REVUE DE STATISTIQUE APPLIQUÉE

I.-C. LERMAN

Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques

Revue de statistique appliquée, tome 54, n° 2 (2006), p. 33-63

<http://www.numdam.org/item?id=RSA_2006__54_2_33_0>

Rev. Statistique Appliquée, 2006, **LIV** (2), 33-63

COEFFICIENT NUMÉRIQUE GÉNÉRAL DE DISCRIMINATION DE CLASSES D'OBJETS PAR DES VARIABLES DE TYPES QUELCONQUES. APPLICATION À DES DONNÉES GÉNOTYPIQUES

I.-C. LERMAN

Irisa - Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cédex
lerman@irisa.fr

RÉSUMÉ

Nous proposons une forme unique pour un indice numérique de discrimination ou d'explication de classes homogènes d'individus ou de catégories par des variables descriptives. Cette forme embrasse les structures de données les plus diverses : numériques, booléennes, de contingence et surtout, relationnelles de différents types. Cette forme est issue des cas de données classiques (e.g. numériques) pour recouvrir le cas des données relationnelles. Cette extension est fondée sur un développement spécifique faisant appel à une famille de coefficients d'association entre variables relationnelles. Une application en cas de données génotypiques liées à l'hémochromatose (surcharge en fer) est considérée.

Mots-clés : *discrimination, classification, rapport de corrélation, variables qualitatives relationnelles, données génotypiques*

ABSTRACT

In order to discriminate homogeneous object clusters by descriptive variables of any kind, a unique form of a numerical coefficient is built. The structural nature of this index is initially provided by numerical variables. Then, it is extended to all types of data : numerical, boolean, contingency and relational data. This generalization is based on a specific development in which a general family of relational association coefficients takes an important part. Application in case of genotypic data related to iron overloading (emochromatosis illness) is studied with respect to the new index.

Keywords : *discrimination, correlation ratio, classification, relational qualitative variables, genotypic data*

1. Introduction

Relativement à une description d'un ensemble O d'objets, par un ensemble A d'attributs (on dit encore variables ou caractères), on suppose établie une partition de O en classes suffisamment homogènes. Cette notion d'homogénéïté est statistique et peut se comprendre de façon intuitive. En prenant deux objets au hasard dans

O, leur similarité par rapport à la description a tendance à être élevée si les deux objets sont dans la même classe. Elle a au contraire tendance à être petite si les deux objets se trouvent dans deux classes distinctes. Une telle partition qu'on peut noter $\pi = \{O_c \mid 1 \leqslant c \leqslant C\}$, où C est le nombre de ses classes, est généralement obtenue à partir d'un algorithme de classification qui opère sur l'ensemble O des objets. Si n est le nombre des objets et p le nombre de variables, on peut noter $O = \{o_i \mid 1 \leqslant i \leqslant n\}$ et $A = \{a^j \mid 1 \leqslant j \leqslant p\}$.

Un problème fondamental de la classification des données consiste à « comprendre » la partiton π et à interpréter chacune de ses classes dans les termes de la description initiale. Considérons le cas de référence le plus classique où l'attribut est numérique [2]. Dans ce cas que nous reprenons au paragraphe 2, un rapport de corrélation permet de mesurer le degré de discrimination de la partition par l'attribut. Un tel indice se présente d'une part, sous la forme d'un rapport de variances et d'autre part, du carré d'un coefficient de corrélation. Nous mentionnerons l'extension d'un tel indice dans le cas de données booléennes ou de contingence (explication d'une classification de lignes à travers les colonnes). Une telle extension tient compte de la représentation géométrique de ces derniers types de données.

Le cas le plus original concerne les données qualitatives de différents types. L'interprétation en termes de préordonnances ou plus généralement de graphes valués sur l'ensemble O va permettre de les inscrire dans un même moule. C'est l'interprétation corrélative de l'indice qui conduira de façon naturelle à traiter le cas de variables qualitatives de toutes sortes, chacune prise dans sa globalité. On a ainsi une ligne directrice commune permettant de traiter ce cas difficile. C'est au paragraphe 3 que nous la mettrons en évidence. Sa mise en oeuvre fait appel à la construction d'une famille très générale de coefficients d'association entre variables relationnelles [12, 13, 19, 20]. Rappelons ici que ce qui est proposé dans la littérature consiste à substituer à une même variable qualitative un ensemble d'attributs booléens dont chacun se trouve défini par une *valeur* (on dit encore *modalité ou catégorie*) de la variable initiale [2, 21]. De la sorte, on perd toute la sémantique qu'il y a derrière l'ensemble des valeurs de la variable. L'application à des données génotypiques concernant le métabolisme du fer est étudiée au paragraphe 4. Le programme v-class réalisant informatiquement les calculs mathématiques y sera mentionné. Ce paragraphe reprend en son sein le résultat des travaux de Véronique Adoue (DESS-CCI (Compétence Complémentaire en Informatique)) [1] que nous avons dirigés. Nous terminerons par une conclusion au paragraphe 5.

2. Cas où les données sont numériques booléennes ou de contingence

Le cas des données numériques est celui le plus classique. L'indice global de discrimination d'une partition est défini par le rapport de corrélation dont nous allons rappeler le principe et les écritures associées.

La donnée est formée d'un couple (π, v) où, sur l'ensemble O des objets, π définit une partition et v, une valuation numérique. n désignant la taille de O, indiquons par $I = \{1, 2, ..., i, ..., n\}$ l'ensemble des indices de :

$$O = \{o_i \mid 1 \leqslant i \leqslant n\} \tag{1}$$

On suppose que la partition π est en C classes :

$$\pi = \{O_c \mid 1 \leqslant c \leqslant C\} \tag{2}$$

et on désigne par

$$\underline{C} = \{1, 2, ..., c, ..., C\} \tag{3}$$

l'ensemble des étiquettes des classes.

La classe O_c sera indexée par le sous ensemble I_c de I. On pose $n_c = |O_c| = |I_c|$, où $|.|$ désigne le cardinal de ., $1 \leqslant c \leqslant C$. On a bien sûr :

$$\sum_{c=1}^{c=C} n_c = n \tag{4}$$

Désignons par :

$$\{x_i \mid 1 \leqslant i \leqslant n\} \tag{5}$$

la suite des valeurs de la variable v sur I. On note la moyenne générale sous la forme

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{i=n} x_i \tag{6}$$

et les moyennes partielles par classe :

$$\overline{X}^c = \frac{1}{n_c} \sum \{x_i \mid i \in I_c\} \tag{7}$$

$1 \leqslant c \leqslant C$.

On associe maintenant à la variable v, la variable v_π qui est constante sur chacune des classes. La valeur de v sur chacun des éléments de la classe c, est égale à sa moyenne $\overline{X}^c$ sur I_c :

$$(\forall i \in I), v_\pi(i) = \overline{X}^c \; ssi \; i \in I_c \tag{8}$$

Introduisons la variance de la variable v qui se décompose selon la formule bien connue en la somme de la variance intra-classe et de celle inter-classe :

$$\frac{1}{n} \sum_{i=1}^{i=n} (x_i - \overline{X})^2 = \sum_{c=1}^{c=C} \frac{n_c}{n} (\frac{1}{n_c} \sum_{i \in I_c} (x_i - \overline{X}^c)^2) + \sum_{c=1}^{c=C} \frac{n_c}{n} (\overline{X}^c - \overline{X})^2 \tag{9}$$

La variance inter-classe, dernier terme du développement précédent, représente la variance de la variable v_π. L'intensité de discrimination de la partition π par v est donnée par le rapport :

$$Dis(\pi/v) = \frac{var(v_\pi)}{var(v)} \tag{10}$$

On peut aisément voir que le coefficient de corrélation entre les variables v_π et v est donné par :

$$Corr(v_\pi, v) = (\frac{var(v_\pi)}{var(v)})^{1/2} \tag{11}$$

Ainsi, le «rapport de corrélation» est donné par :

$$(Corr(v_\pi, v))^2 = Dis(\pi/v) \tag{12}$$

C'est cette interprétation corrélative du coefficient de discrimination $Dis(\pi/v)$, d'une partition par une variable qui constituera un fil directeur fondamental du développement à suivre.

Dans notre typologie relationnelle des attributs de description [7, 12] l'attribut booléen, que nous noterons ici a et la variable numérique, notée v ci-dessus, font partie d'un même premier type I. En effet, si la variable v définit une *valuation* (on dit encore *pondération*) numérique sur l'ensemble O des objets, l'attribut booléen a définit une valuation logique (en termes de $VRAI, FAUX$) sur O. En codant par 1 et 0 ces deux dernières valeurs, la notion de moyenne se trouve être traduite par la notion de proportion. Notons dans ces conditions par p_a la proportion dans l'ensemble O des objets où a est à $VRAI$ et par p_a^c, celle correspondante dans la classe O_c, $1 \leqslant c \leqslant C$. Les expressions (10) et (12) deviennent :

$$Dis(\pi/a) = \frac{var(a_\pi)}{var(a)} = \frac{\sum_{c=1}^{c=C} \frac{n_c}{n}(p_a^c)^2 - p_a^2}{p_a - p_a^2} \tag{13}$$

$$(Corr(a_\pi, a))^2 = Dis(\pi/a) \tag{14}$$

Maintenant, si l'on considère le tableau de contingence croisant la partition π avec celle en deux classes $VRAI$ et $FAUX$ associée à l'attribut booléen. Si $\phi^2 (= \chi^2/n)$ indique l'indice attaché à un tel tableau, on peut montrer que

$$Dis(\pi/a) = \phi^2 \tag{15}$$

L'adaptation dans le cas de données de contingence s'obtient dès lors qu'on considère la représentation qu'en donne l'analyse factorielle des correspondances. Ainsi, soit un tableau k_{IJ} où I joue le rôle de l'ensemble des objets ou individus, sur lequel se trouve

donnée la partition π à discriminer. À la colonne j ($j \in J$) on associe une variable numérique w^j dont la valeur sur le i-ème objet est la j-ème composante du profil de i à travers J. N'allons pas plus loin et laissons le lecteur retrouver l'expression adéquate et spécifique de $Dis(\pi/w^j)$ où il est tenu compte de la métrique du χ^2 dont se trouve muni l'espace de représentation du nuage de points associé à I [14].

3. Cas où les données sont qualitatives de différents types

3.1. Introduction

La donnée est définie ici par la description d'un ensemble O d'objets au moyen de variables de toutes sortes : booléennes, numériques, qualitatives nominales, qualitatives ordinales, qualitatives préordonnances ou graphes valués.

Nous avons montré dans [10, 16] comment la méthode de classification ascendante hiérarchique $\mathcal{AVL}$ permet selon un principe unique la classification d'objets décrits par des attributs de types les plus divers. On se trouve intéressé ici par une seule partition sur l'ensemble O des objets que nous continuons à noter (voir (2)) :

$$\pi(O) = \{O_c \mid 1 \leqslant c \leqslant C\} \tag{16}$$

qu'on peut récolter à un niveau «pertinent» de l'arbre des classifications. On peut repérer ce dernier au moyen d'un critère d'adéquation appelé «Statistique globale des niveaux» [7, 8, 15, 16]. Nous allons maintenant nous intéresser au cas de la discrimination de $\pi(O)$ par une variable qualitative quelle que soit la sémantique ou structure de proximité dont se trouve muni l'ensemble des valeurs de la variable. Désignons par w^q une telle variable et par :

$$\underline{E} = \{1, 2, ..., e, ..., E\} \tag{17}$$

l'ensemble de ses valeurs. Ainsi w^q est une application de l'ensemble O des objets dans $\underline{E}$ associant à l'objet codé i, $1 \leqslant i \leqslant n$, la catégorie e à laquelle il appartient. w^q induit déjà une partition sur O qu'on peut noter :

$$\chi^q(O) = \{O_e \mid 1 \leqslant e \leqslant E\} \tag{18}$$

où $O_e = (w^q)^{-1}(e)$ est formé de l'ensemble des objets pour lesquels la valeur de w^q est e. De plus, le cas très général dans lequel nous nous situons est celui d'une variable relationnelle binaire valuée. En d'autres termes, on se donne un graphe valué sur $\underline{E}$ traduisant la structure de proximité (posée par l'expert). Désignons par :

$$W^q = \{w_{ef} \mid (e, f) \in \underline{E} \times \underline{E}\} \tag{19}$$

la valuation d'un tel graphe. w_{ef} représente une valeur numérique qui indique l'intensité de la similarité sur le couple de *valeurs* (on dit également *modalités* ou

catégories) (e, f) de la variable w^q. Le cas où une telle valuation est symétrique par rapport à (e, f) est très fréquent. Cependant, nous dépasserons ce cadre pour considérer le cas antisymétrique ($w(e, f) = -w(f, e)$) et celui, le plus général.

Dans ce contexte donnons le codage adopté de la variable préordonnance qui, précisément, intervient dans l'application. Dans ce cas la donnée est un préordre total sur l'ensemble que nous notons $\underline{E}^{(2)}$ des couples :

$$\underline{E}^{(2)} = \{(e, f) \mid 1 \leqslant e \leqslant f \leqslant E\} \tag{20}$$

Cet ensemble comprend $E(E + 1)/2$ couples de valeurs et concerne la comparaison en termes de similarité ordinale des paires de modalités en comprenant l'association entre une modalité et elle-même. On peut ainsi considérer l'exemple suivant où :

$$\underline{E} = \{1, 2, 3\} \tag{21}$$

avec la préordonnance :

$$12 < 13 \sim 23 < 33 < 22 < 11 \tag{22}$$

où ef avec $1 \leqslant e \leqslant f \leqslant 3$, indique le couple (e, f).

Le préordre total sur $\underline{E}^{(2)}$ est codé en utilisant la fonction «rang moyen» r_m. Ainsi, on obtient dans le cas précédent le tableau des valeurs :

TABLEAU 1

e/f	1	2	3
1	6	1	2.5
2	1	5	2.5
3	2.5	2.5	4

La formule générale donnant la fonction de «rang moyen» pour coder un préordre total sur un ensemble fini est classique. L'ensemble en question est ici $\underline{E}^{(2)}$ et la somme des rangs est égale à :

$$Somrang = \frac{1}{8}E(E + 1)(E^2 + E + 2) \tag{23}$$

Ainsi, dans l'exemple précédent, $Somrang$ vaut 21. Finalement ici :

$$\{w_{ef} = r_m(e, f) \mid (e, f) \in \underline{E}^{(2)}\} \tag{24}$$

Dans ce cadre très général il est aisé de coder les variables qualitatives nominale ou ordinale, de toute sorte [14].

Les généralisations ou extensions de l'analyse des données procèdent par analogie cohérente des points de vue formel et statistique. Nous allons appliquer ce principe pour passer du cas de la discrimination d'une partition π sur l'ensemble des objets par une variable numérique ou, plus généralement unaire au cas où la variable est qualitative d'un type quelconque définissant une relation binaire valuée de similarité sur l'ensemble $\underline{E}$ des catégories de la variable.

Nous avons vu que l'indice de discrimination d'une partition π sur l'ensemble des objets par une variable numérique v prend la forme du carré d'un coefficient de corrélation entre d'une part, la variable v et d'autre part, une variable issue de v et indicatrice de la partition. Comme nous l'avons déjà mentionné dans l'introduction c'est cette forme corrélative que nous allons précisément généraliser.

À cette fin nous allons commencer par rappeler [6, 7, 9, 12, 13] une forme non paramétrique et combinatoire du coefficient de corrélation entre variables numériques; soit entre deux valuations unaires sur l'ensemble O des objets. Nous considérons ensuite une forme de généralisation d'un tel coefficient au cas de la comparaison de deux valuations binaires; en montrant d'ailleurs que le cas unaire peut être interprété dans le contexte binaire (paragraphe 3.2). Nous pourrons déduire au paragraphe 3.3 le cas de la comparaison de deux variables catégorielles valuées. Le cas qui nous intéresse est celui où l'une des variables catégorielles est nominale, puisqu'associée à une partition.

Dans [20] on propose une normalisation du codage conduisant à une forme synthétique et élégante de calcul des coefficients d'association entre variables relationnelles symétriques, pourvu que ces dernières donnent une même valeur maximale à la comparaison entre une catégorie et elle-même. Nous nous contenterons, dans la situation structurelle considérée, pour des raisons de clarté, de prendre ici des expressions plus directes [6, 7, 9, 12, 13].

3.2. Corrélation entre deux valuations unaires ou binaires

3.2.1. Le cas unaire

Soit (v, w) un couple de variables numériques sur l'ensemble des objets. Désignons par :

$$\{(x_i, y_i) \mid 1 \leqslant i \leqslant n\} \tag{25}$$

la suite des valeurs de (v, w) sur la suite des objets conformément à $I = \{1, 2, ..., i, ..., n\}$:

$$x_i = v(o_i) \text{ et } y_i = w(o_i) \tag{26}$$

$1 \leqslant i \leqslant n$.

Nous définissons l'indice «brut» d'association entre v et w sous la forme :

$$s(v, w) = \sum_{i=1}^{i=n} x_i y_i \tag{27}$$

On lui associe un indice brut aléatoire sous l'une ou l'autre des trois formes suivantes qui sont équivalentes; c'est-à-dire, de même loi de probabilité :

$$s(v, w^*) = \sum_{i=1}^{i=n} x_i y_{\sigma(i)} \tag{28}$$

$$s(v^*, w) = \sum_{i=1}^{i=n} x_{\tau(i)} y_i \tag{29}$$

$$\text{et} \quad s(v^*, w^*) = \sum_{i=1}^{i=n} x_{\tau(i)} y_{\sigma(i)} \tag{30}$$

où σ et τ sont deux permutations aléatoires indépendantes prises uniformément au hasard dans l'ensemble des $n!$ permutations sur I.

Compte tenu de l'équivalence entre les trois indices aléatoires précédents, considérons alors le premier d'entre eux, $s(v, w^*)$, on a pour l'espérance mathématique et la variance les expressions suivantes :

$$\mathcal{E}(s(v, w^*)) = n\mu(v)\mu(w) \tag{31}$$

$$\text{et} \quad var(s(v, w^*)) = \frac{n^2}{n-1} var(v) var(w) \tag{32}$$

où μ et var désignent la moyenne et la variance sur I.

L'indice $s(v, w)$ statistiquement normalisé (centré et réduit) se met sous la forme :

$$Q(v, w) = \sqrt{n-1} \times \frac{\frac{1}{n} s(v, w) - \mu(v)\mu(w)}{\sqrt{var(v) var(w)}} = \sqrt{n-1} \times \rho(v, w) \tag{33}$$

où ρ est le coefficient de corrélation.

Le coefficient de corrélation $\rho(v, w)$ s'obtient à partir de $Q(v, w)$ au moyen d'une normalisation formelle que nous disons *géométrique* [9, 13, 20] :

$$\rho = \frac{Q(v, w)}{\sqrt{Q(v, v) Q(w, w)}} \tag{34}$$

C'est bien cette forme que nous retiendrons lorsqu'il s'agit de discriminer une partition π (sur l'ensemble des objets) par une variable qualitative w^q, sous la forme :

$$(Corr(w^q, \pi))^2 \tag{35}$$

3.2.2. Le cas binaire

Relativement à la comparaison de deux variables w et z, définissant cette fois-ci deux relations binaires valuées sur O, l'indice brut $s(w,z)$ correspondant à celui (27) s'écrit :

$$s(w,z) = \sum_{(i,i')\in I^{[2]}} w(i,i')z(i,i') \tag{36}$$

où $I^{[2]}$ est l'ensemble des couples d'éléments de I à composantes distinctes.

Les trois indices bruts aléatoires de même loi correspondants à (28),(29) et (30) se mettent sous la forme :

$$s(w,z^*) = \sum_{(i,i')\in I^{[2]}} w(i,i')z(\sigma(i),\sigma(i')) \tag{37}$$

$$s(w^*,z) = \sum_{(i,i')\in I^{[2]}} w(\tau(i),\tau(i'))z(i,i') \tag{38}$$

$$\text{et} \quad s(w^*,z^*) = \sum_{(i,i')\in I^{[2]}} w(\tau(i),\tau(i'))z(\sigma(i),\sigma(i')) \tag{39}$$

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

Compte tenu de la forme générale adoptée du coefficient de corrélation (*cf.* (34)), il y a lieu de déterminer pour former un coefficient de type Q, l'espérance mathématique et la variance de l'indice aléatoire (*cf.* (37)) :

$$s(w,z^*) = \sum_{(i,j)\in I^{[2]}} w(i,j)z(\sigma(i),\sigma(j)) \tag{40}$$

où σ est une permutation aléatoire dans l'ensemble muni d'une probabilité uniforme des $n!$ permutations sur I.

L'espérance mathématique se met sous la forme :

$$\mathcal{E}(s(w,z^*)) = n(n-1)\mu(w)\mu(z) \tag{41}$$

où μ indique la moyenne sur $I^{[2]}$. Ainsi,

$$\mu(w) = \frac{1}{n(n-1)} \sum_{(i,j)\in I^{[2]}} w(i,j) \tag{42}$$

$$\text{et } \mu(z) = \frac{1}{n(n-1)} \sum_{(i,j)\in I^{[2]}} z(i,j) \tag{43}$$

L'extension de la variance a été étudiée sous différentes formes dans [6, 9, 13, 19, 20], compte tenu notamment de la spécificité des relations à comparer. Nous ferons ici une mise au point et un développement directement liés à l'application (*cf.* § 4), mais restant de portée très générale.

L'expression de la variance serrant directement la nature combinatoire du calcul [10, 13] nécessite l'introduction des ensembles suivants d'indexation où des lettres différentes indiquent des indices distincts :

$$\begin{aligned} I^{[2]}, D = \{[(i,j),(i,j)]\}, \quad & E = \{[(i,j),(j,i)]\} \\ G_1 = \{[(i,j),(i,k)]\}, \quad & G'_1 = \{[(i,j),(h,i)]\} \\ G_2 = \{[(i,j),(h,j)]\}, \quad & G'_2 = \{[(i,j),(j,k)]\} \\ & H = \{[(i,j),(h,k)]\} \end{aligned} \tag{44}$$

Nous désignons d'autre part par $n^{[r]} = n(n-1)(n-2)...(n-r+1)$, la r-ème puissance factorielle de n. On a :

$$\begin{aligned} var(s(w,z^*) = & \frac{1}{n^{[2]}}(\sum_{I^{[2]}} w_{ij}^2)(\sum_{I^{[2]}} z_{ij}^2) + \frac{1}{n^{[2]}}(\sum_{I^{[2]}} w_{ij}w_{ji})(\sum_{I^{[2]}} z_{ij}z_{ji}) \\ & + \frac{1}{n^{[3]}}(\sum_{G_1} w_{ij}w_{ik})(\sum_{G_1} z_{ij}z_{ik}) + \frac{1}{n^{[3]}}(\sum_{G'_1} w_{ij}w_{hi})(\sum_{G'_1} z_{ij}z_{hi}) \\ & + \frac{1}{n^{[3]}}(\sum_{G_2} w_{ij}w_{hj})(\sum_{G_2} z_{ij}z_{hj}) + \frac{1}{n^{[3]}}(\sum_{G'_2} w_{ij}w_{jk})(\sum_{G'_2} z_{ij}z_{jk}) \\ & + \frac{1}{n^{[4]}}(\sum_{H} w_{ij}w_{hk})(\sum_{H} z_{ij}z_{hk}) - [\frac{1}{n^{[2]}}(\sum_{I^{[2]}} w_{ij})(\sum_{I^{[2]}} z_{ij})]^2 \end{aligned} \tag{45}$$

Une distinction importante concerne les cas symétrique et antisymétrique. Très précisément, le cas symétrique s'exprime par :

$$(\forall (i,j) \in I^{[2]}) \quad w_{ij} = w_{ji} \quad \text{et} \quad z_{ij} = z_{ji} \tag{46}$$

alors que le cas antisymétrique est défini par :

$$(\forall (i,j) \in I^{[2]}) \quad w_{ij} = -w_{ji} \quad \text{et} \quad z_{ij} = -z_{ji} \tag{47}$$

L'expression ci-dessus devient :

$$var(s(w,z^*) = \frac{2}{n^{[2]}}(\sum_{I^{[2]}} w_{ij}^2)(\sum_{I^{[2]}} z_{ij}^2)$$
$$+ \frac{4}{n^{[3]}}(\sum_{G} w_{ij}w_{ik})(\sum_{G} z_{ij}z_{ik}) + \frac{1}{n^{[4]}}(\sum_{H} w_{ij}w_{hk})(\sum_{H} z_{ij}z_{hk})$$
$$- [\frac{1}{n^{[2]}}(\sum_{I^{[2]}} w_{ij})(\sum_{I^{[2]}} z_{ij})]^2 \qquad (48)$$

où G (resp. H) est l'ensemble des tri-uplets $[i,j,k]$ (resp. quadruplets $[i,j,h,k]$) à composantes mutuellement distinctes.

Ces expressions ont le mérite de la clarté formelle mais l'inconvénient de la complexité calcul qui, avec H, est, dans le cas de graphes valués les plus généraux, de complexité n^4. Toutefois, le passage à une forme de complexité quadratique est aisé dans les cas où les valuations à associer sont toutes les deux symétriques (*cf.* (46)). Dans le cas symétrique, on obtient la célèbre formule de Mantel [19] conçue dans un tout autre contexte de régression et que nous ignorions [6]. Dans ce dernier article, nous considérions une expression clairement combinatoire et statistique d'une extension proposée dans [5], qui s'inspirait de [3]; ce dernier travail étant pensé dans le contexte très particulier de la comparaison des coefficients de Spearman et de Kendall.

Les paramètres qu'on introduit pour l'expression «Mantel» de la variance de l'indice brut aléatoire (dans le cas symétrique) sont :

$$A_1 = (\sum_{I^{[2]}} w_{ij})^2,\ A_2 = \sum_{i\in I}(\sum_{j\in I-\{i\}} w_{ij})^2,\ A_3 = \sum_{I^{[2]}} w_{ij}^2$$
$$B_1 = (\sum_{I^{[2]}} z_{ij})^2,\ B_2 = \sum_{i\in I}(\sum_{j\in I-\{i\}} z_{ij})^2,\ B_3 = \sum_{I^{[2]}} z_{ij}^2 \qquad (49)$$

On a :

$$\sum\{w_{ij}w_{ik} \mid ((i,j),(i,k)) \in G\} = (A_2 - A_3)$$
$$\text{et} \quad \sum\{z_{ij}z_{ik} \mid ((i,j),(i,k)) \in G\} = (B_2 - B_3) \qquad (50)$$

D'autre part, dans les cas symétriques ou antisymétrique on a :

$$\sum\{w_{ij}w_{hk} \mid ((i,j),(h,k)) \in H\} = A_1 - 4A_2 + 2A_3$$
$$\text{et} \quad \sum\{z_{ij}z_{hk} \mid ((i,j),(h,k)) \in H\} = B_1 - 4B_2 + 2B_3 \qquad (51)$$

On obtient ainsi dans le cas symétrique :

$$var(s(w,z^*)) = \frac{2}{n^{[2]}}A_3B_3 + \frac{4}{n^{[3]}}(A_2-A_3)(B_2-B_3) + \frac{1}{n^{[4]}}(A_1-4A_2+2A_3)(B_1-4B_2+2B_3) - \frac{1}{(n^{[2]})^2}A_1B_1 \quad (52)$$

Ce type d'expression se simplifie encore dans le cas antisymétrique où $A_1 = B_1 = 0$ et où la somme sur H se réduit à la somme sur $I^{[2]}$ au carré. L'expression de la variance devient :

$$var(s(w,z^*)) = \frac{2}{n^{[2]}}A_3B_3 + \frac{4}{n^{[3]}}(A_2-A_3)(B_2-B_3) \quad (53)$$

Il y a lieu de noter que l'indice de corrélation entre deux variables numériques v et w peut être interprété en termes de comparaison entre deux relations binaires valuées et antisymétriques.

L'indice brut prend la forme :

$$s'(v,w) = \sum_{(i,j)\in I^{[2]}} (x_i - x_j)(y_i - y_j) \quad (54)$$

L'indice brut aléatoire correspondant à (37) s'écrit :

$$s'(v,w^*) = \sum_{(i,j)\in I^{[2]}} (x_i - x_j)(y_{\sigma(i)} - y_{\sigma(j)}) \quad (55)$$

Un calcul simple montre que

$$s'(v,w^*) = 2n\sum_{i\in I} x_i y_{\sigma(i)} - 2n^2\mu(v)\mu(w) \quad (56)$$

Suite au développement précédent (formules(41) et (53)) on obtient :

$$\mathcal{E}(s'(v,w^*)) = 0 \quad (57)$$

$$var(s'(v,w^*)) = \frac{4n^4}{n-1}var(v)var(w) \quad (58)$$

Comme nous l'avons mentionné on pourra trouver dans [20] des expressions synthétiques et élégantes pour différents cas de figure ; mais à condition qu'ils puissent correspondre à l'un des deux cas : symétrique ou antisymétrique, et qu'on adopte dans le cas de la comparaison entre variables qualitatives, la même valeur maximale de l'association entre une catégorie et elle-même. Les expressions que nous considérons sont les plus générales par rapport à la nature des relations à comparer. D'autre part, elles ont le mérite d'être plus directement explicites et notamment, de pouvoir distinguer dans le cas de la comparaison entre variables qualitatives, une variation de l'association entre une catégorie et elle-même, selon la catégorie. Cette propriété est importante pour l'application.

3.3. Comparaison de deux variables catégorielles valuées

Nous avons déjà introduit la notion de variable catégorielle valuée (voir paragraphe 3.1 autour des expressions (18) et (19). Désignons par w^q et w^p les deux variables à associer.

$$\underline{E} = \{1, 2, ..., e, ..., E\} \tag{59}$$

$$\underline{C} = \{1, 2, ..., c, ..., C\} \tag{60}$$

codent respectivement les ensembles de valeurs ou catégories des variables w^q et w^p. Indiquons par :

$$W^q = \{w_{ef} \mid (e, f) \in \underline{E} \times \underline{E}\} \tag{61}$$

$$\text{et} \quad W^p = \{v_{cd} \mid (c, d) \in \underline{C} \times \underline{C}\} \tag{62}$$

les deux valuations que nous supposons symétriques, notamment compte tenu de l'application qui va suivre :

$$(\forall (e, f) \in \underline{E} \times \underline{E}), w_{ef} = w_{fe} \tag{63}$$

$$\text{et} \quad (\forall (c, d) \in \underline{C} \times \underline{C}), v_{cd} = v_{dc} \tag{64}$$

Ainsi les deux relations binaires valuées induites sur I sont elles-mêmes symétriques. Ces dernières se trouvent exprimées comme suit :

$$(\forall (i, j) \in I^{[2]}), w(i, j) = w(w^q(i), w^q(j)) \tag{65}$$

$$\text{et} \quad (\forall (i, j) \in I^{[2]}), v(i, j) = v(w^p(i), w^p(j)) \tag{66}$$

Dans ces conditions, on se trouve dans le cadre de l'application, pour le calcul de la variance, de la formule de Mantel (*cf.* (52)) qu'il importe de compacter, compte tenu de la catégorisation que suppose les variables w^q et w^p.

Désignons par π et χ les partitions sur O induites par les variables v et w :

$$\pi = \{O_{c.} = v^{-1}(c) \mid 1 \leqslant c \leqslant C\} \tag{67}$$

où $O_{c.}$ est l'ensemble des objets pour lesquels la valeur de la variable v est c,

$$\chi = \{O_{.e} = w^{-1}(e) \mid 1 \leqslant e \leqslant E\} \tag{68}$$

où $O_{.e}$ est l'ensemble des objets pour lesquels la valeur de la variable w est e.

Nous désignons par $I_{c.}$ (resp. $I_{.e}$) le sous ensemble de I codant $O_{c.}$ (resp. $O_{.e}$), $1 \leqslant c \leqslant C$, $1 \leqslant e \leqslant E$.

Nous indiquons par :

$$\{n_{c.} \mid 1 \leqslant c \leqslant C\} \tag{69}$$

$$\text{et par} \quad \{n_{.e} \mid 1 \leqslant e \leqslant E\}, \tag{70}$$

respectivement, la suite des cardinaux des classes de la partition π et de celle χ. Il y a lieu maintenant de considérer la table de contingence de croisement des deux partitions π et χ :

$$\{n_{ce} = card(O_{c.} \cap O_{.e}) \mid (c, e) \in \underline{C} \times \underline{E}\} \tag{71}$$

à C lignes et E colonnes.

On peut décomposer l'ensemble d'indexation I conformément à la partition croisée $\pi \wedge \chi$:

$$I = \{I_{ce} = I_{c.} \cap I_{.e} \mid (c, e) \in \underline{C} \times \underline{E}\} \tag{72}$$

Il en résulte une décomposition de l'ensemble $I^{[2]}$ des couples d'indices différents sous la forme :

$$\begin{aligned} I^{[2]} = &\sum \{I_{ce}^{[2]} \mid (c, e)\} + \sum \{I_{ce} \times I_{cf} \mid ((c, e), (c, f))\} \\ &+ \sum \{I_{ce} \times I_{de} \mid ((c, e), (d, e))\} + \sum \{I_{ce} \times I_{df} \mid ((c, e), (d, f))\} \end{aligned} \tag{73}$$

où des lettres différentes indiquent des indices différents.

Dans ces conditions, l'indice brut correspondant à (36) se met sous la forme :

$$\begin{aligned} s(v, w) = &\sum \{n_{ce}(n_{ce} - 1) v_{cc} w_{ee} \mid (c, e)\} \\ &+ \sum \{n_{ce} n_{cf} v_{cc} w_{ef} \mid ((c, e), (c, f))\} \\ &+ \sum \{n_{ce} n_{de} v_{cd} w_{ee} \mid ((c, e), (d, e))\} \\ &+ \sum \{n_{ce} n_{df} v_{cd} w_{ef} \mid ((c, e), (d, f))\}\} \end{aligned} \tag{74}$$

où des lettres différentes indiquent des indices différents.

L'expression de la moyenne (*cf.* (41)) devient :

$$\mathcal{E}(s(v,w^*)) = \frac{1}{n^{[2]}}\Big(\sum_{c=1}^{c=C} n_{c.}^{[2]} v_{cc} + \sum_{1\leqslant c\neq d\leqslant C} n_{c.} n_{d.} v_{cd}\Big) \times \Big(\sum_{e=1}^{e=E} n_{.e}(n_{.e}-1) w_{ee} + \sum_{1\leqslant e\neq f\leqslant E} n_{.e} n_{.f} w_{ef}\Big) \qquad (75)$$

où $x^{[2]} = x(x-1)$.

Les expressions associées à la valuation $W^{[p]}$ (*cf.* (62)) des éléments composants de la formulation de la variance selon Mantel (*cf.* (52)) deviennent :

$$A_1 = \Big(\sum_{c=1}^{c=C} n_{c.}^{[2]} v_{cc} + \sum_{1\leqslant c\neq d\leqslant C} n_{c.} n_{d.} v_{cd}\Big)^2,$$

$$A_2 = \sum_{c=1}^{c=C} n_{c.}\Big((n_{c.}-1) v_{cc} + \sum_{\{d|d\neq c\}} n_{d.} v_{cd}\Big)^2$$

$$\text{et} \quad A_3 = \sum_{c=1}^{c=C} n_{c.}^{[2]} v_{cc}^2 + \sum_{1\leqslant c\neq d\leqslant C} n_{c.} n_{d.} v_{cd}^2. \qquad (76)$$

On développera de la même façon, relativement à la valuation W^q, les expressions correspondantes à A_1, A_2 et A_3, B_1, B_2 et B_3.

On pourra ainsi fournir la formulation compactée de la variance exprimée dans (52) selon Mantel. Ayant déjà donné une telle formulation pour la moyenne (*cf.* (75)), on pourra obtenir les coefficients centrés et réduits $Q(v,w)$, $Q(v,v)$ et $Q(w,w)$ de même conception que (33) :

$$Q(v,w) = \frac{s(v,w) - \mathcal{E}(s(v,w^*))}{\sqrt{var(s(v,w^*))}}, \qquad (77)$$

$$Q(v,v) = \frac{s(v,v) - \mathcal{E}(s(v,v^*))}{\sqrt{var(s(v,v^*))}}, \qquad (78)$$

$$Q(w,w) = \frac{s(w,w) - \mathcal{E}(s(w,w^*))}{\sqrt{var(s(w,w^*))}}, \qquad (79)$$

De la sorte et conformément à (34) nous obtenons notre coefficient de corrélation entre deux variables qualitatives valuées v et w qui induisent deux graphes valués sur l'ensemble O des objets :

$$Corr(v,w) = \frac{Q(v,w)}{\sqrt{Q(v,v)Q(w,w)}} \qquad (80)$$

Un tel coefficient est égal à 1 dans le cas où la structure induite par v est identique à celle induite par w Dans la pratique des données réelles il est compris entre -1 et $+1$. Il reste d'ailleurs bien enserré entre ces valeurs dont il se rapproche -comme on le verra dans l'application- très difficilement. Cependant, il existe des situations mathématiques très particulières où le coefficient peut dépasser l'unité. Il s'est ainsi agi de la comparaison entre deux partitions π et χ où π est une partition en deux classes de même taille et où χ s'obtient à partir de π en découpant l'une de ses classes en classes de même taille. Ainsi χ est, d'une façon très particulière, plus fine que π. Cette constatation résulte directement d'une expérimentation effectuée par A. Guénoche dans une étude de comparaison des comportements de différents coefficients de comparaison entre partitions (communication personnelle). Maintenant, s'il s'agit d'avoir un coefficient nécessairement compris entre -1 et $+1$ quelle que soit la situation mathématique à laquelle on se trouve confrontée et qui reste très près de l'esprit du précédent coefficient, on peut proposer :

$$Corr'(v,w) = \frac{Q(v,w)}{max\{Q(v,v)Q(w,w)\}} \tag{81}$$

Néanmoins, ce dernier coefficient a un caractère dissymétrique et ne fait pas jouer le même rôle à $Q(v,v)$ et à $Q(w,w)$. Nous lui préférons $Corr(v,w)$.

La situation qui nous intéresse particulièrement est celle de la comparaison entre une partition et une variable catégorielle. Il s'agit donc d'un cas particulier de celle considérée ci-dessus où la valuation W^p devient :

$$v_{cd} = \begin{cases} 1 & \text{si } d = c \\ 0 & \text{si } d \neq c \end{cases} \tag{82}$$

$1 \leqslant c, d \leqslant C$.

Ainsi, nous répondons au problème de la discrimination d'une partition π par une variable catégorielle valuée w, à partir de la détermination de $Corr(\pi, w)$, conformément à l'expression générale (12). On se reportera à [14] pour le développement spécifique des expressions calculs nécessaires à cette détermination.

4. Application à des données génotypiques. Le logiciel v-class

4.1. Le problème posé, les données, leur nature et leur structure

Le génotype d'un individu est associé à son ADN qui comprend la suite de ses gènes. Plus exactement, la suite dont chaque composante est définie par deux copies d'un même gène, provenant respectivement des deux parents. À une position donnée (dite *loci* dans le langage des biologistes) d'une copie donnée du gène est inclus un nucléotide (ou base azotée) dans l'ensemble $\{A, C, G, T\}$ des quatre nucléotides. L'énorme majorité des locis est occupée par le même nucléotide pour les deux copies du même gène. Cependant, pour certains, une mutation peut avoir lieu. Un SNP (Single

Nucleotid Polymorphism) correspond au changement d'une seule base azotée dans un gène. Tandis que la plupart de ces mutations sont silencieuses, certaines peuvent jouer un rôle dans l'apparition d'un déséquilibre au niveau de l'organisme, ou encore être responsables du développement d'une maladie. Ainsi par exemple en est-il du SNP $C282Y$ relativement au métabolisme du fer.

Pour un SNP donné et relativement aux deux copies du même gène portées par un même individu, deux valeurs du SNP ou « allèles » sont à considérer. Si la mutation est de la forme $C \rightarrow T$, les configurations suivantes peuvent être envisagées pour un sujet donné : CC, TT et CT. Pour le premier et le deuxième cas, l'individu est dit homozygote (pour l'allèle C dans le premier cas et pour celui T dans le second); alors qu'il est dit hétérozygote dans le troisième cas. Relativement à un *loci* concerné par un SNP nous créons une variable qualitative à 3 valeurs codées $1, 2$ et 3 :

- 1, si l'individu est homozygote pour l'allèle le plus rare du gène concerné;
- 2, si l'individu est homozygote pour l'allèle le plus fréquent du gène concerné;
- 3, si l'individu est hétérozygote.

Avant d'aller plus loin précisons la notion de fréquence pour les deux allèles qu'on notera X et Y d'un SNP qui admet la mutation $X \rightarrow Y$. L'échantillon d'individus sur lequel on suppose travailler et qui contient l'information détermine un ensemble O de taille n. La variable qualitative SNP va déterminer une partition en trois classes qu'on peut noter :

$$\{O_{XX}, O_{YY}, O_{XY}\} \tag{83}$$

où $O_{UV}(U, V = X \text{ ou } Y)$ est l'ensemble des individus où la valeur de l'un des allèles est U et la valeur de l'autre, V. En désignant par n_{xx}, n_{yy} et n_{xy} les cardinaux respectifs de O_{XX}, O_{YY} et O_{XY}, les fréquences relatives de l'allèle X et de celui Y sont donnés par les formules :

$$f_X = \frac{2n_{xx} + n_{xy}}{2n} \tag{84}$$

$$\text{et} \quad f_Y = \frac{2n_{yy} + n_{xy}}{2n} \tag{85}$$

Pour chercher à capturer l'information sémantique la plus riche derrière l'ensemble des valeurs de la variable SNP sur un individu, nous avons codé cette dernière en termes de variable préordonnance. La similarité ordinale que nous avons supposée sur l'ensemble suivant des couples de valeurs :

$$\underline{E}^{(2)} = \{(e, f) \mid 1 \leqslant e \leqslant f \leqslant 3\} \tag{86}$$

$$\text{est} \qquad 12 < 13 \sim 23 < 33 < 22 < 11 \tag{87}$$

où ef, $1 \leqslant e \leqslant f \leqslant 3$, indique le couple (e, f) (*cf.* (20), (21), (22)).

En effet, en l'absence de toute connaissance *a priori*, deux individus homozygotes pour l'allèle le plus rare, sont considérés comme se ressemblant le plus

relativement à cet attribut SNP. Par ordre décroissant du degré de ressemblance (en lisant de droite à gauche (22) ou (87)), vient directement ensuite le cas de deux individus homozygotes pour l'allèle le plus fréquent. Puis le cas de deux individus tous les deux hétérozygotes; en effet, ces deux derniers sont naturellement plus associés que ne le sont un individu homozygote et un individu hétérozygote. Enfin, deux individus homozygotes pour des allèles différents sont considérés les plus dissemblables.

Ce préordre total sur l'ensemble $\underline{E}^{(2)}$ des couples de valeurs (*cf.* (20),(22)) est codé au moyen de la fonction «rang moyen» (voir la table 1 du paragraphe 3.1). La variable SNP sera donc interprétée comme une variable catégorielle valuée (*cf.* § 3.3.). Dans cette étude on considère 8 variables SNPs où ces derniers sont localisés à différents loci sur le gène de la transferrine; la suite des valeurs sur les deux copies («allèles») du même gène chez un patient, définira son génotype.

Le problème général est de relier statistiquement le génotype d'un individu à son bilan martial qui permet d'apprécier la charge en fer du sang. Cinq variables numériques décrivent ce bilan :

- Ferritine (taux en $\mu g/litre$)
- Fer ($\mu mo/litre$)
- Coefficient de saturation de la transferrine (%)
- Transferrine ($grammes/litre$)
- Coefficient de fixation de la transferrine ($\mu mo/litre$)

L'analyse s'est effectuée sur une population «normale» à partir de laquelle nous avons pu extraire un échantillon de 306 individus pour lesquels chacune des 13 variables (8 SNPs et 5 mesures cliniques déterminant le bilan martial) a une valeur présente pour chacun des 306 individus.

4.2. Les traitements effectués et les résultats obtenus

4.2.1. Introduction

Le caractère «normal» de la population étudiée rend difficile la mise en évidence de l'influence d'un profil génotypique sur un profil martial, une fois ces deux derniers dégagés. Cette difficulté est déjà importante même dans le cas où la population étudiée comporte un sous ensemble formé de patients malades d'une surcharge en fer.

L'approche choisie ainsi a consisté à déterminer un couple de partitions (π, χ) sur l'ensemble des individus où π est une «bonne» classification par similarité de comportement à partir du bilan martial et où χ est une «bonne» classification par similarité génotypique à travers les $SNPs$. Il s'agira ensuite de mettre en correspondance les deux partitions au moyen d'indices adéquats.

Maintenant, dans la mesure où une classe C^{χ} de la partition χ se trouve significativement associée à une classe C^{π} de la partition π, on cherchera d'abord à distinguer C^{π} relativement aux variables quantitatives du bilan martial. On déterminera ensuite quelles sont les variables $SNPs$ qui ont un comportement spécifique sur

la classe C^{χ}. Ce comportement particulier sera validé au moyen d'un coefficient de discrimination de la forme $(Corr(\chi, w))^2$ ou plus simplement $Corr(\chi, w)$ (*cf.* (80)) où w est la variable préordonnance associée à un SNP (*cf.* § 4.1 ci-dessus).

L'outil de classification choisi pour déterminer chacune des deux partitions π et χ est le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens) [17, 18]. Il s'agit d'une méthode de CAH (Classification Ascendante Hiérarchique) dont la portée est extrêmement générale par rapport à la structure mathématique du tableau des données qu'il s'agisse d'une description d'objets élémentaires, de classes ou de concepts. La méthode permet aussi bien la classification de l'ensemble des lignes que des colonnes du tableau des données en tenant intimement compte de la structure mathématique de ce dernier.

L'algorithme produit une hiérarchie ordonnée par finesse décroissante de classifications emboîtées. Chaque niveau de la hiérarchie détermine une partition de l'ensemble organisé. La méthode comprend en son sein la reconnaissance des niveaux pertinents ou « significatifs » de l'arbre des classifications à partir de l'évolution le long de la suite des niveaux d'un critère d'adéquation (entre la partition et la similarité sur les données). Ce dernier est dit « Statistique globale des niveaux » . Alors que la « Statistique locale des niveaux » permet la détection des noeuds « significatifs » de l'arbre des classifications [7, 8, 15]. Relativement à la suite des fusions de classes que suppose une CAH, un niveau « significatif » correspond à un maximum local (un pic) observé de la distribution de la « Statistique globale » le long de la suite des niveaux. Il définit (constatation expérimentale) un état d'équilibre dans la synthèse. La « Statistique locale » associée à un niveau l est définie par le taux d'accroissement de la Statistique globale entre les niveaux $l-1$ et l. Un noeud « significatif » est associé à un maximum local observé (un pic) de la distribution de la Statistique locale le long de la suite des niveaux. Il définit (constatation expérimentale) une complétion significative d'une classe. Relativement au tableau suivant concernant l'une des classifications hiérarchiques considérées sur l'ensemble des 306 sujets, le niveau le plus significatif est celui 280, il ne correspond pas à un noeud significatif, qui lui, s'est produit au niveau précédent 279. Néanmoins, la cohérence de la classification a continué à augmenter entre les niveaux 279 et 280. Par contre, pour ce qui est du niveau 289, chacune des deux statistiques des niveaux présente un extrémum. Il s'agit d'un niveau globalement « significatif » . D'autre part, dans la mesure où un seul nouveau noeud apparaît à ce niveau, ce dernier est « significatif » . En effet, il peut arriver que plusieurs nouveaux noeuds émergent à un niveau donné. Ce phénomène résulte du fait que plusieurs paires de classes peuvent réaliser à un niveau donné la plus petite dissimilarité entre classes en présence; et, notre programme de CAH réunit toutes ces paires en « même temps » [11]. Dans le cas concerné ici 300 niveaux d'agrégations ont été distingués, alors que l'ensemble comporte 306 éléments.

TABLEAU 2
Exemple de fichier chavl.lis et de choix d'un niveau de condensation

STATISTIQUES DES NIVEAUX			
	NIVEAU	STATISTIQUE GLOBALE	STATISTIQUE LOCALE
	1	0.9117	0.9117
	2	1.2918	0.3801
	3	1.5854	0.2935
	4	1.8344	0.2490
	5	2.0550	0.2206
1 MAXIMUM	6	2.4415	0.3865
	. . .	. . .	. . .
	276	37.4787	1.8617
27 MAXIMUM	277	39.6873	2.2086
	278	41.6923	2.0050
28 MAXIMUM	279	44.1730	2.4807
	280	44.7177	0.5447 <- max de la Stat. glob.
	281	42.2183	-2.4994
29 MAXIMUM	282	42.9365	0.7182
	283	40.3859	-2.5507
30 MAXIMUM	284	39.3361	-1.0498
	285	35.7743	-3.5618
	286	34.2544	-1.5199
	287	34.8169	0.5625
	288	37.6634	2.8465
31 MAXIMUM	289	43.1313	5.4679
	290	43.0571	-0.0742
	291	42.8137	-0.2434
	292	34.1511	-8.6626
	293	29.8805	-4.2706
	294	33.9193	4.0388
32 MAXIMUM	295	44.1939	10.2747
	296	22.3379	-21.8560
	297	11.8411	-10.4968
33 MAXIMUM	298	14.8079	2.9668
	299	-2.0275	-16.8354
	300	-21.5396	-19.5122

4.2.2. Tableau de contingence

Deux classifications ascendantes hiérarchiques CHAVL sur l'ensemble O des individus ont été opérées. La première ne prend en compte que la description par les 5 variables numériques définissant le bilan martial et la seconde ne considère que la description par les 8 variables préordonnances $SNPs$. Nous avons considéré pour chacun des arbres de classification le niveau le plus significatif au sens de la Statistique globale des niveaux. Nous avons ainsi obtenu une partition π en 6 classes de O relativement à la description numérique et une partition χ en 8 classes de O relativement à la description qualitative préordonnance par les variables $SNPs$. On obtient dans ces conditions le tableau de contingence suivant où la partition π est portée en lignes et où la partition χ est portée en colonnes.

TABLEAU 3

π/χ	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	$n_{c.}$
1	5	12	7	6	11	9	2	6	58
2	8	6	7	5	4	10	3	7	50
3	12	5	4	3	4	7	4	2	41
4	9	4	5	1	8	5	3	1	36
5	7	14	6	6	7	13	4	7	64
6	10	8	6	8	6	13	4	2	57
$n_{.d}$	51	49	35	29	40	57	20	25	306

Son expression formelle est :

$$\{n_{cd} \mid 1 \leqslant c \leqslant C, 1 \leqslant d \leqslant D\} \tag{88}$$

où n_{cd} est le nombre de sujets à l'intersection de la c-ème classe de la partition π et de la d-ème classe de la partition χ. $C = 6$ et $D = 8$ dans notre cas. $n_{c.}$, porté à la c-ème ligne de la marge colonne est le cardinal de la c-ème classe de la partition π, $1 \leqslant c \leqslant C$. $n_{.d}$, porté à la d-ème colonne de la marge ligne est le cardinal de la d-ème classe de la partition χ, $1 \leqslant d \leqslant D$.

À ce tableau nous associons le tableau des «contributions orientées au χ^2» [7] (Chap. 3). Il se présente sous la forme :

$$\{ct(c,d) = \frac{n_{cd} - \frac{n_{c.}n_{.d}}{n}}{\sqrt{\frac{n_{c.}n_{.d}}{n}}} \mid 1 \leqslant c \leqslant C, 1 \leqslant d \leqslant D\} \tag{89}$$

où on a pour le χ^2 associé au tableau de contingence :

$$\chi^2 = \sum_{c=1}^{c=C} \sum_{d=1}^{d=D} (ct(c,d))^2 \tag{90}$$

Dans le cas du tableau précédent il s'agit d'un $\chi^2_{(35)}$ à $5 \times 7 = 35$ degrés de libertés. Sa valeur approximative est 32.0. Il est loin d'être significatif. Il correspond à une légère dépendance positive; puisque par consultation des tables on a :

$$Pr(\chi^2_{(35)} > 32.0) > Pr(\chi^2_{(30)} > 32.0) > Pr(\chi^2_{(30)} > 34.8) = 0.25 \qquad (91)$$

TABLEAU 4
Contributions orientées au χ^2

π/χ	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	–1.5	0.89	0.14	0.21	1.24	–0.55	–0.92	0.58
2	–0.12	0.71	0.54	0.12	–0.99	0.22	–0.15	1.44
3	1.98	–0.61	–0.32	–0.45	–0.59	–0.23	0.81	–0.74
4	1.22	–0.73	0.43	–1.31	1.52	–0.66	0.42	–1.13
5	–1.12	1.17	–0.49	–0.03	–0.47	0.31	–0.09	0.77
6	0.16	–0.37	–0.20	1.12	–0.53	0.73	0.14	–1.23

Néanmoins le caractère plus ou moins prononcé d'une valeur de $ct(c, d)$ peut être situé par rapport à une variable aléatoire normale centrée et réduite. En effet, dans le cadre d'une hypothèse probabiliste d'indépendance entre les deux partitions, respectant les proportions $n_{c.}/n$ et $n_{.d}/n$, $ct(c, d)$ peut être considérée – pour n «assez grand» – comme la réalisation d'une variable aléatoire normale centrée et réduite [4, 7, 12].

Ajoutons deux tableaux très synthétiques concernant d'une part la classification en 6 classes à travers les variables numériques et d'autre part, la classification en 8 classes concernant la description qualitative. Dans le premier cas on fournit les moyennes de chaque variable dans chacune des classes et dans le second cas, il s'agit de valeurs modales.

TABLEAU 5
Valeurs moyennes pour le cas numérique

	FRT	**FER**	**CS**	**TF**	**CF**
Cl1	81	15	23	33	65
Cl2	216	15	25	30	60
Cl3	28	12	18	34	68
Cl4	68	23	44	26	52
Cl5	61	13	18	37	74
Cl6	97	19	32	30	60

TABLEAU 6
Valeurs moyennes générales

	FRT	**FER**	**CS**	**TF**	**CF**
moyenne	93	16	26	32	64

TABLEAU 7
Valeurs modales pour le cas qualitatif

SNPs	**282**	**23**	**47**	**267**	**369**	**524**	**558**	**571**
Cl1	2	1	1	3	2	2	3	2
Cl2	2	3	3	2	3	2	3	3
Cl3	2	3	3	2	3	2	2	3
Cl4	2	3	1	2	2	3	3	2
Cl5	2	3	3	2	2	3	2	2
Cl6	2	2	3	2	3	3	2	3
Cl7	2	2	1	2	2	1	2	2
Cl8	2	2	2	2	1	2	2	1

Très peu de cases correspondent à une contribution orientée au χ^2 suffisamment sensible. Les valeurs positives sont plus «parlantes» que celles négatives. Ainsi, la première classe issue de la description par les variables qualitatives $SNPs$ $(d = 1)$ est positivement liée à la troisième classe issue de la description numérique définissant le bilan martial $(c = 3)$. En effet, la valeur de la contribution orientée $ct(3, 1)$ atteint la valeur 1.98. Si on observe cette classe $(d = 1)$, on s'aperçoit qu'en son sein, deux $SNPs$ ($TF23$ et $TF47$) sont majoritairement représentés par leur allèle le plus rare au niveau de l'échantillon global. Il s'agit d'un point singulier qui interpelle. D'autre part, la classe $(c = 3)$ représente un groupe de sujets ayant un bilan en fer relativement bas en comparaison avec la moyenne de l'échantillon global (notamment, une ferritine de $28\mu g/l$ pour cette classe, contre $93\mu g/l$ pour l'échantillon total). Un lien significatif entre l'allèle rare des $SNPs$ $TF23$ et $TF47$ et un tel profil martial mérite d'être confirmé sur un échantillon plus important. Un autre lien positif se signale entre la classe $(d = 8)$ issue de la description qualitative et celle $(c = 2)$ issue de la description numérique $(ct(2, 8) = 1.44)$. En induisant de la même manière que précédemment, on peut conclure à une augmentation de la ferritine en présence de l'allèle rare des SNPs $TF369$ et $TF571$.

Ainsi et pour nous résumer, notre raisonnement a consisté à se rendre compte si une hyper ou une hypoferritinémie se manifestait dans une classe d'individus formée à partir de la description martiale (numérique). Dans ce cas, on se demande si une telle classe se trouve associée à une classe d'individus issue de la description génotypique, donc qualitative (contribution orientée au χ^2 importante). Si oui, on se trouve alors particulièrement intéressé par les SNPs dont les fréquences relatives des deux allèles

(*cf.* (84) et (85)) se trouvent inversées dans la dernière classe génotypique, par rapport à leurs valeurs sur l'ensemble total des individus.

4.2.3. Indices de discrimination

Maintenant, on peut vouloir lier l'importance d'une variable qualitative SNP à l'intensité avec laquelle elle discrimine la classification retenue sur l'ensemble des individus. Cette classification est – pour le tableau 8 ci-dessous – récoltée au niveau le plus «significatif» de l'arbre des classifications sur l'ensemble des individus décrits par les 8 variables SNPs. Les indices retenus sont ceux $(Corr(\pi, w))^2$ (noté ci-dessous «Coef2») et $Corr(\pi, w)$ (noté ci-dessous «Coef»)[*cf.* (80)]. L'indice $Q(\pi, w)$ est celui noté ci-dessous «Qpiw» On constate que parmi les deux variables $TF23$ et $TF47$ ayant un comportement distinctif sur la classe $(d = 1)$, c'est le SNP $TF23$ qui est le plus structurant. Par ailleurs, en dehors de ce dernier, c'est bien parmi les SNPs : $TF369$, $TF524$ et $TF571$, ayant le rôle le plus tangible dans la discrimination de la partition, qu'on retrouve les deux : $TF369$ et $TF571$, dont le comportement se distingue sur la classe $(d = 8)$.

TABLEAU 8
Indices de discrimination; cas qualitatif

Variable	Coef2	Coef	Qpiw
C282Y	0.001242	0.035242	2.464258
TF23	0.299792	0.547532	102.530493
TF47	0.025318	0.159117	32.678138
TF267	0.010863	0.104225	7.555742
TF369	0.143158	0.378363	48.753293
TF524	0.283431	0.532382	59.081957
TF558	0.022718	0.150727	13.273294
TF571	0.266094	0.515843	69.381109

Maintenant, compte tenu du caractère «normal» de la population dont un échantillon de taille modérée (306) est soumis à l'analyse, on ignore quel est le facteur structurant la classification des individus décrits par leur génotype. Cette structure peut être conditionnée par les différents types de normalité. Curieusement, le SNP $C282Y$ dont la mutation peut permettre à la maladie de l'hémachromatose de se manifester, a et de loin le plus faible pouvoir discriminant relativement à la classification qui a émergé. Le SNP qui se distingue également; mais de façon moins nettement accentuée, par un faible pouvoir discriminant, est $TF267$. De façon surprenante ces deux attributs s'associent ensemble à l'avant dernier niveau; mais avec un noeud «significatif» dans l'arbre des classifications sur l'ensemble des variables SNPs (voir Figure 1). Y a-t-il lieu de chercher du côté de ce dernier SNP?

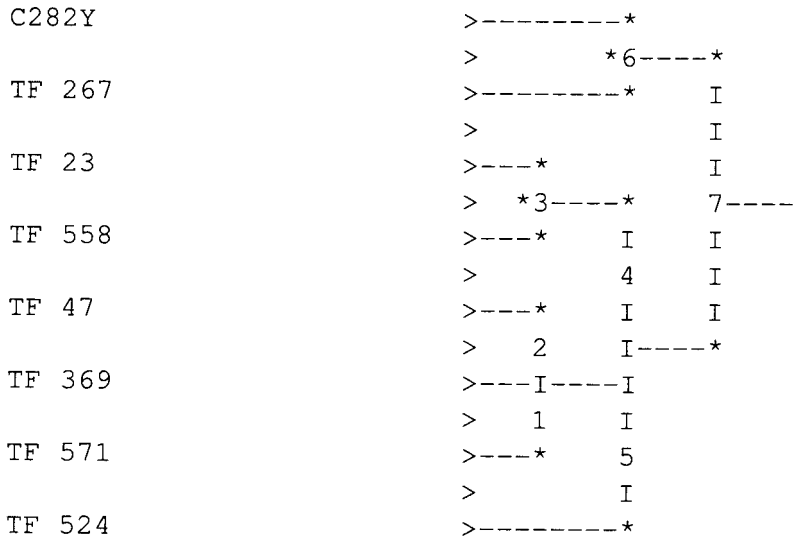

FIGURE 1
Classification hiérarchique des variables SNPs

Ce dernier arbre des classifications sur l'ensemble des variables qualitatives SNPs est extrêmement difficile à interpréter. Il n'en est pas du tout de même pour l'arbre des classifications sur l'ensemble des variables numériques établissant le bilan martial (voir Figure 2).

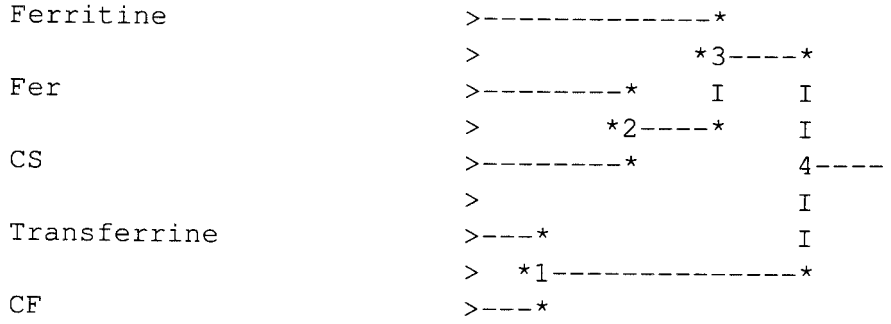

FIGURE 2
Classification hiérarchique des variables du bilan martial

À la différence du cas où les variables sont qualitatives, les coefficients de discrimination des différentes variables [*cf.* (11) et (12)] sont tous plus clairement positifs (voir tableau 9). Ils vont en quelque sorte dans le même sens par rapport à la typologie dégagée.

TABLEAU 9
Coefficients de discrimination; cas des variables numériques

	Coef2	Coef
Variable 1	0.615100	0.784283
Variable 2	0.614265	0.783750
Variable 3	0.624831	0.790462
Variable 4	0.625052	0.790602
Variable 5	0.782992	0.884868

À cet égard, un dernier résultat plein d'intérêt va toujours dans le même sens. Pour ce dernier, deux aspects nouveaux sont intervenus. D'une part, un nouvel ensemble de variables SNPs a été ajouté au précédent, passant ainsi de 8 à 20. D'autre part et surtout, une version spécifique CHAVLh de CHAVL a été utilisée. Elle permet la classification d'objets décrits par un mélange de variables hétérogènes. Il s'agit ici d'un mélange de variables numériques et de variables préordonnances.

On commencera par remarquer que les valeurs de $Coef2$ (représentant le carré du coefficient de corrélation) d'un précédent traitement où seuls les 8 précédents attributs SNPs sont intervenus (*cf.* tableau 8) sont beaucoup plus notables que ceux concernant les variables SNPs du présent traitement (*cf.* tableau 10). C'est que, dans le premier cas, la classification s'est faite au mieux à partir des seuls attributs qualitatifs. C'est lorsque la classification s'effectue en faisant intervenir sur le même pied d'égalité les attributs qualitatifs et ceux, numériques, que le phénomène du rapport des valeurs ci-dessus mentionné est mis en évidence. Il apparaît ainsi que les variables du bilan martial jouent un rôle beaucoup plus important que les SNPs dans la formation des classes.

Précisément, parmi les variables numériques représentant le bilan martial, c'est la ferritine qui a le plus fort pouvoir classificatoire de l'ensemble des individus, compte tenu de son coefficient le plus élevé. Ceci concorde avec la connaissance en biologie, puisque c'est le taux de ferritine qui est utilisé dans de nombreuses études épidémiologiques pour évaluer la charge en fer d'un patient.

Le fait que les SNPs aient un coefficient de corrélation assez faible se comprend à partir de la concomitance de trois raisons différentes. Premièrement, cette population est considérée «normale» , c'est-à-dire, qu'elle est composée d'individus *a priori* sains, dont le bilan martial, même s'il diffère d'un individu à l'autre, reste dans la norme. Il est donc difficile de mettre en évidence un lien entre une augmentation de la charge en fer et un SNP particulier. D'autre part, un autre facteur serait que les SNPs étudiés n'ont pas un rôle tangible dans un éventuel déséquilibre du métabolisme du fer. Enfin, le type de coefficient de «corrélation» proposé est plus rigoureux que dans le cas numérique, en ce sens qu'il s'écarte plus difficilement de 0.00.

TABLEAU 10
Coefficients de discrimination; cas des variables numériques/préordonnances

Variable	Coef2	Coef	Qpiw
Variable 1	0.515617	0.718065	-------
Variable 2	0.286276	0.535047	-------
Variable 3	0.142609	0.377636	-------
Variable 4	0.098659	0.314100	-------
Variable 5	0.374676	0.612107	-------
Variable 6	0.000857	0.029271	2.07220
Variable 7	0.007744	0.087999	17.58031
Variable 8	0.003300	0.057447	12.547731
Variable 9	0.002462	0.049619	3.772409
Variable 10	0.022951	0.151496	21.848252
Variable 11	0.001986	0.044569	5.049361
Variable 12	0.015162	0.123136	11.345709
Variable 13	0.024573	0.156757	23.377040
Variable 14	0.015085	0.122821	11.316533
Variable 15	0.015539	0.124655	21.230163
Variable 16	0.014207	0.119193	10.675383
Variable 17	0.000255	0.015969	1.108735
Variable 18	0.021895	0.147971	28.237780
Variable 19	0.002510	0.050096	3.478194
Variable 20	0.016842	0.129776	22.969455
Variable 21	0.001558	0.039472	7.484577
Variable 22	0.007386	0.085945	10.108278
Variable 23	0.007399	0.086018	8.437941
Variable 24	0.000326	0.018063	-1.242385
Variable 25	0.000137	0.011715	1.983601

4.2.4. Conclusion

Nous avons pu montrer l'application d'un outil d'analyse classificatoire des données adapté et puissant, permettant de mettre en évidence sur une population d'individus des profils martiaux et des profils génotypiques, en cherchant à les associer.

Cependant, les résultats au niveau biologique sont assez ténus et ne peuvent, comme nous nous y attendions, être très significatifs. Comme nous l'avons déjà exprimé, la population étudiée était composée de sujets *a priori* normaux d'un point

de vue biologique. D'autre part, la présence de données manquantes (qui ne sont pas acceptées par CHAVL) nous a amenés à ne retenir qu'une partie seulement des individus d'origine, réduisant ainsi l'échantillon à une taille de l'ordre de 300.

Toutefois, cet essai par les perspectives qu'il offre est stimulant. Il invite d'abord à une large expérimentation par un accroissement important de la taille de l'échantillon. Dans ces conditions, il importe alors d'injecter un sous ensemble d'individus surchargés en fer. Il faut néanmoins s'attendre que même dans ce cas, la mise en évidence de l'influence du phénomène génotypique soit difficile, compte tenu déjà du faible pourcentage de développement de la maladie de l'hémachromatose en cas de mutation du SNP $C282Y$.

Une autre approche, passant par un programme d'haplotypage, peut fournir une autre facette dans l'interprétation biologique. Un tel programme permet de décomposer l'ensemble des génotypes d'un échantillon de sujets en leurs haplotypes respectifs. Ainsi, chaque génotype se trouve décomposé en les deux haplotypes des deux parents dont il est issu. On peut craindre un poucentage d'erreurs non nécessairement négligeable avec l'usage d'un tel programme. Pour une solution fournie, la structure mathématique des données est plus simple que ci-dessus, où on avait à gérer directement les génotypes globaux. Cependant, si on veut tenir compte d'un risque d'erreur dans la prédiction de la répartition des deux allèles d'un même SNP entre les deux haplotypes parentaux, la structure des données devient complexe. Néanmoins, la méthodologie classificatoire de l'Analyse de la Vraisemblance des Liens, est en mesure de prendre en charge cette forme de données.

4.3. Le logiciel v-class

Le but de ce programme est la mise en oeuvre des calculs des indices de discrimination présentés ci-dessus des classes d'une partition d'un ensemble d'objets (les objets représentent des individus dans notre cas). Il s'agit de fournir, pour chaque variable descriptive, quelle que soit sa nature, un corespondant à un coefficient de corrélation au carré entre la variable et une partition donnée de l'ensemble des objets issue d'un algorithme de classification. Trois types de variables sont actuellement pris en considération : le numérique, le booléen et le qualitatif préordonnance.

Ce programme est actuellement accroché au logiciel CHAVL. On applique une telle CAH (Classification Ascendante Hiérarchique) sur l'ensemble des sujets pour recueillir à certains de ses «niveaux significatifs» une partition. En fait, on considère la partition la plus significative (celle réalisant la plus grande valeur de la «Statistique globale»); puis, de part et d'autre, on considère le niveau qui réalise la plus grande valeur de la statistique mentionnée. Trois partitions sont ainsi récoltées et «expliquées» variable par variable. Ainsi, le nombre de fichiers résultats est multiplié par trois.

Les calculs les plus élaborés concernent les coefficients mettant en correspondance une partition de l'ensemble des individus – telle que l'une de celles recueillies à un niveau donné de l'arbre des classifications – et une variable catégorielle préordonnance (*cf.* § 3.3.). La structure mathématique de base pour le calcul est un tableau de contingence croisant la partition récoltée à un niveau donné de l'arbre des classifications et la partition induite par la variable qualitative; qui, ici, possède 3 valeurs.

On calcule ensuite des statistiques combinatoires liées à la distribution de la variable préordonnance à l'intérieur des classes de la partition et également entre les classes. On peut noter ce tableau de contingence sous la forme :

TABLEAU 11
Croisement entre la partition et la variable qualitative w

π/w	$\mathbf{e=1}$	$\mathbf{e=2}$	$\mathbf{e=3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$classe\ c$	n_{c1}	n_{c2}	n_{c3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Les deux arguments du calcul sont d'une part, ce dernier tableau de contingence (où n_{ce} est le nombre d'objets de la classe c possédant la modalité e de la variable w) et d'autre part, la valuation de la préordonnance définie par la variable w. Des variables temporaires ont dû être introduites ; notamment, pour le calcul de l'indice brut $s(\pi, w)$. Un calcul intermédiaire aboutit à la détermination de $Q(\pi, w)$. Ce coefficient est normalisé pour obtenir $Corr(\pi, w)$ dont on produit le carré $Corr(\pi, w)^2$. Un tel coefficient est porteur de l'importance d'une variable dans la formation des classes de la partition π. L'ensemble de ces expressions est détaillé dans [14], elles correspondent à la situation définie dans (82). Les résultats sont édités sous la forme d'un fichier résultat précisant le numéro de la variable ainsi que certains calculs intermédiaires.

Pour une variable donnée, la décomposition du coefficient global $Corr(\pi, w)^2$ a à être considérée donnant ainsi pour chaque classe d'une partition l'importance de la variable pour sa formation [14].

5. Conclusion générale

Au terme de cette étude, nous avons bâti et validé un outil général pour mesurer l'importance du rôle d'une variable descriptive relativement à une classification d'objets sur lesquels la variable a été mesurée et cela quelle que soit sa nature.

Pour la conception de cet outil une variable est interprétée en termes de relation valuée sur l'ensemble des objets. Une base fondamentale consiste alors en l'élaboration d'un coefficient d'association entre variables relationnelles. C'est une méthode générale que nous avons mise au point et développée qui est considérée. Elle permet d'intégrer les cas classiques des variables unaires et retrouve les coefficients associés pour ces variables, qui peuvent par ailleurs admettre une représentation géométrique. La méthode a été spécifiée dans les différents cas de figure pour ce qui concerne les variables relationnelles binaires. Le cas d'intérêt est celui des variables catégorielles où une valuation compare les couples de catégories (§ 3.3.). Le cas qui nous a plus particulièrement concerné est celui où l'une des variables catégorielles est nominale en étant associée à la partition produite par un algorithme

de classification. Pour ce cas, le logiciel v-class a été élaboré. Il est appelé à être repris et développé en recouvrant l'aspect local qui concerne la reconnaissance de la part d'une même classe de la partition à «expliquer» dans la valeur du coefficient global de discrimination [14].

Nous avons validé notre construction en montrant comment procéder dans le cadre d'un problème important de biologie moléculaire, reliant génotype et phénotype du métabolisme du fer. À cet égard, on se reportera à notre conclusion du paragraphe 4 (§ 4.2.4.). Il est clair que la portée applicative de notre travail peut être très largement étendue. Par ailleurs, sur le plan purement méthodologique, les travaux à mener peuvent correspondre à des comparaisons avec d'autres types de coefficients ou concerner des valeurs incertaines ou entachées d'erreur.

Remerciements

• Nous tenons à remercier le Professeur Yves Deugnier, Service des Maladies du Foie et Centre d'Investigation clinique, INSERM 0203 (CHU de Rennes) et le Professeur Jean Mosser de l'UMR-CNRS 6061 «Génétique et Développement» pour les discussions que nous avons pu avoir et la mise à disposition des données.

• Nous remercions également Benjamin Enriquez (Professeur de Mathématiques à l'Université Louis Pasteur de Strasbourg) pour les échanges que nous avons pu avoir sur des aspects calculs et qui ont pu préciser nos associations d'idées.

• Nous remercions vivement Pierre Cazes pour sa lecture très attentive qui a conduit à améliorer la présentation et à corriger quelques points.

Références

[1] ADOUE V. (septembre 2003), Élaboration d'un logiciel d'explication de classes pour une classification de données génotypiques, Rapport DESS-CCI, IFSIC, Université de Rennes 1.

[2] CELEUX G., DIDAY E., LECHEVALLIER Y. and RALAMBONDRAINY H. (1998), *Classification automatique des données*, Dunod.

[3] DANIELS H.E. (1994), The relation between measures of correlation in the universe of sample permutations, *Biometrika*, vol. 33, p. 129-135.

[4] LANCASTER H.O. (1969), *The chi squared distribution*, Wiley.

[5] LECALVE G. (1976), Un indice de similarité pour des variables de types quelconques, *Statistique et Analyse des Données*, (01-02), p. 39-47.

[6] LERMAN I.C. (1977), Formal analysis of a general notion of proximity between variables, In J.R. Barra, editor, *Congrès Européen des Statisticiens, Grenoble, Septembre 1976*, North Holland.

[7] LERMAN I.C. (1981), *Classification et analyse ordinale des données*, Dunod.

[8] LERMAN I.C. (1983), Sur la signification des classes issues d'une classification automatique, *In* J. Felsenstein, editor *Numerical Taxonomy*, p. 179-198, Springer-Verlag.

[9] LERMAN I.C. (juillet 1987), Analyse de la forme limite de coefficients statistiques d'association entre variables relationnelles, Rapport de recherche, INRIA.

[10] LERMAN I.C. (1987), Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème de consensus en Classification, *Revue de Statistique Appliquée*, XXXV(2), p. 39-60.

[11] LERMAN I.C. (1989), Formules de réactualisation en cas d'aggrégation multiple, *RAIRO, série R.O.*, vol. 23 n° 2, p. 151-163.

[12] LERMAN I.C. (1992), Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles I, *Mathématique Informatique & Sciences Humaines*, (118), p. 35-52.

[13] LERMAN I.C. (1992), Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles II, *Mathématique Informatique & Sciences Humaines*, (119), p. 75-100.

[14] LERMAN I.C. (décembre 2004), Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques, Publication Interne 1652, IRISA-INRIA.

[15] LERMAN I.C. and GHAZZALI N. (1991), What do we retain from a classification tree, In E. Diday and Y. Lechevallier, editors, *Symbolic-Numeric data analysis and learning*, p. 27-42, Nova Science.

[16] LERMAN I.C. and PETER PH. (octobre 2003), Indice probabiliste de vraisemblance du lien entre objets quelconques; analyse comparative entre deux approches, *Revue de Statistique Appliquée*, LI(1), p. 5-35.

[17] LERMAN I.C., PETER PH. and LEREDDE H. (décembre 1993), Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens). Première partie, *La Revue de Modulad*, (13), p. 33-70.

[18] LERMAN I.C., PETER PH. and LEREDDE H. (juin 1994), Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens). Deuxième partie, *La Revue de Modulad*, (14), p. 63-90.

[19] MANTEL N. (1967), Detection of disease clustering and a generalized regression approach, *Cancer Research*, vol. 27, n° 2, p. 209-220.

[20] OUALI-ALLAH M. (décembre 1991), *Analyse en préordonnances des données qualitatives. Application aux données numériques et symboliques*, Université de Rennes 1.

[21] YOUNESS G. (juillet 2004), *Contributions à une méthodologie de comparaisons de partitions*, Université de Paris 6.