En fait, dans l'analyse des données qualitatives ou quantitatives que nous avons eues à embrasser, il a été suffisant de considérer $q = 1, 2$ ou 4. Une même relation R_j est représentée par une partie structurée de O^q ; mais cela n'exclut pas qu'elle puisse être valuée. Le cas envisagé dans l'exemple est celui non valué pour $q = 1$ (cf. première partie figure 4).

Le fait que la méthode soit fondée sur la notion de similarité lui confère une grande souplesse en cas de données manquantes. S'agissant de la comparaison des variables de description et donc, des relations induites, la comparaison de deux relations R_j et R_k $(1 \leq j < k \leq p)$ s'appuiera sur le sous ensemble des objets O_{jk} où chacune des deux relations est complète. D'autre part, quelle que soit l'arité commune des deux relations, la forme d'un indice tel que (21) (première partie) permet de dégager l'influence du cardinal de O_{jk}, qu'on peut alors aisément neutraliser, avant la réduction globale des similarités [cf. (62) première partie § 3.5] [Lerman 1992$_b$].

Un autre type d'extension concerne la forme même de l'hypothèse d'absence de liaison. À cet égard trois formes fondamentales ont été mises en évidence [Lerman 1981, 1992$_a$] et étudiées (Daudé 1992, Lerman 1984, 1992$_b$, Ouali-Allah 1991].

1.2.3 Cœfficients d'association entre variables descriptives calculés dans CHAVL

Dans l'étape ASVAR ci-dessus mentionnée, on établit la demi matrice inférieure des cœfficients d'association entre variables descriptives, centrés et réduits, de même type que (21) ci-dessus. Si j est l'indice courant d'une ligne de cette matrice et k, celui d'une colonne, cette demi matrice prend, après réduction globale [cf. (62) (première partie)], la forme suivante :

$$\{Q_g(a^j, a^k)/1 \leq k < j \leq p\} \tag{89}$$

La diagonale de la demi matrice $\{(j,j)/1 \leq j \leq p\}$ peut servir à contenir des éléments de calcul permettant précisément la normalisation statistique.

Nous allons maintenant expliciter ce que devient $Q_l(a^j, a^k)$ [cf. (21 (première partie)] dans chacun des cas de figure considérés au paragraphe 1.2.1.

* Les variables sont quantitatives-numériques :
Désignons ici par (v^j, v^k) le couple de variables descriptives à associer. x_i^j (resp. x_i^k) est la mesure de la variable v^j (resp. v^k) sur le i-ème objet, $1 \leq i \leq n$ x^j (resp. x^k) est la moyenne de la variable v^j (resp. v^k) sur l'ensemble des objets. Dans ces conditions, le cœfficient de corrélation empirique

$$R(v^j, v^k) = \frac{\sum_{1 \leq i \leq n} (x_i^j - x^j)(x_i^k - x^k)}{\left\{ \left[\sum_{1 \leq i \leq n} (x_i^j - x^j)^2 \right] \left[\sum_{1 \leq i' \leq n} (x_{i'}^k - x^k) \right] \right\}^{1/2}} \tag{90}$$

est, au facteur $1/\sqrt{n-1}$ près, le cœfficient $Q_l(v^j, v^k)$ pour l'une des formes de l'h.a.l. [Lerman 1981, 1992$_a$].

C'est la matrice des cœfficients de corrélation $R(v^j, v^k), 1 \leq k < j \leq p$, qui est calculée avant d'être soumise à la "réduction globale des similarités" [cf. (62) (première partie)].

* Les variables sont des attributs booléens de présence-absence :
Ce cas a été largement considéré ci-dessus. Comme nous l'avons déjà mentionné dans un autre contexte [cf. première partie) § 3.3], le cœfficient d'association entre deux attributs booléens a^j et a^k, $1 \leq k < j \leq p$, est normalisé par rapport à la forme Poissonnienne de l'h.a.l. [Lerman 1981, 1992$_a$]. Rappelons ici une fois de plus l'expression du cœfficient [cf (57) (première partie)] :

$$Q_l(a^j, a^k) = \frac{n(a^j \wedge a^k) - [n(a^j)n(a^k)/n]}{\sqrt{n(a^j)n(a^k)/n}} \tag{91}$$

avec des notations que l'on comprend :

$$n(a^j) = card[O(a^j)], n(a^k) = card[O(a^k)]$$

et

$$n(a^j \wedge a^k) = card[O(a^j) \cap O(a^k)]$$

* Les variables sont qualitatives nominales :
Pour ne pas traîner des indices supérieurs, notons π et χ les deux partitions sur l'ensemble O des objets, respectivement induites par deux variables qualitatives nominales α^j et α^k, $1 \leq k < j \leq p$. On écrira

$$\pi = \{E_i/1 \leq i \leq h\} \text{ et } \chi = \{F_j/1 \leq j \leq k\}, \tag{92}$$

où E_i (resp. F_j) est la i-ème (resp. j-ème) classe de la partition π (resp. χ), qu'on supposera -sans restreindre la généralité- qu'elle est en classes étiquetées, $1 \leq i \leq k$ (resp. $1 \leq j \leq k$).

$$\begin{aligned} &t(\pi) = (m_i/1 \leq i \leq h) \\ &[\text{resp. } t(\chi) = (n_j/1 \leq j \leq k)] \end{aligned} \tag{93}$$

où $m_i = card(E_i)$ [resp. $n_j = card(F_j)$], indique le type de la partition π (resp. χ).

En se référant à (78) ci-dessus, les représentations respectives de π et de χ comme sous ensembles de $O^{\{2\}}$, sont définies par :

$$\begin{aligned} R(\pi) &= \Sigma\{E_i^{\{2\}}/1 \leq i \leq h\} \\ \text{et } R(\chi) &= \Sigma\{F_j^{\{2\}}/1 \leq j \leq k\} \end{aligned} \tag{94}$$

On a, en désignant par $\pi \wedge \chi$, la partition résultant du croisement des deux partitions π et χ :

$$\begin{aligned} s(\pi,\chi) &= card[R(\pi) \cap R(\chi)] = card[R(\pi \wedge \chi)] \\ &= card[\Sigma\{(E_i \cap F_j)^{\{2\}}/1 \leq i \leq h, 1 \leq j \leq k\}] \\ &= \Sigma\{n_{ij}(n_{ij}-1)/2/1 \leq i \leq h, 1 \leq j \leq k\}, \end{aligned} \tag{95}$$

où nous avons noté $\pi \wedge \chi$ le croisement des deux partitions π et χ et n_{ij} le cardinal de $E_i \cap F_j$, $1 \leq i \leq h, 1 \leq j \leq k$. Ce cardinal est un des composants du tableau de contingence associé à $\pi \wedge \chi$.

La forme la plus classique de l'hypothèse d'absence de liaison (h.a.l.) consiste à associer à la partition π (resp. χ) une partition aléatoire π^* (resp. χ^*), élément dans l'ensemble -muni d'une probabilité uniforme- de toutes les partitions sur O de type $t(\pi)$ (resp. $t(\chi)$). D'autre part, π^* et χ^* sont indépendantes. Dans ces conditions, on obtient les expressions suivantes pour la moyenne et la variance de l'indice brut aléatoire $s(\pi^*,\chi^*)$ [Lerman 1973 dans 1981] :

$$E[s(\pi^*,\chi^*] = \lambda\mu \text{ et } var[s(\pi^*,\chi^*] = \lambda\mu + \rho\sigma + \theta\zeta - \lambda^2\mu^2 \tag{96}$$

où

$$\begin{aligned} \lambda &= \sum\{m_i(m_i-1)/\sqrt{2n(n-1)}/1 \leq i \leq h\}, \\ \rho &= \sum\{m_i(m_i-1)(m_i-2)/\sqrt{n(n-1)(n-2)}/1 \leq i \leq h\}, \\ \theta &= \left\{\left[\sum m_i(m_i-1)\right]^2 - 2\sum_i m_i(m_i-1)(2m_i-3)\right\} \\ &\quad /2\sqrt{n(n-1)(n-2)(n-3)} \end{aligned}$$

et où les expressions de μ, σ et ζ ont respectivement la même forme que λ, ρ et θ ; les m_i de $t(\pi)$, étant remplacés par les n_j de $t(\chi), 1 \leq j \leq k$.

L'indice qui nous intéresse est, comme toujours, celui centré et réduit :

$$Q_l(\alpha^j,\alpha^k) = \frac{s(\pi,\chi) - \lambda\mu}{\sqrt{\lambda\mu + \rho\sigma + \theta\zeta - \lambda^2\mu^2}} \tag{97}$$

* Les variables sont qualitatives ordinales :

Nous notons ici ω et $\bar{\omega}$ les deux préordres totaux sur l'ensemble O des objets, respectivement

induits par deux variables qualitatives ordinales β^j et β^k, $1 \leq k < j \leq p$. On notera les classes de ω (resp. $\bar{\omega}$) de la même façon que les classes de π (resp. χ) [cf. (92)].

$$t(\omega) = (m_i / 1 \leq i \leq h) [resp\ t(\bar{\omega}) = (n_j / 1 \leq j \leq k)] \tag{98}$$

désigne la composition du préordre total ω (resp. $\bar{\omega}$).

En se référant aux expressions (84) et (85) ci-dessus, on comprend la nature des ensembles de représentation $R(\omega)$ et $R(\bar{\omega})$. L'indice brut prend alors la forme

$$\begin{aligned} s(\omega, \bar{\omega}) &= card[R(\omega) \cap R(\bar{\omega})] \\ &= card(\Sigma\{(E_i \cap F_j) \times (E_{i'} \cap F_{j'}) / 1 \leq i < i' \leq h, \\ &\quad 1 \leq j < j' \leq k\}), \end{aligned} \tag{99}$$

où nous notons $i, i', \ldots$ (resp. $j, j', \ldots$) des étiquettes de classes de ω (resp. $\bar{\omega}$). On a [Lerman 1973 dans 1981,1983], sous la forme la plus combinatoire de l'hypothèse d'absence de liaison :

$$\begin{aligned} E[S(\omega^*, \bar{\omega}^*] &= \lambda\mu \text{ et } var[S(\omega^*, \bar{\omega}^*)] = \lambda\mu + \rho_{cc}\sigma_{cc} + \rho_{ff}\sigma_{ff} \\ &\quad + 2\rho_{cf}\sigma_{cf} + \theta\zeta - \lambda^2\mu^2. \end{aligned} \tag{100}$$

où

$$\lambda = \frac{1}{\sqrt{n(n-1)}} \Sigma\{m_i, m_{i'} / 1 \leq i < i' \leq h\},$$

$$\rho_{cc} = \frac{1}{\sqrt{n(n-1)(n_2)}} \Sigma\{m_i\ m_{i'} [m_{c(i)} - 1]/2 \leq i \leq h\},$$

$$\rho_{ff} = \frac{1}{\sqrt{n(n-1)(n-2)}} \Sigma\{m_i\ m_{f(i)} [m_{f(i)} - 1] / 1 \leq i \leq (h-1)\},$$

$$\rho_{cf} = \frac{1}{\sqrt{n(n-1)(n-2)}} \Sigma\{m_i\ m_{c(i)}\ m_{f(i)} / 2 \leq i \leq (h-1)\}$$

$$\begin{aligned} \theta &= \frac{1}{\sqrt{n(n-1)(n-2)(n-3)}} \times \\ &\quad \Sigma\{m_i\ m_{i'} [\Sigma\{m_p\ m_{p'} / 1 \leq p < p' \leq h\} + m_i + m_{i'} - 2n + 1] / 1 \leq i < i' \leq h\} \end{aligned}$$

où on note $m_{c(i)} = \Sigma\{m_{i'} / i' < i\}$ et $m_{f(i} = \Sigma\{m_i / i' > i\}$.

D'autre part, les expressions de $\mu, \sigma_{cc}, \sigma_{ff}, \sigma_{cf}$ et ζ sont respectivement de même forme que celles $\lambda, \rho_{cc}, \rho_{ff}, \rho_{cf}$ et θ ; si les premières sont relatives à la composition $t(\omega) = (m_1, m_2, \ldots, m_h)$, les secondes sont relatives à la composition $t(\bar{\omega}) = (n_1, n_2, \ldots, n_k)$.

L'indice centré et réduit se met sous la forme :

$$Q_I(\beta^j, \beta^k) = \frac{s(\omega, \bar{\omega}) - \lambda\ \mu}{\sqrt{\lambda\mu\ +\ \rho_{cc}\sigma_{cc}\ +\ \rho_{ff}\sigma_{ff}\ +\ 2\rho_{cf}\sigma_{cf}\ +\ \theta\zeta\ -\ \lambda^2\mu^2}} \tag{101}$$

1.2.4 Extensions

L'extension la plus efficace, qui a été largement expérimentée, concerne l'élaboration de la matrice des cœfficients d'association entre variables qualitatives préordonnances [cf. § 1.2.2]. Plus précisément, on suppose que l'échelle des valeurs d'une variable qualitative se trouve munie d'une structure de préordonnance. Cette dernière doit ensuite être codée ou représentée. Dans ce contexte, les représentations précédentes des variables qualitatives les plus usuelles, peuvent paraître d'un point de vue analytique comme particulières ; cependant, elles ont une justification ensembliste propre.

Le programme AVARE, ci-dessus évoqué, permet la construction de la matrice des cœfficients d'association centrés et réduits conformément à AVL, entre variables qualitatives de toutes sortes, interprétées en termes de préordonnances. La portée de ce programme est très générale en analyse des données qualitatives. Il est écrit conformément aux normes Modulad et devrait pouvoir rejoindre la bibliothèque Modulad. Alors que AVARE concerne la forme la plus combinatoire de l'hypothèse d'absence de liaison, le programme AVRP (Association entre Variables Relationnelles sous le modèle Poissonnien), mis au point par F. Daudé dans le cadre de sa thèse [Daudé 1992], mais non conforme aux normes Modulad, est dévolu au modèle poissonnien de l'h.a.l.

Considérons maintenant la structure descriptive la plus générale d'un ensemble O d'objets élémentaires, telle qu'elle se trouve définie par un système de Tarski [cf. (88)) ci-dessus]. Nous sommes ici concernés par la classification AVL de l'ensemble des relations qu'on suppose de même arité q. Dans [Lerman 1992$_b$] nous montrons comment établir le cœfficient d'association centré et réduit entre deux telles relations.

1.2.5 Structures de données prises en compte dans CHAVL pour la classification de O et indices de similarité associés

En plus des structures de données considérées au paragraphe IV 2.1, le type "préordonnance" des variables qualitatives de description est également pris en compte (cf. § 1.2.2). Mais ici, nécessairement, la préordonnance est un préordre total sur H [cf. (86) et (87)]. Ce préordre total est codé par la notion de "rang moyen" pour laquelle la somme des rangs est constante quelle que soit la finesse du préordre. Ainsi, si $L = card(H)$, cette somme des rangs est égale à $L(L+1)/2$; nous y reviendrons ci-dessous.

Le calcul de la table des indices de similarité entre objets (chargement de la demi matrice inférieure) :

$$\{Q_g(i,i')/1 \leq i' \leq i \leq n\} \tag{102}$$

[cf. (69) (première partie)], est calculé dans l'étape SIMOB qui résulte de [Lerman & Peter 1985].

Le diagramme général du calcul est exprimé au paragraphe III 6. On se rend compte que, pour un type donné de variable descriptive, il faut et il suffit de définir de façon adéquate, la contribution brute $s_j(o_i, o_{i'})$ de la j-ième variable descriptive a^j à la comparaison de deux objets o_i et $o_{i'}$, $1 \leq j \leq p, 1 \leq i < i' \leq n$.

Dans ces conditions, nous allons donner, pour chaque type concerné de variable, l'expression de l'indice brut $s_j(o_i, o_{i'})$.

* Type quantitatif-numérique :
Nous notons le tableau des données sous la forme

$$\{x_i^j \ / 1 \leq i \leq n,\ 1 \leq j \leq p\} \tag{103}$$

conformément à l'expression (90) ci-dessus.

Au tableau (103) on associera celui

$$\{\xi_i^j \ / \ 1 \leq i \leq n,\ 1 \leq j \leq p\} \tag{104}$$

où

$$\xi_i^j \ = \frac{x_i^j}{\sqrt{\sum_{1 \leq j' \leq p} (x_i^j)^2}} \tag{105}$$

On pose alors pour l'indice brut $s_j(i,i') = s_j(o_i, o_{i'})$:

$$s_j(i,i') = \frac{1}{p} - \frac{1}{2}(\xi_i^j \ - \ \xi_{i'}^j)^2 \tag{106}$$

Considérons ici la représentation géométrique classique de l'ensemble O des objets par un nuage de points dans l'espace $\mathbb{R}^p$, où l'objet o_i est représenté par le point i de coordonnées $(x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^p)$. Dans ces conditions, on a les propriétés suivantes :

a) $\Sigma\{s_j(i,i')/1 \leq j \leq p\} = Cos(\widehat{iOi'})$, où O désigne l'origine ;

b) $s_j(i,i')$ est maximal si i et i' sont homothétiques par rapport à l'origine ; de plus, cette valeur maximale est la même quel que soit $j = 1, 2, \ldots, p$.

On reprendra ici la forme de l'indice $S_j(o_i, o_{i'})$ [cf. (66)] qui définit la contribution normalisée de la j-ième variable v^j à la comparaison des deux objets o_i et $o_{i'}$. Ainsi, dans l'indice de forme (67) (première partie) :

$$S(o_i, o_{i'}) \ = \sum_{1 \leq j \leq p} S_j(o_i, o_{i'}), \tag{107}$$

qui précède la réduction globale des similarités [cf. (69) (première partie)], chacune des variables a le même pouvoir discriminant.

Après ces propriétés générales, précisons ici l'expression explicite de

$$S_j(o_i, o_{i'}) = \frac{s_j(i,i') - \mu_j}{\sigma_j}, \tag{108}$$

où il reste à fournir μ_j et σ_j^2 qui sont, respectivement, la moyenne et la variance de $s_j(i,i')$ [cf. (106)] sur $I \times I$, où $I = \{1, 2, \ldots, i, \ldots, n\}$ indexe l'ensemble des objets.

Relativement au tableau (104), associons à la j-ème colonne la somme $T_r(j)$ des puissances r-èmes de ses composantes :

$$T_r(j) \ = \sum_{1 \leq i \leq n} (\xi_i^j)^r \tag{109}$$

En considérant $r = 1, 2, 3$ et 4 ; on a :

$$\mu_j = \frac{1}{p} - \frac{1}{n} T_2(j) + \frac{1}{n^2} [T_1(j)]^2 \tag{110}$$

et

$$\sigma_j^2 = \frac{1}{2n} T_4(j) - \frac{2}{n2} T_3(j)T_1(j) + \frac{1}{2n^2} [T_2(j)]^2 + \frac{2}{n^3} T_2(j)[T_1(j)]^2 - \frac{1}{n^4} [T_1(j)]^4 \tag{111}$$

* Type attribut logique de présence-absence :
D'un point de vue analytique, ce type de données qui a illustré l'exemple (cf. §III.6) sera traité de façon identique au type précédent. Il peut y avoir à cela quelques éléments de justification théorique que nous ne pouvons développer ici.

Le tableau (103) est ici remplacé par le tableau des valeurs zéros ou uns :

$$\{\xi_i^j / 1 \leq i \leq n,\ 1 \leq j \leq p\} \tag{112}$$

Le tableau (104) est alors remplacé par celui noté

$$\{\eta_i^j / 1 \leq i \leq n,\ 1 \leq j \leq p\} \tag{113}$$

Ces deux tableaux ont déjà été introduits au-dessus de la formule (64) du paragraphe § 3.6 de la première partie qu'il s'agit de suivre en relation avec le traitement ci-dessus du cas quantitatif.

* Type qualitatif nominal :
Désignons par π^j l'une des variables qualitatives de description, $1 \leq j \leq p$. On peut, de façon naturelle, poser

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

En d'autres termes, s_j correspond à la fonction indicatrice de l'ensemble $Cr(\pi^j)$ des couples d'objets réunis par la partition induite sur l'ensemble O des objets par π^j [cf. (77) § 1.2.1]. On a :

$$Cr(\pi^j) = \Sigma\{E_i \times E_i / 1 \leq i \leq h\} \tag{114}$$

Conformément aux notations déjà introduites [cf. (79) [cf. (79) § 1.2.1], désignons par $p_i = (m_i/n)$ la proportion d'objets qui se trouvent dans la i-ième classe E_i de la partition $\pi^j(O)$, $1 \leq i \leq h$. La proportion de couples d'objets réunis par $\pi^j(O)$ est

$$\sum_{1 \leq i \leq h} p_i^2 ;$$

de sorte que, la moyenne μ_j et la variance σ_j^2 de $s_j(i, i')$ sur $I \times I$, sont respectivement déterminées par

$$\mu_j = \sum_{1 \le i \le h} p_i^2$$

et

$$\sigma_j^2 = \left(\sum_{1 \le i \le h} p_i^2 \right) \left(1 - \sum_{1 \le i \le h} p_i^2 \right) \tag{115}$$

* Type qualitatif ordinal :

Ici w^j désigne une des variables qualitatives ordinales de description. Relativement aux notations du paragraphe §1.2.1 ci-dessus, on se trouve conduits [cf. Lerman 1981, chap. 2] à prendre comme contribution brute de w^j à la comparaison de deux objets o_i et $o_{i'}$, la quantité :

$$s_j(i, i') = (h_j - 1) - |w^j(o_i) - w^j(o_{i'})|, \tag{116}$$

où nous avons noté h_j le nombre de modalités de la variable w^j, $1 \le j \le p$.

Le calcul donne pour la moyenne μ_j et la variance σ_j^2 de $s_j(i, i')$ [défini par (116)] sur $I \times I$:

$$\mu_j = h_j - 1 - \frac{2}{n^2} \sum_{1 \le i < i' \le h} m_i \, m_{i'} (i' - i) \tag{117}$$

et

$$\sigma_j^2 = \frac{2}{n^2} \sum_{1 \le i < i' \le h} m_i \, m_{i'} (i' - i)^2 - \frac{4}{n^4} \left[\sum_{1 \le i < i' \le h} m_i \, m_{i'} (i' - i) \right]^2 \tag{118}$$

* Type qualitatif préordonnance :

w^j indiquera une des variables qualitatives préordonnances de description. Conformément aux notations du paragraphe 1.2.2, nous supposons établi un préordre total $w(H)$ sur l'ensemble H [cf. (87)]. Ce préordre est établi de façon à respecter les ressemblances perçues entre les modalités de la variable. Ainsi, les couples de la forme (i, i) ont un rang maximal. Pour être tout à fait explicite, soit l'exemple illustratif suivant issu d'un cas réel d'une variable préordonnance à 9 modalités ($h = 9$) :

$$\begin{aligned} &15 \sim 16 \sim 17 \sim 18 \sim 19 \sim 25 \sim 26 \sim 27 \sim 28 \sim 29 \sim 36 \sim 37 \sim 38 \sim 39 \\ &\sim 46 \sim 48 \sim 49 \sim 57 \sim 58 \sim 59 \sim 67 \sim 68 \sim 69 \sim 78 \sim 79 \sim 89 \\ &< 13 \sim 23 \sim 35 \sim 45 < 14 \sim 24 \sim 34 \sim 56 < 12 \\ &< 11 \sim 22 \sim 33 \sim 44 \sim 55 \sim 66 \sim 77 \sim 88 \sim 99, \end{aligned} \tag{119}$$

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

A chaque élément de l'ensemble totalement préordonné H, on associe un "rang". Pour définir précisément la fonction ordinale "rang" désignons par $(l_1, l_2, \ldots, l_k)$ la suite des cardinaux de la suite ordonnée des classes du préordre total. Dans l'exemple ci-dessus $k = 5$ et

$$(l_1, l_2, l_3, l_4, l_5) = (27, 4, 4, 1, 9).$$

Le rang d'un élément appartenant à la j-ème classe, est égal à

$$\sum_{1 \leq i \leq (j-1)} l_i + (l_j + 1)/2 \tag{120}$$

Ainsi, dans l'exemple précédent, le rang d'un élément de la troisième classe est $(27 + 4 + 5/2) = 33,5$. La structure descriptive sera donc fondée sur le tableau des rangs ainsi calculés que nous notons

$$\{r(g, g')/(g, g') \in H\} \tag{121}$$

Etant donnés, maintenant deux objets o_i et o_i', pour lesquels, respectivement, $w^j(o_i) = g$ et $w^j(o_{i'}) = g'$, on posera pour l'indice brut de comparaison entre o_i et $o_{i'}$ relativement à la variable w^j:

$$s_j(i, i') = r[w^j(0_i), w^j(o_{i'}] = r(g, g') \tag{122}$$

$\{E_g/1 \leq g \leq h\}$ indique comme d'habitude la partition de l'ensemble O des objets induite par la variable qualitative w^j. Compte tenu des notations adoptées (cf. ci-dessus dans "type qualitatif nominal"), introduisons les proportions suivantes :

$$\rho_g = p_g^2 \text{ et } \sigma_{fg} = 2p_g p_h \tag{123}$$

$p_g (resp. \sigma_{fg})$ est la proportion de couples d'objets réunis (resp. séparés) dans la classe E_g (dans la paire de classes $\{E_f, E_g\}$), $1 \leq f \leq g \leq h$.

$$\rho = \sum_{1 \leq g \leq h} \rho_g \text{ et } \sigma = \sum_{1 \leq f < g \leq h} \sigma_{fg} \tag{124}$$

sont respectivement la proportion de couples d'objets réunis et celle des couples d'objets séparés par la partition ci-dessus mentionnée.

En notant σ_{gg} pour ρ_g, l'indice brut $s_j(i, i')$ [cf. (122)] centré et réduit par rapport à l'ensemble de tous les couples d'objets (indexé par $I \times I$) peut se mettre sous la forme [Lerman et Peter 1985] :

$$S_j(o_i, o_{i'}) = \frac{\sum\{\sigma_{ef}[r(g, g') - r(e, f)]/1 \leq e \leq f \leq h\}}{\left\{\sum_{d \leq e} \sigma_{de} \left[\sum_{f \leq g} \sigma_{fg}[r(d, e) - r(f, g)]\right]^2\right\}^{1/2}} \tag{125}$$

Signalons avant de terminer ce paragraphe 1.2.5 qui concerne la classification d'un ensemble O d'objets , que la construction de l'indice de similarité sous la forme d'une somme de contributions normalisées, permet la prise en compte de différents types de variables, dans le cadre d'une même description.

Nous avons déjà évoqué que les données manquantes ne sont pas prises en compte dans CHAVL, mais qu'il était aisé de les ignorer dans les comparaisons deux à deux entre variables. Il est également aisé des les ignorer dans les comparaisons deux à deux entre objets. A cette fin, la normalisation statistique d'une même variable relationnelle se basera sur le sous ensemble des objets où elle a valeur. D'autre part, dans la comparaison de deux objets donnés, on prendra la moyenne (au lieu de la somme) des contributions normalisées des variables qui ont valeur sur chacun des deux objets.

1.3 Cas où la description concerne un ensemble C de catégories

1.3.1 Le cas le plus simple d'un tableau de contingence

Relativement à la définition générale d'un tableau de données que nous avons proposée [cf. figure 1, § 2 de la première partie et autour], le cas d'un tableau de contingence rentre dans le cadre (i, i) de la description d'un ensemble C de catégories, classes ou concepts. Il s'agit en fait du cas où l'ensemble A comprend une seule variable de type qualitatif nominal que nous notons a. D'autre part, l'ensemble

$$C = \{c_i / i \in I = \{1, 2, \ldots, n\}\} \tag{126}$$

de concepts est observé sur un ensemble d'objets élémentaires, généralement déterminé par un échantillon E de la population à laquelle on s'intéresse. La taille de ce dernier pourra être notée - comme il est d'usage - k . On peut, sans ambiguïté, désigner par c_i la classe de l'ensemble des objets qui relèvent de la catégorie $c_i, 1 \leq i \leq n$.

Dans ces conditions en désignant par :

$$A = \{a_j / j \in J = \{1, 2, \ldots, j, \ldots, p_a\}\} \tag{127}$$

l'ensemble des modalités de la variable a, la valeur $a(c_i)$ est une distribution statistique sur l'ensemble des valeurs que nous venons de définir [cf. (127)]. $a(c_i)$ peut donc être représenté par un vecteur d'entiers

$$(k_{ij} \ : 1 \leq j \leq p_a) \tag{128}$$

où k_{ij} est le nombre d'objets de la classe c_i qui possèdent la modalité a_j. On peut associer à a_j (resp. c_i) un attribut modalité α_j (resp. γ_i) qui est un attribut booléen qui est à "vrai" sur un objet x, si et seulement si x possède la modalité a_j du caractère a (resp. si x relève de la catégorie c_i). On a alors :

$$k_{ij} = card\{[\sigma_i^{-1}(1)] \cap [\alpha_j^{-1}(1)]\} \tag{129}$$

$1 \leq i \leq n, 1 \leq j \leq p_a$. Finalement, le tableau qu'on note $k_{I \times J}$:

$$\{k_{ij}/(i,j) \in I \times J\} \tag{130}$$

est de contingence et peut être considéré comme résultant du croisement sur E entre deux variables qualitatives nominales dont C [cf. (126)] est l'ensemble des valeurs de la première et dont A [cf. (127)] est l'ensemble des valeurs de la seconde. Il y a alors lieu d'analyser -par la classification ici- I à travers J. Mais il peut également s'agir dualement et exactement de la même façon d'analyser J à travers I ; car la structure mathématique du tableau de contingence est parfaitement symétrique.

Cependant, pour effectuer l'analyse -par exemple et pour fixer les idées de I à travers J- il faut nécessairement que dans la représentation mathématique-logique, on introduise une dissymétrie. Cette dissymétrie est celle qui sépare la notion de variable descriptive de celle d'objet décrit. Il s'agit de façon cohérente de faire jouer à I le rôle de l'ensemble des objets et à J, celui de l'ensemble des variables ; ou bien, de faire jouer à I le rôle de l'ensemble des variables et à J, celui de l'ensemble des objets. C'est exactement ce qui est fait dans la représentation géométrique que propose l'analyse factorielle des correspondances. Dans le premier cas, on associe à I, le nuage de points, qu'on note classiquement :

$$N(I) = \{(f_j^i, p_i)/i \in I\} \tag{131}$$

Dans l'espace géométrique $\mathbb{R}^p$, où $p = card(J)$. f_j^i est le point de $\mathbb{R}^p$ qui représente i; sa j-ème composante est la proportion relative :

$$f_j^i = \frac{k_{ij}}{k_i}\ ,\ 1 \leq j \leq p \tag{132}$$

où k_i est la somme pour j des $k_{ij}, 1 \leq i \leq n$. On a

$$k_i = card[\gamma_i^{-1}(1)] \tag{133}$$

Ainsi ici, i de I peut être matérialisé par un objet O_i qu'on représente dans $\mathbb{R}^p$ par le point dont la suite des composantes est définie par (132); ce qui suppose implicitement que chaque j de J, se trouve représenté par une variable quantitative v^j qui est à son tour représentée par la j-ème forme linéaire coordonnée de $\mathbb{R}^p$. Dans ces conditions

$$v^j(\sigma_i) = f_j^i, \tag{134}$$

$1 \leq i \leq n, 1 \leq j \leq p$.

Le point f_j^i est affecté du poids

$$p_i = \frac{k_i}{k}, \tag{135}$$

$1 \leq i \leq n$.

Enfin, l'espace $\mathbb{R}^p$ est muni -pour des raisons algèbriques et statistiques- de la métrique diagonale du χ^2 :

$$(1/p_j\ ; 1 \leq j \leq p), \tag{136}$$

où, avec des notations que l'on comprend :

$$p_j = \frac{k_j}{k}, \tag{137}$$

avec

$$k_j = card[\alpha_j^{-1}(1)], \tag{138}$$

$1 \leq j \leq p$.

C'est la représentation que nous venons de développer qui prévaut dans CHAVL pour la classification de I, conformément au point de vue développé au paragraphe 1.2.5 dans le cas où les variables sont de type quantitatif-numérique. Puisque nous plaçons l'origine au centre de gravité du nuage $N(I)$ et ; en tenant compte de la métrique [cf. (136)], le correspondant de ξ_i^j [cf. (105)] devient ici (on remarquera à gauche l'inversion des positions des deux indices i et j :

$$\varphi_i^j = \frac{(f_j^i - p_j)/\sqrt{p_j}}{\sqrt{\sum_{1 \leq h \leq p} (f_h^i - p_h)^2/p_h}} \tag{139}$$

La contribution brute de j à la comparaison de i et de i' est alors

$$s_j(i,i') = \frac{1}{p} - \frac{1}{2}(\varphi_i^j - \varphi_{i'}^j)^2 \tag{140}$$

Maintenant, en introduisant pour $r = 1, 2, 3$ et 4, le r-ème moment absolu :

$$M_r(j) = \sum_{1 \leq i \leq n} p_i \ (\varphi_i^j)^r, \tag{141}$$

on obtient la moyenne μ_j et la variance σ_j^2 de $s_j(i,i')$ sur $I \times I$ convenablement pondéré ; c'est à dire, par $\{p_i \times p_{i'} / (i,i') \in I \times I\}$:

$$\mu_j = \frac{1}{p} - M_2(j) + [M_1(j)]^2 \tag{142}$$

et

$$\begin{aligned} \sigma_j^2 \ = \ & \frac{1}{2}M_4(j) - 2M_3(j)M_1(j) + \frac{1}{2}[M_2(j)]^2 \\ & +2M_2(j)[M_1(j)]^2 - [M_1(j)]^4 \end{aligned} \tag{143}$$

On obtient alors l'indice $S_j(i,i')$ qui résulte de la normalisation statistique de $s_j(i,i')$, conformément à la formule (107) ci-dessus.

La suite des calculs est conforme à ce que qui a été maintes fois exprimé [cf. expressions (67) à (71) première partie].

La demi matrice inférieure des indices centrés et réduits globalement [de même forme que (69) première partie] est calculée dans le cadre de l'étape SIMOB.

Dans nos précédents programmes, nous avons avec B. Tallur [Lerman & Tallur 1980] fait jouer à l'ensemble à classifier (I ou J) le rôle de l'ensemble des variables, qui sont dans notre cas,

quantitatives [cf. (134)]. S'il s'agit comme ci-dessus, de la classification de I, on considère le nuage $N(J)$, dual de celui de $N(I)$ ci-dessus. Dans ces conditions, le coefficient d'association qui est une corrélation se met sous la forme initiale suivante :

$$R(i,i') = \frac{\sum_j p_j \left(f_i^j - p_i\right)\left(f_{i'}^j - p_{i'}\right)}{\left\{\left[\sum_j p_j \left(f_i^j - p_i\right)^2\right]\left[\sum_j p_j \left(f_{i'}^j - p_{i'}\right)^2\right]\right\}^{1/2}} \tag{144}$$

Nous démontrons dans [Lerman & Peter 1985] qu'un indice de type corrélation dans l'un des espaces est un indice de type cosinus dans l'espace dual. Ainsi, $R(i,i')$ est le cosinus de l'angle $\widehat{f_j^i g_J f_j^{i'}}$, où g_J est le centre de gravité du nuage $N(I)$. Et, précisément, nous sommes partis de ce cosinus pour l'élaboration de l'indice de similarité implanté dans SIMOB (cf. § 1.2.5).

1.3.2 Extensions.

Une structure de données qui se rencontre fréquemment est celle d'une juxtaposition "horizontale" de tableaux de contingence de la forme :

$$I \times (J^{(1)} \cup J^{(2)} \cup \ldots \cup J^{(l)} \cup \ldots \cup J^{(L)} \tag{145}$$

où I (resp. $J^{(l)}, 1 \leq l \leq L$) se trouve défini par l'ensemble des modalités d'une variable qualitative nominale. Par rapport à la définition générale proposée d'un tableau de données [cf. figure 2 première partie, § 2 et autour], il s'agit du cas, s'inscrivant dans le cadre (ii) où A est formé de variables qualitatives nominales, en nombre L ; conformément à (145) ci-dessus :

$$A = \{a^l / 1 \leq l \leq L\} \tag{146}$$

Si on désigne par n le nombre d'éléments de I et par p_l le nombre de modalités de a^l, il faut une table de

$$n \times (p_1 + p_2 + \ldots + p_l + \ldots + p_L) \tag{147}$$

nombres pour consigner l'information statistique.

Les différents indices élaborés dans le contexte d'une seule table de contingence peuvent naturellement être étendus à des structures de données qui généralisent cette dernière :

(i) double juxtaposition "verticale" et "horizontale" de tables de contingence de la forme :

$$(I^{(1)} \cup \ldots \cup I^{(k)} \cup \ldots \cup I^{(K)}) \times (J^{(1)} \cup \ldots \cup J^{(l)} \cup \ldots \cup J^{(L)}) \; ; \tag{148}$$

(iii) données cubiques ou à plusieurs dimensions de contingence.

Ces travaux, dans l'optique corrélationnelle qui nous concerne, sont développés dans [Lerman & Tallur 1980], [Lerman 1984] et [Tallur 1988].

Toujours dans une optique corrélationnelle et relativement à une seule table de contingence $I \times J$ et à la famille de distributions $\{f_j^i / i \in I\}$ sur J associée [cf. (130),(131)(132)], on peut considérer l'indice brut suivant

$$\sum_{1 \leq j \leq p} \sqrt{f_j^i f_j^{i'}} \tag{149}$$

de comparaison entre i et i' qui correspond au cœfficient d'affinité de Matusita [Matusita 1967] et est sous jacent à la métrique de Hellinger où une même distribution f_j^i est représentée par le point

$$(\sqrt{f_j^i} / 1 \leq j \leq p) \tag{150}$$

à coordonnées positives de la sphère unité H. Bacelar-Nicolaü [icolaü 1988] a particulièrement développé l'optique "vraisemblance du lien" en partant de (149). Mais, dans ce cas là, on peut ne pas pouvoir tenir compte des poids respectifs des différentes distributions ; soit

$$\{p_i / i \in I\} \tag{151}$$

Rappelons maintenant que nous avons vu [cf. (88)] que, dans le cas de la description d'un ensemble O d'objets élémentaires, le tableau de données représente la concrétisation d'un système de Tarski. D'autre part, nous venons d'exprimer ci-dessus que le cas (145) est particulier de la structure générale d'un tableau de données tel que nous l'envisageons pour la description d'un ensemble C de concepts (classes ou catégories). Dans ce dernier cas, le tableau des données matérialise un système de la forme :

$$S = < C ; R_1, R_2, \ldots, R_j, \ldots, R_p >, \tag{152}$$

où les R_j sont des relations de même arité -éventuellement valuées- et où on se donne pour chaque relation R_j, la suite de ses distributions statistiques sur la suite des concepts (ou classes). R_j est induite par la j-ème variable, $1 \leq j \leq p$.

Le programme AVAND (Association entre VAriables à modalités Non Disjointes) mis au point par M. Ouali-Allah [Ouali-Allah 1991] conformément aux normes Modulad, peut être considéré comme l'élaboration de la matrice de cœfficients d'association entre variables relationnelles préordonnances décrivant un ensemble C de concepts [cf. (152)]. La description peut également concerner un ensemble O d'objets élémentaires dans le cas (non couvert par AVARE) où les modalités d'une même variable ne sont pas logiquement exclusives. Ainsi, un même objet (ou individu) possède un sous ensemble -éventuellement vide ou plein- de modalités d'une même variable préordonnance $w^j, 1 \leq j \leq p$.

L'introduction du système S [cf. (152)] s'est imposée à nous à partir de la description la plus complexe effectivement traitée par AVL. Il s'agit d'une base de connaissance (due à J. Lebbe, J.P. Dedet et R. Vignes, Université Paris 6, Institut Pasteur de la Guyane française) d'un

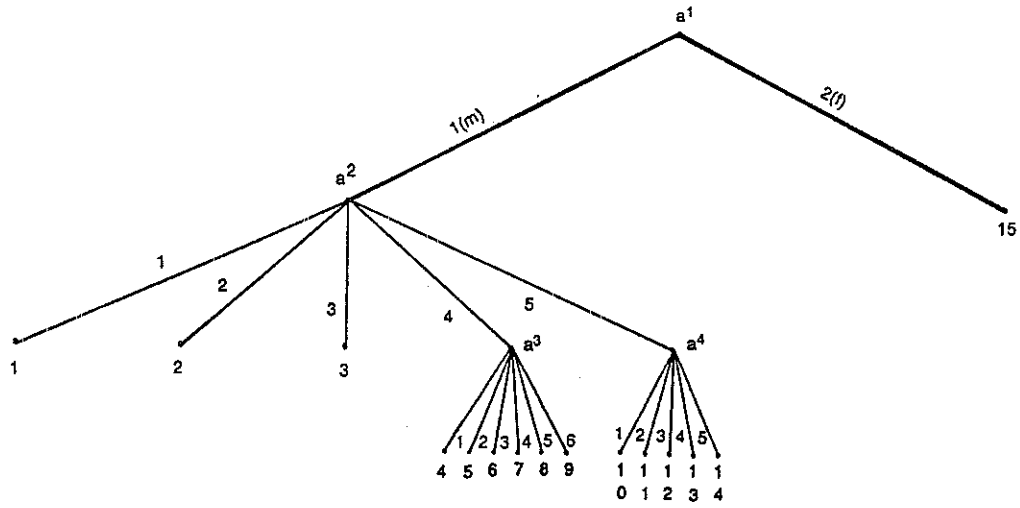

FIG. 1 - Variable "préordonnance taxinomique à choix multiple"

ensemble d'espèces de phlébotomes de la Guyane française. La structuration de la base nécessite la classification de l'ensemble d'espèces, donc d'un ensemble de concepts [Lerman & Peter 1989]. Pour garder toute la richesse de l'information descriptive, nous avons introduit la notion de "variable préordonnance taxinomique à choix multiple". Une telle macro-variable résulte de l'organisation d'une suite de variables qualitatives, partiellement ordonnée selon la relation mère-fille. On suppose que l'ensemble des modalités d'une même variable qualitative -se situant à un même niveau de la taxinomie- se trouve muni d'une préordonnance totale (traduisant de façon ordinale les ressemblances entre modalités). Par ailleurs, la valeur d'une telle variable sur un concept donné (il s'agit ici d'une espèce) est définie par une formule logique portant sur l'ensemble de ses modalités. Dans notre cas, cette variable a été bien formalisée par la notion de variable à modalités non disjointes, ci-dessus introduite.

Exemple : Nous notons a^1, a^2, a^3 et a^4 les variables 1, 18, 19 et 20 de la base de connaissance ci-dessus mentionnée. a^1 est le sexe, a^2 est le nombre d'épines du style, a^3 est la disposition des 4 épines du style et a^4 est la disposition des 5 épines du style. Nous obtenons la structure taxinomique de la figure 1 ci-dessus. La variable préordonnance taxinomique associée définit une même relation R_j de (152).

2 Pour aller plus loin dans l'optique d'AVL.

Quelle que soit la structure mathématico-logique des données traitées [système T (cf. (88)) ou système S (cf. (152))] et en nous référant aux figures 2 et 3 de la première partie, le schéma de la démarche générale de traitement par l'AVL est le suivant

- FAIRE: **E = A**

ARBA = AVL(E)

- FAIRE: **E = O (resp. C)**

selon la nature du tableau des données.

ARBO = AVL(E)

Si **E = A** $\Rightarrow$ poser **dual(E) = O (resp. C)**,

selon la nature du tableau des données.

Si **E = O (resp. C)** $\Rightarrow$ poser **dual (E) = A**.

Le premier aspect que nous mentionnerons est celui de croisement. En effet, une étape très importante de l' "explication" des classes suppose qu'on peut "situer" au moyen de cœfficients d'association, un système organisé de classes sur une partie de E [resp. $dual(E)$], par rapport à un système organisé de classes sur une partie de $dual(E)$ (resp. E).

Dans ces conditions, on pourra par exemple parler du "degré de responsabilité" de telle classe d'attributs dans la formation de telle classe d'objets.

La gestion de la situation relative **ARBA** par rapport à **ARBO** (cf. ci-dessus) permet aux techniques d'intelligence artificielle d'intervenir plus efficacement dans l' "explication" de la synthèse automatique

Citons ici quelques références importantes qui se situent dans le cadre de ces cœfficients de croisement [Lerman 1983_b, 1984_b], [Lerman, Hardouin & Chantrel 1980], [Moreau 1985] et [Ouali-Allah 1991].

les cœfficients d'association entre variables relationnelles ci-dessus considérés ont un caractère "total". Mais on peut vouloir dans l'organisation des liaisons entre variables (production de ARBA dans le cadre du schéma ci-dessus), neutraliser l'influence d'une variable ou même d'un groupe de variables exogènes [Lerman 1983_c, 1983_d]. Des expériences très intéressantes ont été menées par A. Sbii (ancien thésard de 3ème cycle). Il s'agit là d'un deuxième aspect d'importance, permettant l'extension de l'approche. Nous avons bien exprimé que, quelle que soit la nature de l'ensemble E à organiser, nous aboutisssons à une table

$$\{P(x,y)/(x,y) \in P_2(E)\} \tag{153}$$

d'indices probabilistes de similarité (cf. première partie fig. 6). Or, de nombreuses méthodes de représentation en analyse des données supposent la donnée d'un indice de dissimilarité, voire de distance. Néanmoins, on peut aisément passer d'un indice de dissimilarité à un indice de distance par l'addition d'une constance minimale [Caillez 1983]. L'indice de dissimilarité qui nous paraît le plus naturellement se déduire de l'indice probabiliste $P(x,y)$, est la quantité d'information de l'évènement dont la probabilité est $P(x,y)$ Très précisement, on substituera à la table (153) ci-dessus, la table

$$\{-\log_2[P(x,y)]/\{x,y\} \in P_2(E)\} \tag{154}$$

avant de s'engager dans une méthode de représentation, dont l'argument est une matrice de dissimilarités. L'analyse du comportement d'une telle matrice pour l'analyse des données correspond à un sujet très actuel de nos travaux.

Avant de clore cette notion de similarité symétrique, signalons que dans l'élaboration d'un coefficient d'association entre deux variables relationnelles, conformément à l'optique AVL, l'indice brut centré est réduit au moyen de l'écart-type de l'indice brut aléatoire [cf. formules (21) (première partie), (97), (101)]. Une autre optique consiste à réduire au moyen de la valeur maximale que peut atteindre l'indice centré sous contrainte de l'hypothèse d'absence de liaison. Cette autre optique conduit parfois à des problèmes très difficiles d'optimisation combinatoire [Lerman 1987_b], [Lerman et Peter 1988] et [Messatfa 1990].

Nous allons maintenant aborder les autres structures de représentation. Avant de quitter la structure d'arbre, soulignons un intérêt particulier de nos coefficients d'association pour la construction d'arbres de décision. Il s'agit là d'un sujet sur lequel des travaux sont initialisés.

En les rangeant de façon croissante selon leur caractère métrique, ces structures de représentation synthétique sont, parmi les plus considérées :

(i) les ordres ou préordres (totaux ou partiels) ;

(ii) les arbres additifs ;

(iii) les analyses factorielles sur tableaux de distances.

Nous avons déjà fait entrevoir au moyen de (154) comment utiliser nos indices dans les cas (ii) [Barthélémy et Guénoche 1988] et (iii).

Pour ce qui est de (i), plaçons nous pour fixer les idées, dans le cas où les attributs sont booléens, de présence-absence. Le problème de la recherche d'un ordre ou préordre total sur l'ensemble O des objets élémentaires ; et, dualement, sur l'ensemble A des attributs, est apparu dans les méthodes de recherche d'une "sériation" sur des objets archéologiques. Il s'agit de trouver un couple de permutations sur l'ensemble des lignes et des colonnes du tableau d'incidence des données, de façon à faire apparaître une forme diagonale, la plus chargée de valeurs 1. Nous avons déjà exprimé nos références à la suite du tableau 14 (§ 3.4 première partie). Citons également en France [Guénoche 1987, Marcotorchino 1991].

Les ordres ou préordres partiels sont apparus dans notre environnement de recherche -comme structure de représentation- sous la forme de graphes d' "implication", entre stimuli ou classes de stimuli, dans le cadre de la recherche en didactique. Il s'agit en effet dans ce cas de remplacer la notion symétrique d'indice de similarité probabiliste, par celle orientée d'indice d'implication probabiliste. Ce dernier permet d'évaluer, dans quelle mesure une réponse à vrai à un attribut booléen a, implique une réponse à vrai à un attribut booléen b [Lerman, Gras et Rostam 1981] et [Gras et Larher 1992]. Dans ce dernier travail on propose bien une structure en arbre de classification, mais il s'agit d'un arbre "implicatif" entre classes de stimuli. Dans ce cas, le lien -représenté par un nœud- entre deux classes adjacentes, est orienté.

Nous avons bien exprimé au paragraphe 2 de la première partie que tout élément de la famille de critères d'agrégation de la vraisemblance du lien maximal était contractant. Ce qui permet d'envisager la construction d'algorithmiques de classification de "très gros" ensembles (quelques dizaines de milliers d'éléments). Les techniques vont utiliser de façon principale des notions qui portent les noms suivants :

(i) "voisins réciproques" et "voisinages réductibles" [Bruynooghe 1989] ;

(ii) "classification hiérarchique en parallèle" [Lerman et Peter 1984] et [Peter 1987].

Nous avons récemment (1991) [dans *"votre thèse à l'Irisa"*, brochure éditée par l'Irisa (Inria-Cnrs, 1991)] proposé un sujet de thèse où il y a lieu de combiner de façon optimale les précédentes idées algorithmiques.

Les derniers problèmes que nous évoquerons sont de stabilité et de validité statistique de la classification. Conformément au schéma ci-dessus (début du paragraphe V), désignons par E l'ensemble à classifier ; si $E = O$ (resp. C), on posera $E^* = A$; et si $E = A$, on posera $E^* = O$ (resp. C). Les problèmes de stabilité et de validité statistique se situent à deux niveaux :

(i) relativement à la taille de E^* ;

(ii) relativement à l'adjonction (resp. extraction) d'éléments à E (resp. de $E*$).

Pour (i) nous avons surtout étudié le cas où E est l'ensemble A des attributs ou variables de description [Lerman 1986]. Mais le problème se pose également si E est l'ensemble O (resp. C) des objets élémentaires (resp. des concepts ou classes) décrits.

Deux points restent à indiquer avant d'aborder les références bibliographiques.

Le présent texte a bénéficié d'aspects notables d'un précédent article [Lerman 1993]. Cependant, ce que nous exprimons ici est davantage dans un esprit Modulad et pour un public plus averti.

D'autre part, nos références bibliographiques sont très concentrées autour de nos travaux qui concernent AVL. Une image plus équilibrée consisterait en l'ensemble des bibliographies respectives de ces travaux. De toute façon, ces derniers se situent dans le cadre de l'Analyse Combinatoire et Statistique des Données qui est un domaine maintenant bien établi et d'une très grande richesse. Tout en étant à une certaine distance de prétendre à l'exhaustivité, l'impressionnante revue synthétique d'Arabie et Hubert [Arabie et Hubert 1992] regroupe près de 500 références - comprenant plusieurs ouvrages - et relatives pour la plupart, à ces dix dernières années.

Références

[Arabie et Hubert 92] Arabie (P.) et Hubert (L.J.). – Combinatorial data analysis. *Annu. Rev. Psychol.*, n° 43, 1992, pp. 169–203.

[Barthélémy et Guénoche 88] Barthélémy (J.P.) et Guénoche (A.). – *Les arbres et les représentations des proximités*. – Paris, Masson, 1988.

[Bruynooghe 89] Bruynooghe (M.). – *Nouveaux algorithmes en classification automatique applicables aux très gros ensembles de données, rencontrés en traitement d'images et en reconnaissance des formes*. – Thèse, Université de Paris VI, 1989.

[Caillez 83] Caillez (F.). – The analytical solution of the additive constant problem. *Psychometrika*, vol. 48, n° 2, 1983, pp. 305–308.

[Daude 92] Daude (F.). – *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*. – Thèse, université de Rennes 1, 1992.

[Ghazzali 92] Ghazzali (N.). – *Comparaison et réduction d'arbres de classification, en relation avec des problèmes de quantification en imagerie numérique*. – Thèse, Université de Rennes I, 1992.

[Giakoumakis et Monjardet 87] Giakoumakis (V.) et Monjardet (B.). – Coefficients d'accord entre deux préordres totaux. *Rev. Statistique et Analyse des Données*, n° 12, 1987, pp. 46–99.

[Gras et Larher 92] Gras (R.) et Larher (A.). – L'implication statistique, nouvelle méthode d'analyse des données. *Rev. Math., Inf. Sci. Hum.*, n° 120, 1992

[Guenoche 87] Guénoche (A.). – Méthodes combinatoires de sériation à partir d'une dissimilarité. Inria, Cinquièmes Journées Internationales "Analyse des Données Informatique", 29 sept-2 oct. 1987 – Versailles, Inria, North-Holland, 1987.

[Jaccard 08] Jaccard (P.). – Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, n° 44, 1908, pp. 223–2790.

[Leredde 79] Leredde (H.). – *La méthode des pôles d'attraction ; la méthode des pôles d'agrégation : deux nouvelles familles d'algorithmes en classification automatique et sériation*. – Thèse, Université Paris VI, 1979.

[Lerman 72] Lerman (I.C.). – Analyse du phénomène de la "sériation". *Rev. Math. Sci. Hum.*, n° 38, 1972, pp. 39–57.

[Lerman 81] Lerman (I.C.). – *Classification et analyse ordinale des donnies*. – Paris, Dunod, 1981.

[Lerman 83a] Lerman (I.C). – Sur la signification des classes issues d'une classification automatique. *Numerical Taxonomy, NATO ASI Series vo. G1*, éd. par Felsenstein (J.), pp. 179–198. – Springer Verlag, 1983.

[Lerman 83b] Lerman (I.C.). – Association entre variables qualitatives ordinales nettes ou flowes. *Rev. Statistique et Analyse des Données*, vol. 8, n° 7, 1983, pp. 41–73.

[Lerman 83c] Lerman (I.C.) – Indices d'association partielle entre variables qualitatives nominales. *Rairo/Oper. Res.*, n° 17, 1983, pp. 213–259.

[Lerman 83d] Lerman (I.C.) – Indices d'association partielle entre variables qualitatives ordinales. *Publ. Inst. Stat. univ. Paris*, n° 28, 1983, pp. 7–46.

[Lerman 84a] Lerman (I.C) – Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. *Publ. Inst. Stat. univ. Paris*, n° 29, 1984, pp. 27–57.

[Lerman 84b] Lerman (I.C.). – Analyse classificatoire d'une correspondance multiple, typologie et régression. *Data Analysis and Informatics, III*, E. Diday. North Holland, 1984, pp. 193–221. –

[Lerman 86] Lerman (I.C.) – Rôle de l'inférence statistique dans une approche de l'analyse classificatoire des donnies. *Journal de la société de statistique de Paris*, n° 4, 1986, pp. 238–252.

[Lerman 87a] Lerman (I.C.) – Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. application au problème du consensus en classification. *Rev. Stat. Appl.*, n° 35, 1987, pp. 39–60.

[Lerman 87b] Lerman (I.C.) – Maximisation de l'association entre deux variables qualitatives ordinales. *Rev. Math. Sci. Hum.*, n° 100, 1987, pp. 49–56.

[Lerman 89] Lerman (I. C.) – Formules de réactualisation en cas d'agrégations multiples. *Rairo série R.O.*, vol. 23, n° 2, 1989, pp. 151–163.

[Lerman 91] Lerman (I.C.) – Foundations of the likelihood linkage analysis (lla) classification method. *Applied Stochastic Models and Data Analysis John Wiley*, vol. 7, 1991, pp. 69–76.

[Lerman 92a] Lerman (I.C.) – Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles i. *Rev. Math. Infor. & Sci. Hum., 30e année, Paris*, n° 118, 1992, pp. 35–52.

[Lerman 92b] Lerman (I.C.) – Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles II. *Rev. Math. Infor. & Sci. Hum., 30e année, Paris*, n° 119, 1992, pp. 75–100.

[Lerman 93] Lerman (I.C.) – *Likelihood linkage analysis (LLA) classification method: An example treated by hand*, Biochimie, Elsevier editions, 1993, volume 75, pp. 379–397. –

[Lerman, Gras et Rostam 81] Lerman (I.C), Gras (R.) et Rostam (H.) – Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II. *Rev. Mat. Sci. Hum.*, n° 74 et 75, 1981, pp. 5–35 et 5–47.

[Lerman, Hardouin et Chantrel 80a] Lerman (I.C.), Hardouin (M.) et Chantrel (T.) – Analyse de la situation relative entre deux classifications floues. *Data Analysis and Informatics*. Secondes Journées Internationales Analyse des Données et Informatique, Versailles 17-19 octobre 1979, North Holland, 1980, pp. 523–552. –

[Lerman et Tallur 80b] Lerman (I.C.) et Tallur (B.). – Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence. *Rev. de Statistique Appliquée*, n° 28, 3, Paris 1980, pp. 5–28.

[Lerman et Peter 84] Lerman (I.C.) et Peter (P.). – *Analyse d'un algorithme de classification hiérarchique "en parallèle" pour le traitement de gros ensembles.* – rapport de recherche n° 339, Inria, Le Chesnay 1984.

[Lerman et Peter 85] Lerman (I.C.) et Peter (P.). – *Elaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème du consensus en classification.* – publication interne, 72 pages n° 262, IRISA, Rennes, juillet 1985.

[Lerman et Peter 88] Lerman (I.C.) et Peter (P.). – Structure maximale pour la somme des carrés d'une contingence aux marges fixées ; une solution algorithmique programmée. *Rairo/Oper. Res. 22*, 1988, pp. 83–136.

[Lerman et Peter 89] Lerman (I.C.) et Peter (P.). – Classification of concepts described by taxonomic preordonnance variables with multiple choice. application to the structuration of a species set of phlebotomine. *Data Analysis, Learning Symbolic and Numeric Knowledge*, Diday Ed. Inria. Proceed of the Conf. Antibes, sept. 11-14 1989, pp. 73–86. – Nova Science Publishers Inc., New-York, 1989.

[Lerman et Ghazzali 91] Lerman (I.C.) et Ghazzali (N.). – What do we retain from a classification tree? an experiment in image coding. *Symbolic-Numeric Data Analysis and Learning* , E. Diday et Y. Lechevallier, Inria, Versailles, sept. 18-20, 1991, pp. 27–42. –

[Marcotorchino 91] Marcotorchino (F.). – Seriation problems : an overview. *Applied Stochastic Models and Data Analysis, John Wiley*, n° 7, 1991, pp 139–151.

[Matusita 67] Matusita (K.). – On the notion of affinity of several distributions and some of its applications. *Ann. Math. Stat.*, vol. 19, n° 2, 1967, pp. 181–192.

[Messatfa 90] Messatfa (H.). – *Unification relationnelle des critères et structures optimales des tables de contingence.* – Thèse, Université de Paris VI, 1990.

[Mollière 86] Mollière (J.L.). – What's the real number of clusters? *Classification as a Tool of Research*, éd. par Gaul (W.) et Schader (M.). – North Holland, 1986.

[Moreau 85] Moreau (A.). – *Elaboration et calcul d'indices d'association entre variables qualitatives "nettes" ou "floues". Application à une forme d'interprétation d'une classification de paramètres épidémiologiques* – Thèse, Université de Rennes I, 1985.

[Nicolau 80] Nicolau (F.). – *Critérios de analise classificatoria hiérarquica baseados na furgao de distribuigao.* – Thèse, Faculté des Sciences de Lisbonne, 1980.

[Nicolaü 88] Nicolaü (H. Bacelar). – Two probabilistic models for classification of variables in frequency tables. The First Conference of the IFCS, Technical Univ. of Aachen, FRG 29 june-1 July 1987, 1988, pp. 181–186. –

[Ouali-Allah 91] Ouali-Allah (M.). – *Analyse en préordonnances des données qualitatives. Applications aux données numériques et symboliques.* – Thèse, Université de Rennes I, 1991.

[Peter 87] Peter (P.). – *Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations assistées par ordinateur*. – Thèse, Université de Rennes I, 1987.

[Tallur 88] Tallur (B.). – *Contribution l'analyse exploratoire de tableaux de contingence par la classification*. – Thèse, Université de Rennes I, 1988.

[Tarski 54] Tarski (A.). – Contribution to the theory of models, I et II. *Indagationes Mathematicae*, n° 16, 1954, pp. 572–588.