REVUE DE STATISTIQUE APPLIQUÉE

CATHERINE DUPONT-GATELMAND

Une méthode de classification automatique sur variables hétérogènes

Revue de statistique appliquée, tome 27, n° 2 (1979), p. 23-37

<http://www.numdam.org/item?id=RSA_1979__27_2_23_0>

UNE MÉTHODE DE CLASSIFICATION AUTOMATIQUE SUR VARIABLES HÉTÉROGÈNES (*)

Catherine DUPONT-GATELMAND

Attachée de recherche au C.N.R.S. Lamsade – Université PARIS IX

RESUME

Nous proposons une méthode de classification non hiérarchique d'un ensemble d'individus (ou d'objets) caractérisés par des variables hétérogènes (nominales, ordinales ou numériques). Purement descriptive, elle permet la constitution de groupes homogènes dans la population étudiée en respectant la nature de chacune des variables.

Cette méthode est itérative et consiste à effectuer alternativement un processus d'agrégation autour de centres variables et une série d'analyses canoniques pour déterminer les codages : ces derniers ne sont donc pas fixés a priori.

INTRODUCTION

La plupart des méthodes d'analyse des données ont été élaborées pour utiliser un type particulier de données (qualitatives, ordinales ou quantitatives). Or, dans la réalité, le praticien désire souvent utiliser conjointement, à l'aide d'une seule méthode, des données à la fois qualitatives nominales (situation de famille, sexe, catégories socio-professionnelles, etc.), ordinales (préférences pour des produits, attitudes à l'égard de phénomènes sociaux, etc.) et quantitatives (revenu, chiffre d'affaires, etc.).

Afin d'étudier et de traduire de telles observations, il est indispensable de considérer le problème à la base et de s'interroger sur le *codage* des données, inévitable pont entre celles-ci et une méthode de traitement n'acceptant qu'un seul type homogène de variables.

De nombreuses recherches ont déjà été menées sur ce domaine dans diverses directions (cf. SAPORTA [15]). Certaines études portant plus spécifiquement sur le codage des variables qualitatives ont montré que des techniques fort classiques (analyse de la covariance, analyse factorielle discriminante, . . .) n'étaient en fait que des méthodes de codage de variables qualitatives ayant comme objectif l'optimisation d'un critère d'ajustement (R.P. MAC DONALD [13], CAILLIEZ F. et PAGES J.P. [3]). De son côté, M. MASSON [14] a souligné la nécessité de faire de l'analyse non linéaire et a ainsi étudié les liaisons non linéaires des variables numériques au moyen d'une analyse canonique généralisée.

(*) Nous avons donné le nom de "CATY" à cette méthode de Codage Adaptatif en TYpologie.

Mots-clés : **Classification automatique, codage, ordinal, qualitatif**

Par ailleurs, Y. TAKANE, F.W. YOUNG et J. De LEEUW ont systématisé l'intégration des variables qualitatives dans des traitements conçus initialement pour des variables quantitatives. Ils ont ainsi, depuis trois ou quatre ans, mis au point :

– l'analyse en composantes principales avec codage : PRINCIPALS [20] ;

– la régression multiple et l'analyse canonique avec codage : MORALS/ CORALS [21] ;

– l'analyse de la variance avec codage : ADDALS [22] ;

– l'analyse des proximités avec codage : ALSCAL [23] ;

– et plus généralement le projet HOMALS de J. de LEEUW.

Dans cette même direction, M. Tenenhaus [18] a réalisé l'analyse en composantes principales d'un ensemble de variables nominales ou numériques. Toutefois, au contraire de PRINCIPALS, son programme PRINQUAL ne part pas d'une solution arbitraire, converge pratiquement plus rapidement et fournit un seul codage par variable nominale.

Il convient enfin de signaler quelques ouvrages dus à G. SAPORTA [16], J.M. BOUROCHE [2], CAZES, BAUMERDER, BONNEFOUS ET PAGES [4] qui ont également apporté une importante contribution aux problèmes soulevés par le "codage".

1. POSITION DU PROBLEME

On considère un ensemble I d'individus de cardinal n sur lequel on mesure p variables $x_1, \ldots, x_p$ qui peuvent être soit qualitatives nominales ou ordinales, soit quantitatives (J désigne l'ensemble de ces p variables).

Le problème peut être posé de la façon suivante : trouver simultanément la partition de l'ensemble des individus et le codage optimal des données au sens d'un critère qui mesure l'adéquation des individus aux différentes classes et respecte les contraintes de nature des données (qualitatives, ordinales et quantitatives).

1.1. – Quelques rappels sur le "codage"

Les résultats généraux que nous exposons ont déjà été explicités dans de nombreux ouvrages ou publications (cf. BOUROCHE [2] ; CAILLIEZ, PAGES [3] ; CAZES, BAUMERDER, BONNEFOUS, PAGES [4] ; DROUET D'AUBIGNY [7] ; HAYASHI [9] ; MASSON [13] ; SAPORTA [15] ; TENENHAUS [17], [18], YOUNG, De LEEUW, TAKANE [20] à [23]).

On notera $\mathcal{X}_j$ l'ensemble des valeurs possibles d'une variable x_j (application définie sur I et à valeurs dans $\mathcal{X}_j$).

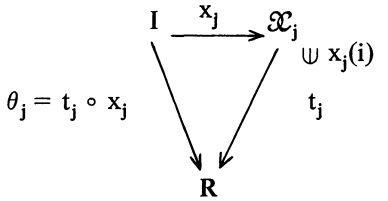
$$x_j : I \to \mathcal{X}_j$$

$$\forall i \in I \quad x_j(i) \in \mathcal{X}_j \quad \text{pour tout j appartenant à J}$$

On pose $|\mathcal{X}_j| = m_j$.

1.1.1. ***Définition du "codage" d'une variable***

Un codage d'une variable x_j est obtenu à partir d'une application $t_j : \mathfrak{X}_j \to \mathbf{R}$.



$$\cup\, \theta_j(i) = t_j(x_j(i))$$

La variable codée $\theta_j = t_j \circ x_j$ est une variable quantitative réelle.

Il y a une infinité de codages possibles et la recherche d'un t_j particulier dépend de l'objectif poursuivi : le critère à optimiser est défini comme on le verra en fonction de la méthode. De plus, des contraintes de codage sont imposées selon la nature des variables considérées.

1.1.2. ***Contraintes selon la nature des variables***

– *Cas d'une variable qualitative nominale*

Soit X_j^h la variable indicatrice associée à la modalité h :

$$X_j^h(i) = \begin{cases} 1 & \text{si } x_j(i) = h \\ 0 & \text{si } x_j(i) \neq h \end{cases}$$

On introduit ainsi m_j variables indicatrices :

$$X_j^1, \ldots, X_j^h, \ldots, X_j^{m_j}.$$

La contrainte imposée par ce type de variable peut être formulée de la façon suivante :

Pour tout i et i′ appartenant à I

$$\left.\begin{array}{l} x_j(i) = x_j(i') \end{array}\right\} \Leftrightarrow \begin{cases} \theta_j(i) = \theta_j(i') \\ (t_j \circ x_j)(i) = (t_j \circ x_j)(i') \end{cases}$$

Autrement dit, à une modalité d'une variable nominale, il ne doit correspondre qu'un seul codage même si celle-ci a été prise par deux individus différents.

– *Cas d'une variable qualitative ordinale*

Le codage d'une variable ordinale peut respecter les contraintes suivantes :

$$\begin{cases} h = h' \Rightarrow t_j(h) = t_j(h') \\ h < h' \Rightarrow t_j(h) \leqslant t_j(h') \end{cases}$$

– *Cas d'une variable quantitative*

L'ensemble $\mathfrak{X}_j$ des valeurs prises par une variable quantitative étant l'ensemble des réels, toute fonction t_j de **R** dans **R** définit un codage de la variable x_j.

On peut toutefois noter que, dans la mesure où l'on désire absolument modifier les données initiales, il serait plus judicieux de choisir pour t_j une fonction monotone croissante qui respecterait ainsi les conditions imposées par les variables ordinales que de prendre, à l'instar de F.W. YOUNG, Y. TAKANE et J. de LEEUW, une fonction polynôminale.

1.1.3. ***Ensembles des codages admissibles pour chaque type de variable***

– *Cas d'une variable qualitative nominale*

On montre (cf. J.P. PAGES [3]) que définir un codage d'une variable qualitative nominale revient à définir une combinaison linéaire de variables indicatrices.

Si X_j désigne la matrice, de dimension (n, m_j), des indicatrices $(X_j^1, \ldots, X_j^h, \ldots, X_j^{m_j})$, on définit ainsi un sous-espace vectoriel de $\mathbf{R}^n$, W_j engendré par les m_j variables indicatrices et le vecteur codage c_j de x_j appartient à :

$$W_j = \{c_j \in \mathbf{R}^n / c_j = X_j\, b_j\,,\, b_j \in \mathbf{R}^{m_j}\}$$

$$b_j \in \mathbf{R}^{m_j} \text{ est un codage de } x_j \qquad b_j = \begin{bmatrix} b_j^1 \\ \cdot \\ \cdot \\ \cdot \\ b_j^h \\ \cdot \\ \cdot \\ \cdot \\ b_j^{m_j} \end{bmatrix}$$

et c_j vérifie les contraintes imposées par la nature nominale de la variable x_j.

– *Cas d'une variable qualitative ordinale*

De même, définir un codage d'une variable ordinale revient à définir une combinaison linéaire de variables indicatrices sous contraintes.

Si X_j est la matrice des indicatrices associée à la variable ordinale x_j, un codage de x_j est un élément du cône convexe $\mathcal{C}_j$ de $\mathbf{R}^{m_j}$.

$$\mathcal{C}_j = \{b_j \in \mathbf{R}^{m_j} / b_j^1 \leqslant \ldots \leqslant b_j^h \leqslant \ldots \leqslant b_j^{m_j}\}$$

X_j est une application définie sur $\mathcal{C}_j$ et à valeurs dans $\mathbf{R}^n$.

On définit alors le cône convexe $\mathbf{C}_j$ de $\mathbf{R}^n$ engendré par les m_j variables indicatrices et le vecteur codage c_j de x_j appartient à :

$$\mathbf{C}_j = \{c_j \in \mathbf{R}^n / c_j = X_j\, b_j\,,\, b_j \in \mathcal{C}_j\}.$$

L'ensemble $\mathbf{C}_j$ des codages est un cône polyédrique convexe (cf. M. TENENHAUS et A. MACQUIN [19]).

– *Cas d'une variable quantitative*

Le codage c_j d'une telle variable est tel que :

$$c_j \in U_j = \{c_j \in \mathbf{R}^n / c_j = t_j(x_j)\}.$$

U_j est un sous-espace vectoriel de $\mathbf{R}^n$.

1.2. **Formulation mathématique du problème**

On munit l'espace des variables de la métrique euclidienne des poids D_n. La matrice associée à l'isomorphisme D_n est :

$$D_n = \begin{pmatrix} p_1 & & & \\ & \ddots & & \\ & & p_i & \\ & & & \ddots \\ & & & & p_n \end{pmatrix} \quad \text{avec} \quad \sum_{i=1}^{n} p_i = 1$$

et on suppose l'espace $\mathbf{R}^p$ muni de la métrique unité.

De plus, on munit l'espace $\mathbf{R}^{np}$ des matrices à n lignes et p colonnes de la métrique définie par :

$$\|A\|^2 = \sum_i \sum_j p_i (a_{ij})^2 = \sum_j \|a_j\|^2_{D_n}$$

où a_{ij} $(1 \leqslant i \leqslant n\ ,\ 1 \leqslant j \leqslant p)$ est le terme général de A,

et a_j le $j^{ème}$ vecteur colonne (n , 1) de A.

Le problème général est de trouver simultanément :

– la partition P de l'ensemble I des individus en q classes (avec q fixé a priori) de manière à grouper les individus les plus semblables dans une même classe (autrement dit de manière à minimiser les inerties intra-classes) ;

– et le codage quantitatif des variables hétérogènes : ce codage doit "s'améliorer" (au sens des moindres carrés).

Les recherches du codage et de la partition s'effectuent en même temps de façon itérative.

On note par :

$P = (P_1, \ldots, P_l, \ldots, P_q)$ la partition en q classes de I

$Y = (y_1, \ldots, y_l, \ldots, y_q)$ la matrice de dimension (n , q) des indicatrices associées à P

$G = (g_1, \ldots, g_j, \ldots, g_p)$ la matrice de dimension (q , p) des centres de gravité des q classes de P

$C = (c_1, \ldots, c_j, \ldots, c_p)$ la matrice de dimension (n , p) des codages des p variables

On cherche alors à résoudre le problème suivant :

$$\min_{(C,G,Y)} \|C - Y G\|^2$$

soit

$$\min \sum_{j=1}^{P} \|c_j - Y g_j\|^2_{D_n} \qquad (1)$$

La résolution de ce problème répondra ainsi aux deux objectifs fixés :

– adéquation du codage aux données initiales ;

– homogénéité des groupes autour de leur centre de gravité.

1.3. Schéma de résolution

La solution est obtenue par l'articulation itérative de deux processus :

– le "processus interne" qui permet l'obtention d'un nouveau codage à partition fixée ;

– le "processus externe" qui permet la formation d'une nouvelle partition à partir de la solution issue du processus interne.

2. LE "PROCESSUS INTERNE" : OBTENTION D'UN NOUVEAU CODAGE A PARTITION FIXEE

A partir de la matrice $\underset{(n,p)}{\overline{C}^0}$ des codages initiaux centrés réduits et de la matrice $\underset{(n,q)}{Y^0}$ associée à P^0 partition initiale de I, le problème consiste à rechercher une nouvelle matrice des codages C^1 vérifiant les contraintes imposées par la nature des différentes variables et une matrice G^1 telles que :

$$\underset{(C,G)}{\text{Min}} \; ||\overline{C} - Y\,G||^2$$

Chaque *phase* du processus interne peut être décomposée en deux étapes :

– recherche d'un codage intermédiaire indépendant du type de variable considérée ;

– recherche d'un nouveau codage fonction de la nature des différentes variables.

Remarque : On ne modifiera pas les variables quantitatives et ainsi, seule l'étude de la transformation des variables qualitatives sera détaillée.

Il convient également de noter que, pour la recherche de ces codages, Y étant fixé, il est équivalent de minimiser (1) ou de minimiser chacun des termes $||c_j - Y\,g_j||_{D_n}$ indépendamment des autres i.e. de rechercher c_j et g_j minimisant la quantité précédente.

2.1. Recherche d'un codage intermédiaire

Nous noterons par $A^0 = (a_1^0, \ldots, a_j^0, \ldots, a_p^0)$ la matrice, de dimension (n , p), des codages intermédiaires et $\Phi^0 = \{Y^0\,g / g \in \mathbf{R}^q\}$ le sous-espace vectoriel engendré par les indicatrices associées à P^0, autrement dit engendré par les colonnes de Y^0.

On définit le codage intermédiaire a_j^0 d'une variable par la projection D_n orthogonale de c_j^0 sur Φ^0. Ainsi $A^0 = Y^0\,G$ est solution du problème :

$$\underset{G}{\text{Min}} \; ||\overline{C}^0 - Y^0\,G||^2$$

On peut remarquer que le projecteur $T^0 = Y^0 (Y^{0'}\,D_n\,Y^0)^{-1}\,Y^{0'}\,D_n$ conserve le centrage. Ainsi, si le codage initial d'une variable x_j est centré, alors son codage intermédiaire a_j^0 est centré.

2.2. Recherche d'un nouveau codage

Le problème est traité variable par variable et on étudie successivement les cas correspondant aux différents types de variable.

2.2.1. *Cas d'une variable qualitative nominale*

On a vu que le codage c_j d'une variable nominale x_j était un élément du sous-espace vectoriel W_j de $\mathbf{R}^n$ défini par :

$$W_j = \{c_j \in \mathbf{R}^n / c_j = X_j\, b_j\,,\, b_j \in \mathbf{R}^{m_j}\}.$$

On s'intéresse en fait à un codage centré réduit $\bar{c}_j$ de x_j. Si $W_j = 1 \oplus \bar{W}_j$, le vecteur codage centré doit être un élément de $\bar{W}_j$.

Le problème est alors de trouver $\bar{c}_j^1$ solution de :

$$\underset{\bar{c}_j \in \bar{W}_j \cap S}{\text{Min}} \; \|\bar{c}_j - a_j^0\|^2$$

avec $S = \{y / \|y\|^2 = 1\}$

ou de manière équivalente, en utilisant un théorème démontré par J. de Leeuw (cf. [5]), il s'agit de trouver c_j^1 solution de :

$$\underset{c_j \in \bar{W}_j}{\text{Min}} \; \|c_j - a_j^0\|^2$$

puis de normer le résultat.

La projection D_n orthogonale de a_j^0 sur $\bar{W}_j \cap S$ donne ainsi

$$c_j^1 = X_j\,(X_j'\,D_n\,X_j)^{-1}\,X_j'\,D_n\,a_j^0 \quad \text{et} \quad \bar{c}_j^1 = c_j^1 / \|c_j^1\|^2.$$

Comme précédemment, on remarque que le projecteur $\tilde{T}_j = X_j\,(X_j'\,D_n\,X_j)^{-1}\,X_j'\,D_n$ conserve le centrage : si a_j^0 est centré, $\bar{c}_j^1$ est bien un élément de $\bar{W}_j$.

2.2.2. *Cas d'une variable qualitative ordinale*

Le codage c_j d'une variable ordinale x_j étant un élément du cône polyédrique $\mathbf{C}_j$ de $\mathbf{R}^n$ défini par :

$$\mathbf{C}_j = \{c_j \in \mathbf{R}^n / c_j = X_j\, b_j\,,\, b_j \in \mathcal{C}_j\},$$

on s'intéresse au codage centré réduit $\bar{c}_j$ de x_j tel que $\bar{c}_j \in \bar{\mathbf{C}}_j$ avec $\mathbf{C}_j = 1 \oplus \bar{\mathbf{C}}_j$.

Le problème est alors de trouver $\bar{c}_j^1$ solution de :

$$\underset{\bar{c}_j \in \bar{\mathbf{C}}_j \cap S}{\text{Min}} \; \|\bar{c}_j - a_j^0\|^2$$

ou de

$$\underset{c_j \in \bar{\mathbf{C}}_j}{\text{Min}} \; \|c_j - a_j^0\|^2$$

et de normer le résultat (cf. théorème de J. de Leeuw précité).

L'algorithme de Kruskal, dans sa version généralisée (cf. [10], [11]), permet d'obtenir le résultat.

La conservation du centrage résultant de la projection sur le cône polyédrique reste vérifiée.

2.3. **Analyse de la solution obtenue à l'issue du "processus interne"**

Le processus interne se définit comme une méthode de moindres carrés alternés, à partition fixée, dans laquelle les deux étapes précédemment décrites : "recherche d'un codage intermédiaire" et "recherche d'un nouveau codage selon

la nature des variables" se répètent successivement. Il correspond ainsi à une application successive de deux opérateurs de projection et converge vers le couple solution $(\bar{c}_j^*, g_j^*)$.

Cette procédure peut être rapprochée de la recherche des vecteurs propres d'une matrice par la méthode de HOTELLING. Ainsi, le codage obtenu(*) à la convergence du processus interne $\bar{c}_j^*$ est vecteur propre associé à la plus grande valeur propre de la matrice $(\tilde{T}_j \circ T^0)$.

En fait, on remarque que le processus interne est comparable à une analyse canonique(*) entre le tableau des indicatrices associées à x_j et celui des indicatrices associées à la partition P^0. Ainsi c_j^1 est le vecteur canonique de $\tilde{T}_j \circ T^0$ associé à la plus grande valeur propre λ^1 et a_j^1 est le vecteur canonique de $T^0 \circ \tilde{T}_j$ associé à λ^1. A ces deux vecteurs canoniques, on fait correspondre respectivement les éléments :

$$b_j^1 \in \mathbf{R}^{m_j^*} \quad \text{et} \quad g_j^1 \in \mathbf{R}^{q^*} \quad \text{tels que} \quad c_j^1 = X_j b_j^1 \quad \text{et} \quad a_j^1 = Y^0 g_j^1 .$$

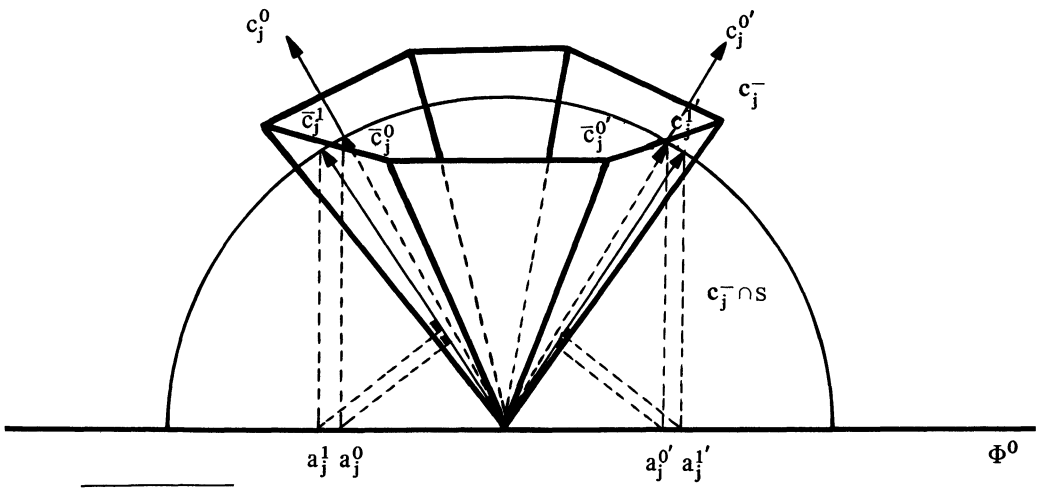
Notons que b_j^1 et g_j^1 sont également les premiers facteurs d'une analyse des correspondances réalisée sur le tableau $Y^{0\prime} X_j$.

Il est à souligner que la condition de centrage-réduction des variables est indispensable si l'on désire éviter la solution triviale, obtenue dès la première itération et égale à la moyenne générale sur l'ensemble de la partition des codages initiaux.

Conséquences

1) Pour toute variable qualitative nominale x_j, son codage $\bar{c}_j^*$ obtenu à la convergence du processus interne est indépendant du codage initial c_j^0 de celle-ci. Fonction des fréquences des différentes modalités dans chaque classe, il dépend de la partition initiale (et du nombre de classes fixé a priori).

2) Cette propriété n'est pas vérifiée pour les variables ordinales : le codage obtenu à la convergence du processus interne n'est pas indépendant du codage initial. En effet, on recherche dans ce cas deux vecteurs appartenant à $\mathbf{C}_j$ et à Φ^0 et formant un angle minimum. Or, la méthode des moindres carrés alternés ne conduit pas à une solution unique. Le contre-exemple suivant le montre. Supposons que la partition a priori P^0 soit une partition en deux classes (P_1^0, P_2^0) ; Φ^0 est alors une droite.



(*) Dans le cas d'une variable qualitative nominale (i.e. quand on n'impose pas de contrainte d'ordre).

On considère deux codages distincts d'une variable ordinale x_j : soit c_j^0 et $c_j^{0'}$ les deux vecteurs de C_j correspondants.

La première étape du processus interne est une projection orthogonale de $\bar{c}_j^0$ et $\bar{c}_j^{0'}$ sur Φ^0 donnant a_j^0 et $a_j^{0'}$.

La seconde étape consiste à projeter orthogonalement a_j^0 et $a_j^{0'}$ sur C_j^- déterminant ainsi un nouveau codage que l'on normalise : $\tilde{c}_j^1$ et $\tilde{c}_j^{1'}$. Une nouvelle projection sur Φ^0 détermine a_j^1 et $a_j^{1'}$ qui, projetés sur $C_j^- \cap S$, redonnent les codages $\tilde{c}_j^1$ et $\tilde{c}_j^{1'}$.

Ainsi, à deux codages initiaux différents d'une même variable x_j, il correspond deux codages $\bar{c}_j^1$ et $\bar{c}_j^{1'}$ distincts.

Contre-exemple numérique

L'ensemble des individus I est de cardinal n = 35 et la partition a priori P^0 en deux classes est tirée au hasard (cf. [8] (annexe 2)). Soit x_i une variable ordinale à 4 modalités :

* $c_j^0 = (1, 2, 3, 4)$ le premier codage a priori.

Le codage obtenu est alors :

$$\bar{c}_j^1 = (-0{,}63246\ ;\ -0{,}63246\ ;\ -0{,}63246\ ;\ 1{,}158114).$$

* $c_j^{0'} = (7, 10, 15, 16)$ le second codage a priori.

Le codage obtenu est alors :

$$\bar{c}_j^{1'} = (-1{,}83712\ ;\ 0{,}54433\ ;\ 0{,}54433\ ;\ 0{,}54433)$$

Ainsi, dans un cas, les trois premières modalités de la variable se trouvent "codées" de la même façon et, dans l'autre cas, ce sont les trois dernières modalités qui possèdent le même "codage".

Remarque : Il est évident qu'en raison du centrage-réduction au début de la procédure, le codage d'une variable ordinale est indépendant des codages fixés a priori si ceux-ci sont identiques à une transformation affine près.

3. LE PROCESSUS EXTERNE : CONSTITUTION D'UNE NOUVELLE PARTITION

A partir de la solution issue du processus interne, on va définir une nouvelle partition P ou, de manière équivalente, une nouvelle matrice d'indicatrices Y associée à P.

Il s'agit ainsi de trouver Y solution du problème :

$$\begin{cases} \underset{Y \in \mathcal{Y}}{\text{Min}} \ \|\bar{C} - Y\,G\|^2 \\ \\ \text{avec } \mathcal{Y} = \{(y_k^i)\ i = 1, \ldots, n \in \{0, 1\} \text{ et } / \sum_{k=1}^{q} y_k^i = 1 \ \ \forall i = 1, \ldots, n\} \end{cases}$$

Si c^i désigne la $i^{ème}$ ligne de C et y^i la $i^{ème}$ ligne de Y ($1 \leqslant i \leqslant n$), $\|C - YG\|^2$ s'écrivant $\sum_{i=1}^{n} p_i \|c^i - y^i\,G\|^2\ p^i$, la recherche de la solution peut donc se faire ligne par ligne.

On a à résoudre pour tout i élément de I

$$\begin{cases} \underset{y^i \in R^q}{\text{Min}} \; ||\bar{c}^i - y^i G||^2 \\ \text{avec} \; \sum_{k=1}^{q} y_k^i = 1 \; \text{et} \; y_k^i \in \{0, 1\} \; \text{pour tout} \; k = 1, \ldots, q \end{cases}$$

La solution est de la forme :

$$y_k^i = 1 \;,\; y_{k'}^i = 0 \quad \text{pour } k' \neq k$$

avec $$||\bar{c}^i - y^i G||^2 = ||\bar{c}^i - g^k||^2 = \underset{k'}{\text{Min}} \; ||\bar{c}^i - g^{k'}||^2$$

g^k (resp $g^{k'}$) étant la $k^{ème}$ (resp $k'^{ème}$) ligne de G.

En d'autres termes, cette phase tend à regrouper les individus les plus semblables.

Les q classes de la nouvelle partition seront telles que :

$$P_k = \{i \in I / ||\bar{c}^i - g^k||^2 \leqslant ||\bar{c}^i - g^{k'}||^2 \;\; \forall \, k' \neq k\}.$$

3.1. Algorithme général

1) *Etape d'initialisation*

On part :

– d'un codage initial des variables établi a priori (pour les variables qualitatives) c^0 qui est centré-réduit devenant ainsi $\bar{c}^0$

– et d'une partition P^0 en q classes, tirée au hasard ou choisie a priori. A cette partition est associée Y^0 la matrice des indicatrices.

2) *Première application du "processus interne"*

Cette phase comporte en alternance deux minimisations au sens des moindres carrés et se déroule à partition fixée P^0.

α) Détermination du codage intermédiaire, résolution du problème :

$$\underset{G}{\text{Min}} \; ||\bar{C} - Y^0 G||^2$$

donnant les solutions $a_j^0 = Y^0 g_j \; \forall \, j \in \{1, \ldots, p\}$.

β) Détermination du nouveau codage (variable par variable) :

$$\underset{\bar{c}_j \in W_j \cap S}{\text{Min}} \; ||\bar{c}_j - a_j^0||^2 \rightarrow \text{cas nominal}$$

$$\underset{\bar{c}_j \in C_j \cap S}{\text{Min}} \; ||\bar{c}_j - a_j^0||^2 \rightarrow \text{cas ordinal}$$

Les étapes α) et β) se répètent jusqu'à la convergence du "processus interne", analogue à une analyse canonique effectuée entre Y^0 et X_j tableau des indicatrices associé à chaque variable qualitative x_j.

Soit $(\bar{c}^1 , g^1)$ la solution obtenue à la convergence de ce processus.

3) *Première application du "processus externe"*

A partir de $(\bar{c}^1 , g^1)$, on cherche à identifier une nouvelle partition P^1 dont les q classes sont construites de manière à être plus homogènes qu'à l'itération précédente.

On cherche donc une nouvelle matrice d'indicatrices Y^1 telle que Y^1 soit solution de :

$$\underset{Y \in \mathcal{Y}}{\text{Min}} \; ||\bar{c} - Y G||^2$$

4) *A l'étape n*

On applique le processus interne à partition fixée P^{n-1}. On obtient ainsi le couple solution $(\bar{c}^n , g^n)$. On applique alors le processus externe pour former une nouvelle partition P^n.

5) *A la convergence,* on a le triplet solution :

$$(\bar{C}^* , G^* , Y^* \text{ ou } P^*).$$

3.2. Etude de la convergence

Appelons F la fonction définie sur $\mathbf{R}^{np} \times \mathbf{R}^{nq} \times \mathcal{Y}$ et à valeurs dans $\mathbf{R}^+$ telle que :

$$F(C , G , Y) = ||\bar{C} - Y G||^2.$$

Montrons que la suite des valeurs prises par cette fonction décroît au cours de l'algorithme et converge vers une limite correspondant à la valeur de la fonction pour le triplet solution $(\bar{C}^* , G^* , Y^*)$.

Montrons en premier lieu que :

$$F(\bar{C}^n , G^n , Y^n) \leqslant F(\bar{C}^n , G^n , Y^{n-1}) \leqslant F(\bar{C}^n , G^{n-1} , Y^{n-1}) \leqslant$$

$$\leqslant F(\bar{C}^{n-1} , G^{n-1} , Y^{n-1}).$$

– $\bar{C}^n$ solution du problème $\underset{C}{\text{Min}} \; ||\bar{C} - A^{n-1}||^2$

ou $\underset{C}{\text{Min}} \; ||\bar{C} - Y^{n-1} G^{n-1}||^2$

$$\Rightarrow F(\bar{C}^n , G^{n-1} , Y^{n-1}) \leqslant F(\bar{C}^{n-1} , G^{n-1} , Y^{n-1})$$

– G^n solution du problème $\underset{G}{\text{Min}} \; ||\bar{C}^n - Y^{n-1} G||^2$

$$\Rightarrow F(\bar{C}^n , G^n , Y^{n-1}) \leqslant F(\bar{C}^n , G^{n-1} , Y^{n-1})$$

– Y^n solution du problème $\underset{Y}{\text{Min}} \; ||\bar{C}^n - Y G^n||^2$

$$\Rightarrow F(\bar{C}^n , G^n , Y^n) \leqslant F(\bar{C}^n , G^n , Y^{n-1}).$$

La suite des valeurs prises par la fonction F est donc décroissante. Comme elle est toujours minorée par zéro (étant formée d'une somme de fonctions positives), elle est convergente.

$$\exists N \, \forall n > N \quad F(\bar{C}^n, G^n, Y^n) = F(\bar{C}^N, G^N, Y^N)$$

La fonction F étant injective (par la définition de G comme matrice des centres de gravité des classes et par la détermination du codage), on a :

$$\exists N \, \forall n > N \quad (\bar{C}^n, G^n, Y^n) = (\bar{C}^N, G^N, Y^N)$$

On montre dans [8] que la suite des valeurs prises par la fonction F converge en un nombre fini d'itérations.

4. UN EXEMPLE

Cet exemple porte sur des données provenant d'un laboratoire de Rhône-Poulenc.

4.1. Nature des données

275 malades ont été examinés très attentivement et des dossiers fort complets ont été constitués : l'information recueillie est conséquente... 250 variables ont été prises en compte. Toutefois, sur les conseils du médecin responsable de l'étude, on en a sélectionné 14 qui lui semblaient particulièrement intéressantes pour une tentative de classification des malades. Elles nous renseignent sur les caractéristiques de la maladie (diagnostic, symptômes de début) et sur l'effet produit par l'absorption du produit T (résultat, tolérance). On observe notamment l'évolution de certains troubles après la prise du médicament (variables 3 à 11).

Les natures des variables sont diverses :

– les variables 1 à 10 sont qualitatives ordinales ;

– les variables 11 à 13 sont qualitatives nominales ;

– la variable 14 est quantitative ; elle est exprimée en milligrammes.

4.2. Le problème

On cherche à former des groupes homogènes de malades qui auraient éventuellement des réactions analogues à la prise du médicament T compte tenu des troubles spécifiques qui les caractérisent. Ceci permettrait peut-être d'établir des contre-indications à l'utilisation de T si l'on découvrait des tolérances mauvaises pour certaines sous-populations.

Les variables sont par nature hétérogènes : on aurait pu, en découpant la variable quantitative, utiliser la méthode des Nuées Dynamiques en prenant une distance du χ^2 sur modalités, mais alors on aurait perdu l'information ordinale apportée par les 10 premières variables. On applique donc la méthode CATY "avec codage adaptatif".

4.3. Analyse des résultats obtenus par une typologie avec codage adaptatif sur variables hétérogènes.

Nous sommes partis d'un tirage au hasard sur la population déterminant ainsi une partition à trois classes.

Nous avons examiné 4 tirages au hasard et nous avons choisi la solution ($\bar{C}$*, G*, P*) qui minimisait la valeur de la fonction F. Le programme donne à la fois la description des classes (repérage des individus) et le codage des modalités des différentes variables.

Analyse de la partition P* : On l'obtient par des tableaux de contingence réalisés entre la partition P* et chacune des variables. On décrit ainsi les trois classes avec leurs caractéristiques respectives sur les variables considérées. On remarque que les trois classes sont bien différenciées par le résultat de l'expérience et la tolérance au médicament :

– *la classe 1* rassemble des individus sur lesquels le médicament T n'a aucun effet bénéfique puisqu'ils conservent les mêmes troubles qu'avant la prise du médicament T. De plus, ils l'ont mal supporté ;

– *la classe 2* est une classe intermédiaire : la tolérance est bonne mais les individus composant cette classe n'étaient atteints que de troubles très légers. On constate toutefois que le médicament n'induit aucun déséquilibre supplémentaire ;

– *la classe 3* comprend des malades plus sérieux et pour lesquels le médicament T est le remède miracle.

Une analyse détaillée des 36 dossiers médicaux caractérisant la première classe permettrait sans doute au médecin d'établir certaines contre-indications à la prise du médicament T.

Quelques compléments d'analyse des données

On s'est intéressé, par souci de comparaison, à l'obtention d'une partition P* de l'ensemble des malades, à l'aide de la méthode classique des Nuées Dynamiques. Toutefois, afin de pouvoir l'appliquer, il fallait a priori "coder" les variables de manière homogène. Disposant de 13 variables qualitatives, on les a transformées en tableaux disjonctifs complets et on a considéré la distance du χ^2 sur modalités entre les individus. Soulignons que l'information apportée par les dix variables ordinales prises en compte était dès lors perdue : il n'y avait plus aucune distinction entre les variables nominales et ordinales.

Après quatre tirages au hasard, on a choisi la partition qui donnait une valeur de la fonction F minimum : l'analyse de celle-ci a mis en évidence des classes moins homogènes et l'interprétation en a été rendue très délicate.

Toujours pour mettre en parallèle les deux méthodes, on a calculé les χ^2 de contingence entre chaque variable et les partitions optimales obtenues dans les deux cas. On a constaté d'une manière générale que ceux-ci étaient beaucoup plus significatifs lorsque le codage des variables était "adaptatif".

Pour conclure sur les Nuées Dynamiques, on est parti dans "CATY" de la solution optimale donnée par les Nuées Dynamiques : cette solution a été améliorée au cours de l'algorithme, dans tous les exemples traités.

A partir du codage homogène des variables par la méthode "CATY", on a effectué une analyse discriminante entre la partition optimale et les variables codées correspondant à la solution obtenue à la convergence de l'algorithme. Le pourcentage d'individus bien classés a été de 94,5 %.

REFERENCES

[1] BARLOW R.E., BARTHOLOMEW D.J., BREMNER J.M., BRUNK H.D. – "Statistical inference under order restrictions". Wiley, 1972.

[2] BOUROCHE J.M. – "Développement de l'utilisation des méthodes d'analyse des données qualitatives". Note de travail n° 26, COREF.

[3] CAILLIEZ F., PAGES J.P. – "Introduction à l'analyse des données". SMASH, 1976.

[4] CAZES P., BAUMERDER A., BONNEFOUS S., PAGES J.P. – "Codage et analyse des tableaux logiques – Introduction à la pratique des variables qualitatives". *Cahier n° 27, BURO,* 1977.

[5] DE LEEUW J. – "A normalized cone regression approach to alternating least squares algorithm". University of Leiden, The Netherlands, 1976.

[6] DIDAY E. – "La méthode des nuées dynamiques et la reconnaissance des formes". *Cahiers de l'IRIA,* 1970.

[7] DROUET D'AUBIGNY G. – "Description statistique des données ordinales : analyse multidimensionnelle". Thèse de 3e cycle, Université de Grenoble, 1975.

[8] DUPONT-GATELMAND C. – "Deux méthodes d'analyse des données qualitatives : typologie et codage adaptatif et modèle de décomposition additive des préférences". Thèse de 3e cycle, Université Paris IX, 1978.

[9] HAYASHI L. – "On the quantification of qualitative data from the mathematico-statistical point of view". *Annals. Inst. Stat. Math 2,* 1950.

[10] KRUSKAL J.B. – "Analysis of factorial experiments by estimating monotone transformation of the data". *Journal of the Royal Statistical Society, Series B,* 27, 1965.

[11] KRUSKAL J.B. – "Monotone regression, continuity and differentiability properties". *Psychometrika,* Tome 36, n° 1, 1971, p. 57-62.

[12] LECHEVALLIER Y. – "Classification automatique optimale sous contrainte d'ordre total". Rapport de recherche n° 200, IRIA-LABORIA, 1976.

[13] MAC DONALD R.P. – "A unified treatment of the weighting problem". *Psychometrika,* 33, 1968, p. 351-381.

[14] MASSON M. – "Analyse non linéaire des données". *Compte rendu à l'Académie des Sciences de Paris,* Tome 278, Série A, 1974, p. 803-806.

[15] SAPORTA G. – "Liaison entre plusieurs ensembles de variables et codages de données qualitatives". Thèse de 3e cycle, Université de Paris VI, 1975.

[16] SAPORTA G. – "Le traitement de variables qualitatives par codage". Note de travail n° 11, COREF.

[17] TENENHAUS M. – "Etude des programmes MORALS et CORALS de F.W. YOUNG, J. de LEEUW et Y. TAKANE". Note de travail n° 15, COREF.

[18] TENENHAUS M. – "Analyse en composantes principales d'un ensemble de variables nominales et numériques". *Revue de Statistiques Appliquées,* Vol. XXV, n° 2, 1977.

[19] TENENHAUS M., MACQUIN A. – "Le modèle linéaire ordinal". Note CESA, 1978.

[20] YOUNG F.W., DE LEEUW J., TAKANE Y. – "How to use principals". *Psychometrika Laboratory,* 1975.

[21] YOUNG F.W., DE LEEUW J., TAKANE Y. – "Regression with qualitative and quantitative variables : An Alternating Least Squares method with optimal scaling features". *Psychometrika,* 41, 1976.

[22] YOUNG F.W., DE LEEUW J., TAKANE Y. – "Additive structure in qualitative data : an Alternating Least Squares method with optimal scaling features (ADDALS)". *Psychometrika,* 41, 1976.

[23] YOUNG F.W., DE LEEUW J., TAKANE Y. – "Non metric individual differences multidimensional scaling : an Alternating Least Squares method with optimal SCALing features (ALSCAL)". *Psychometrika,* 42, 1977.