REVUE DE STATISTIQUE APPLIQUÉE

P. CAZES

Codage d'une variable continue en vue de l'analyse des correspondances

Revue de statistique appliquée, tome 38, nº 3 (1990), p. 35-51

<http://www.numdam.org/item?id=RSA_1990__38_3_35_0>

Rev. Statistique Appliquée, 1990, **XXXVIII** (3), 35-51

CODAGE D'UNE VARIABLE CONTINUE EN VUE DE L'ANALYSE DES CORRESPONDANCES

P. CAZES

CERESTA
Université Paris Dauphine

RÉSUMÉ

Nous comparons ici divers codages flous d'une variable continue utilisés en vue de l'analyse des correspondances. Ces codages qui se font sur deux ou trois colonnes sont des cas particuliers ou des variantes du codage barycentrique (ou semi-linéaire). Le cas d'un ensemble de notes fourni par des experts est également détaillé pour examiner comment on peut dans une certaine mesure s'affranchir de la façon de noter de chaque expert.

Mots clés : *Analyse des Correspondances - Analyse en Composantes Principales - Codages Barycentrique, Disjonctif Complet, Flou - Fenêtre Glissante.*

Indices de classification *STMA : 06-070 , 06-110.*

SUMMARY

We compare several fuzzy codages of a continuous variable used in correspondence analysis. These codages split the variable in two or three items; they are particular cases of barycentric codage. The case of a set of notes is studied for see how shake off the manner in which each person notes.

Key Words : *Correspondence Analysis - Principal Component Analysis - Barycentric, Complete Disjunctive, Fuzzy Codages - Sliding Window.*

I Introduction

Nous proposons ici divers codages d'une variable continue en vue de l'analyse des correspondances. Ces codages qui peuvent s'apparenter à un codage flou se font sur deux ou trois colonnes, et permettent comme le codage disjonctif complet usuel de traiter simultanément ces variables avec des variables qualitatives codées en 0-1. Contrairement au codage disjonctif complet, les codages proposés ont l'avantage d'une part de ne pas faire perdre d'information puisque l'on ne perd pas la valeur de la variable prise pour ne garder que l'appartenance à une tranche; et d'autre part de ne pas créer de distance artificielle entre deux individus proches mais situés de part et d'autre de la frontière entre deux classes.

La plupart des codages développés ici correspondant à un dédoublement, nous faisons d'abord des rappels sur l'analyse des correspondances d'un tableau dédoublé. Nous proposons ensuite cinq codages d'une variable continue (quatre sur deux colonnes, le dernier sur trois colonnes) qui ont été utilisés avec succès ou qui dérivent de codages utilisés en pratique, et nous montrons que ces codages sont des cas particuliers ou des variantes du codage barycentrique (encore appelé semi-linéaire); puis nous examinons le cas d'un système de notes et étudions comment s'affranchir de la façon de noter de chaque sujet. Nous appliquons alors les résultats obtenus au cas d'un codage par rangs et à l'analyse des préférences. Nous étudions ensuite comment on peut ramener un codage disjonctif complet trop fin et risquant de donner des résultats d'analyse instables à un codage flou plus robuste, avant de donner un aperçu sur la technique de fenêtre glissante en codage disjonctif complet.

II Rappels sur le dédoublement en analyse des correspondances

Soit k_{0IJ} un tableau rectangulaire défini sur le produit de deux ensembles I et J et $k_{IJ'}$ le tableau déduit de k_{0IJ} par dédoublement et défini par :

$$J' = J^+ U J^- \quad ; \quad J^+ = J^- = J$$

$$\forall i \epsilon I, \forall j+ \ \epsilon J^+, \forall j- \ \epsilon J^- : \left.\begin{array}{l} k(i,j+) = k_0(i,j) \\ k(i,j-) = d_j - k(i,j+) \end{array}\right\} \tag{1}$$

d_j étant une quantité connue.

Posons : $D = \sum \{d_j / j \epsilon J\}$

$$\left.\begin{array}{lll} \forall i \epsilon I : & k(i) & = \sum \{k(i,j)/j \epsilon J'\} = D \\ \forall j \epsilon J : & k_0(j) & = \sum \{k_0(i,j)/i \epsilon I\} \end{array}\right\} \tag{2}$$

Alors (cf [2a] et [7]) l'analyse des correspondances du tableau $k_{IJ'}$ est équivalente(*) à l'analyse en composantes principales (A.C.P.) du tableau k_{0IJ}, chaque point i étant muni de la masse 1/n (où n = CardI) et l'espace R_J étant muni de la métrique définie par :

$\forall i, i' \epsilon I : d^2(i,i') =$

$$\frac{n^2}{D} \sum \left\{ \frac{d_j}{k_0(j)(nd_j - k_0(j))} \left(k_0(i,j) - k_0(i',j)\right)^2 / j \epsilon J \right\} \tag{3}$$

$d^2(i,i')$ désignant le carré de la distance entre i et i'.

En particulier, nous traiterons le cas où :

$$k_0(i,j) = c_j + b_j x_{ij} \tag{4}$$

x_{ij} étant la valeur de la variable x_j pour l'individu i, c_j et b_j étant des constantes indépendantes de i.

La formule de distance précédente s'écrit alors :

$$d^2(i,i') = \sum \{q_j \left(x_{ij} - x_{i'j}\right)^2 / j \epsilon J\} \tag{5}$$

(*) Dans le sens où l'on obtient la même représentation de I sur les axes factoriels issus des deux analyses.

avec :

$$q_j = \frac{d_j b_j^2}{D(c_j + b_j \bar{x}_j)(d_j - c_j - b_j \bar{x}_j)} \tag{6}$$

$\bar{x}_j$ qui est défini en (8) désignant la moyenne (sur I) des x_{ij}.

L'analyse des correspondances du tableau k_{IJ} est donc équivalente à l'A.C.P. du tableau X des x_{ij}, R_J étant muni de la métrique définie par (5) et (6).

Dans les cas que nous traiterons, $d_j = 1$ et donc D est égal au nombre Card$J$ de variables. Posant $p = CardJ$, l'expression de q_j se simplifie et s'écrit alors :

$$q_j = \frac{b_j^2}{p(c_j + b_j \bar{x}_j)(1 - c_j - b_j \bar{x}_j)} \tag{7}$$

III Quelques exemples de codage d'une variable continue

Nous supposons ici que l'on a p variables quantitatives x_j $(j \epsilon J)$ mesurées sur un ensemble I de n individus, et nous désignerons par x_{ij} la valeur de la variable x_j pour l'individu i. Nous désignerons par X le tableau des x_{ij} et nous poserons :

$$\left.\begin{aligned} M_j &= \text{Max } \{x_{ij}/i\epsilon I\} \\ m_j &= \text{Min } \{x_{ij}/i\epsilon I\} \\ \bar{x}_j &= (1/n)\sum \{x_{ij}/i\epsilon I\} \\ \sigma_j^2 &= (1/n)\sum \{(x_{ij} - \bar{x}_j)^2/i\epsilon I\} \\ A_j &= \text{Max } (M_j - \bar{x}_j, \bar{x}_j - m_j) \end{aligned}\right\} \tag{8}$$

III.1 Codage de x_j suivant deux modalités $j+$ et $j-$

III.1.1 Les codages proposés

Comme au paragraphe précédent, nous poserons $J' = J^+ U J^-$, avec $J^+ = J^- = J$.

Nous suggérons quatre codages possibles. Chacun de ces codages revient à définir un tableau de base k_{0IJ} dont le terme général est de la forme donnée par (4) et à le dédoubler par rapport à 1, les codages différant par les valeurs b_j et c_j adoptées.

L'analyse des correspondances du tableau dédoublé associé est donc équivalente à l'A.C.P. de X, la métrique dans R_J dépendant du codage adopté et étant définie par (5) et (7).

Le choix du codage semble donc équivalent au choix des coefficients q_j de la métrique. Nous verrons en fait que ce n'est pas tout à fait le cas du fait de la représentation des modalités $j+$ et $j-$ de la variable x_j (cf § III.1.3, 2$^{\text{ème}}$ remarque).

• 1$^{\text{er}}$ *codage* :

Ce codage est défini par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\left.\begin{aligned} k_1(i,j+) &= (x_{ij} - m_j) \,/\, (M_j - m_j) \\ k_1(i,j-) &= (M_j - x_{ij}) \,/\, (M_j - m_j) = 1 - k_1(i,j+) \end{aligned}\right\} \tag{9}$$

Avec ce codage assez naturel, les valeurs 0 et 1 sont atteintes pour $k_1(i,j+)$ (et donc pour $k_1(i,j-)$). Le coefficient de métrique q_j s'écrit alors d'après (4), (7) et (9) :

$$q_j = (p(\bar{x}_j - m_j)(M_j - \bar{x}_j))^{-1} \tag{10}$$

• 2ème*codage* :

Il est donné par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\left.\begin{aligned} k_2(i,j+) &= (1 + (x_{ij} - \bar{x}_j)/A_j)/2 \\ k_2(i,j-) &= (1 - (x_{ij} - \bar{x}_j)/A_j)/2 \end{aligned}\right\} \tag{11}$$

Ce codage est suggéré par le codage d'un ensemble de notes introduit par A. DUBOIS dans sa thèse (cf [9]), pour s'abstraire de la façon de noter de chaque sujet, codage dont le principe est rappelé au §III.4.2. On peut noter qu'en général, $k_2(i,j+)$ n'atteint que l'une des deux valeurs 0 ou 1. Le coefficient q_j associé s'écrit d'après (4), (7) et (11) :

$$q_j = 1/(pA_j^2) \tag{12}$$

Remarque :

Posons : $A = \text{Max}\,\{A_j / j \epsilon J\}$.

Si dans la formule (11) on remplace A_j par A, le coefficient q_j devient indépendant de j. L'analyse des correspondances du tableau $k_{2IJ'}$ est alors équivalente (au coefficient $1/\left(pA^2\right)$ près) à l'A.C.P. sur matrice variance du tableau X.

• $3^{ème}$ *codage* :

Il est défini par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\left.\begin{aligned} k_3(i,j+) &= (1 + (x_{ij} - \bar{x}_j)/\sigma_j)/2 \\ k_3(i,j-) &= (1 - (x_{ij} - \bar{x}_j)/\sigma_j)/2 \end{aligned}\right\} \qquad (13)$$

Ce codage a été introduit par B. ESCOFIER (cf [10]) pour analyser conjointement des variables qualitatives et quantitatives. On peut noter que contrairement aux codages précédents, on peut obtenir des valeurs extérieures à l'intervalle [0,1], et en particulier des valeurs négatives, ce qui n'est pas gênant au niveau de l'analyse des correspondances du tableau $k_{3IJ'}$, compte tenu de ce que tous les éléments de marge sont strictement positifs. Le coefficient q_j valant $1/(p\sigma_j^2)$,cette analyse est équivalente (au facteur $1/p$ près) à l'A.C.P. sur matrice de corrélation du tableau X .

• $4^{ème}$ *codage* :

Contrairement aux codages précédents, ce codage n'est valable que si la variable x_j est positive. Il est défini par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\left.\begin{aligned} k_4(i,j+) &= x_{ij}/M_j \\ k_4(i,j-) &= 1 - x_{ij}/M_j \end{aligned}\right\} \qquad (14)$$

Le coefficient q_j associé s'écrit :

$$q_j = (p\bar{x}_j(M_j - \bar{x}_j))^{-1} \qquad (15)$$

III.1.2 Comparaison des codages précédents

Les codages précédents reviennent à normaliser la variable x_j puis à associer à la variable normalisée $j+$, la variable $j-$ obtenue en dédoublant $j+$ par rapport à 1. On peut remarquer, en comparant les coefficients q_j, que ces codages introduisent la dispersion de la variable x_j de façon différente. Les deux premiers codages se fondent sur l'écart entre la moyenne $\bar{x}_j$ et les valeurs extrêmes m_j et M_j. On peut noter que ces deux codages sont identiques si $\bar{x}_j$ est à mi-distance de m_j et M_j. Le troisième codage se fonde sur l'écart-type, tandis que le dernier codage (où la variable x_j est supposée positive) est équivalent au premier à condition de remplacer m_j par la valeur nulle (qui est une borne de x_j, pas forcément atteinte sur l'échantillon I considéré).

Une autre comparaison de ces codages est donnée au § III.3 à partir de la notion de codage barycentrique.

III.1.3 Remarques

1) Dans le cas où l'on a un ensemble de notes, si a_j désigne la valeur maximale possible pour la note x_j (valeur pas forcément atteinte) on peut dans le quatrième codage remplacer M_j par a_j (de la même façon que dans ce codage, on a par rapport au premier codage remplacé la valeur minimale m_j par la valeur minimale possible à savoir zéro).

2) Si la variable x_j ne prend que les valeurs 0 ou 1 (auquel cas $m_j = 0, M_j = 1$), le premier et le quatrième codage sont identiques. Si donc toutes les variables sont binaires, les tableaux k_1 et k_4 sont identiques et correspondent au dédoublement classique du tableau X. On sait alors (cf [7]) que l'analyse des correspondances du tableau k_1 (ou k_4) est équivalente à l'A.C.P. sur matrice de corrélation du tableau X, ce que l'on peut vérifier immédiatement à partir de (10) compte tenu de ce que $(\bar{x}_j - m_j)(M_j - \bar{x}_j) = \bar{x}_j(1-\bar{x}_j)$ qui n'est autre, puisque l'on a des données 0-1 que la variance σ_j^2 de x_j. On obtient donc le même coefficient q_j pour les codages 1, 3 et 4, et donc les mêmes représentations pour l'ensemble I. Par contre les représentations de $j+$ et de $j-$ ne sont pas identiques dans l'analyse de k_1 (ou k_4) et dans celle de k_3. De façon précise, soit $r_{j\alpha}$ la corrélation entre la variable x_j et le facteur α dans l'A.C.P. sur matrice de corrélation de X. Alors dans l'analyse des correspondances du tableau k_1, les points $j+$et $j-$ sont représentés sur le $\alpha^{\grave{e}me}$ axe factoriel par des points dont l'abscisse est donnée (cf [7]) par :

$$G_\alpha(j+) = \sqrt{(1-\bar{x}_j)/\bar{x}_j} r_{j\alpha}; G_\alpha(j-) = -\sqrt{\bar{x}_j/(1-\bar{x}_j)} r_{j\alpha}$$

tandis que dans l'analyse des correspondances du tableau k_3, les représentations associées sont données (cf [10]) par :

$$G_\alpha(j+) = r_{j\alpha} \qquad ; \qquad G_\alpha(j-) = -r_{j\alpha}$$

3) Dans l'analyse des correspondances des tableaux $k_m (m = 1, 4)$ tous les individus ont le même poids $1/n$; de même le poids de chaque variable x_j (qui est la somme des poids de $j+$ et de $j-$) est égal à $1/p$ et ne dépend donc pas de j. Si on veut donner un poids p_i ($\sum p_i = 1$) à chaque individu i, et une pondération p_j à chaque variable x_j ($\sum p_j = 1$) il suffit de considérer le codage plus général suivant défini par :

$$\forall m = 1, 4, \forall i \epsilon I, \forall j \epsilon J : k'_m(i,j) = p_i p_j k_m(i,j)$$

L'expression (6) du coefficient de métrique q_j doit alors être remplacée par l'expression suivante :

$$q_j = \frac{p_j d_j b_j^2}{\bar{d}(c_j + b_j\bar{x}_j)(d_j - c_j - b_j\bar{x}_j)} \qquad (6')$$

où $\bar{d} = \sum_j p_j d_j$ et $\bar{x}_j = \sum_i p_i x_{ij}$.

Dans les cas qui nous intéressent, $d_j = \bar{d} = 1$, et donc q_j s'écrit :

$$q_j = \frac{p_j b_j^2}{(c_j + b_j \bar{x}_j)(1 - c_j - b_j \bar{x}_j)} \tag{7'}$$

Tous les résultats établis précédemment restent valables à condition de calculer la variance σ_j^2 en affectant des poids p_i à chaque observation i, comme on l'a fait pour le calcul de la moyenne $\bar{x}_j$.

III.2 Codage de x_j suivant trois modalités $j-, j=$ et $j+$

Ce codage est inspiré du codage d'un ensemble de notes suivant une formule tenant compte de la façon de noter de chaque sujet, codage introduit par P. LOSLEVER et alt. (cf [12]) et par T. BEHRAKIS et alt. (cf [3]) et dont le principe est rappelé au § III.4.3.

Ce codage se présente ici sous la forme suivante :

Si $x_{ij} < \bar{x}_j$:

$$\left.\begin{array}{ll} k(i,j+) & = 0 \\ k(i,j=) & = (x_{ij} - m_j)/(\bar{x}_j - m_j) \\ k(i,j-) & = 1 - k(i,j=) = (\bar{x}_j - x_{ij})/(\bar{x}_j - m_j) \end{array}\right\} \tag{16}$$

Si $x_{ij} \geq \bar{x}_j$:

$$\left.\begin{array}{ll} k(i,j+) & = (x_{ij} - \bar{x}_j)/(M_j - \bar{x}_j) \\ k(i,j=) & = 1 - k(i,j+) = (M_j - x_{ij})/(M_j - \bar{x}_j) \\ k(i,j-) & = 0 \end{array}\right\} \tag{17}$$

Ce codage vérifie les deux propriétés suivantes :

$$\left.\begin{array}{rcrcr} k(i,j-) & + & k(i,j=) & + & k(i,j+) = 1 \\ m_j k(i,j-) & + & \bar{x}_j k(i,j=) & + & M_j k(i,j+) = x_{ij} \end{array}\right\} \tag{18}$$

Ces relations faciles à vérifier traduisent simplement le fait que le barycentre de $m_j, \bar{x}_j$ et M_j affectés respectivement des masses $k(i,j-)$, $k(i,j=)$ et $k(i,j+)$ est x_{ij}.

Remarque : Le codage précédent revient à recadrer la variable x_j entre -1 et +1 en posant $y_{ij} = -k(i,j-)$ si $x_{ij} < \bar{x}_j$ et $y_{ij} = k(i,j+)$ si $x_{ij} \geq \bar{x}_j$, puis à éclater y_{ij} suivant les trois modalités $j-, j=$ et $j+$ en posant :

Si $x_{ij} < \bar{x}_j$: $k(i,j+) = 0$; $k(i,j=) = 1 + y_{ij}$; $k(i,j-) = -y_{ij}$

Si $x_{ij} \geq \bar{x}_j$: $k(i,j+) = y_{ij}$; $k(i,j=) = 1 - y_{ij}$; $k(i,j-) = 0$

III.3 Codage barycentrique de x_j en r modalités

Ce codage est un codage linéaire par morceaux. Tous les codages examinés précédemment sont des cas particuliers ou des variantes de ce codage dont le principe est le suivant : on se donne r pivots, i.e. r valeurs $t_1, t_2, ..., t_r$ $(t_1 < t_2 < t_3 < ... < t_r)$. Si $(k(i,j_1), k(i,j_2), ..., k(i,j_r))$ désigne le codage de x_j pour l'individu i, on pose :

Si $x_{ij} \leq t_1$: $k(i,j_1) = 1$; $k(i,j_s) = 0$ si $s \neq 1$

Si $x_{ij} \geq t_r$: $k(i,j_r) = 1$; $k(i,j_s) = 0$ si $s \neq r$

Si $t_m \leq x_{ij} \leq t_{m+1}$:

$$\begin{aligned} k(i,j_m) &= (t_{m+1} - x_{ij}) / (t_{m+1} - t_m) \\ k(i,j_{m+1}) &= (x_{ij} - t_m) / (t_{m+1} - t_m) \\ k(i,j_s) &= 0 \text{ si } s \neq m \text{ ou } m+1 \end{aligned}$$

Avec ce codage, où au plus deux des valeurs $k(i,j_s)$ sont non nulles, on a les relations, pour x_{ij} appartenant à l'intervalle $[t_1, t_r]$:

$$\left.\begin{aligned} &\sum \{k(i,j_s) / s = 1, r\} = 1 \\ &\sum \{k(i,j_s) t_s / s = 1, r\} = x_{ij} \end{aligned}\right\} \qquad (19)$$

relations qui généralisent les formules (18) et qui traduisent le fait que x_{ij} est le barycentre des points t_s affectés des masses $k(i,j_s)$.

Tous les codages définis au § III.1.1 (sauf le troisième) sont des cas particuliers de ce codage quand l'on a deux pivots, ces pivots étant respectivement donnés par $t_1 = m_j$, $t_2 = M_j$ pour le premier codage, $t_1 = \bar{x}_j - A_j$, $t_2 = \bar{x}_j + A_j$ pour le deuxième, $t_1 = 0$, $t_2 = M_j$ pour le quatrième. Le troisième codage est très voisin d'un codage barycentrique : il correspond aux pivots $t_1 = \bar{x}_j - \sigma_j$, $t_2 = \bar{x}_j + \sigma_j$, et est identique à ce codage entre t_1 et t_2. Par contre, à l'extérieur de (t_1, t_2) il en diffère, $k(i,j+)$ n'étant pas égal à 0 (resp. 1) si x_{ij} est plus petit que t_1 (resp. plus grand que t_2). On peut du reste noter que ce codage vérifie (19) quelle que soit la valeur de x_{ij}. En fait ce codage cesse d'être barycentrique quand $k(i,j+)$ et $k(i,j-)$ sortent de l'intervalle (0,1). On pourrait se ramener au codage barycentrique en posant $k(i,j+) = 1$ si $x_{ij} > \bar{x}_j + \sigma_j$, et $k(i,j+) = 0$ si $x_{ij} < \bar{x}_j - \sigma_j$, $k(i,j+)$ restant donné par (13) quand x_{ij} est compris entre $\bar{x}_j - \sigma_j$ et $\bar{x}_j + \sigma_j$, $k(i,j-)$ étant bien sûr dans tous les cas égal à $1 - k(i,j+)$.

Le codage en trois modalités du § III.2 est également un cas particulier du codage barycentrique avec trois pivots définis par $t_1 = m_j, t_2 = \bar{x}_j$, $t_3 = M_j$. Ce codage peut être considéré comme dérivant du premier codage à deux modalités en rajoutant le pivot intermédiaire $\bar{x}_j$, ce qui permet d'envisager immédiatement d'autres codages en rajoutant aux trois autres codages à deux modalités ce pivot intermédiaire $\bar{x}_j$.

L'introduction de ce codage barycentrique permet donc de comparer aisément tous les codages introduits auparavant.

Remarque : On trouve dans GALLEGO (cf [11]) d'autres types de codages flous, et en particulier une modification du codage barycentrique (que GALLEGO appelle semi-linéaire) rendant injectif ce codage pour les valeurs de x_j extérieures à l'intervalle $[t_1, t_r]$ défini par les pivots extrêmes.

III.4 Cas d'un ensemble de notes

III.4.1 Introduction

On suppose ici que l'ensemble J des variables est un ensemble de notes. L'analyse du tableau individus × notes fait souvent ressortir des caractéristiques propres à la notation des individus. Suivant le sujet considéré, la notation peut être sévère, indulgente, avec une plage de variation plus ou moins grande. Pour s'affranchir de l'équation personnelle de chaque individu, on peut considérer les codages suivants, analogues à ceux donnés précédemment, mais où les calculs de moyenne, maximum, minimum etc... se font pour chaque individu sur l'ensemble des notes. On supposera que la plage de variation de chaque note est la même 0-a (par exemple 0-20 ou 0-10, ou encore 0-1) quitte à affecter le cas échéant des pondérations à certaines notes (cf Remarque 3 du § III.1.3.).

On posera avec des notations analogues à celles définies par (8) :

$$\left.\begin{array}{lll} \forall i \epsilon I: & M_i & = \text{Max } \{x_{ij}/j\epsilon J\} \\ & m_i & = \text{Min } \{x_{ij}/j\epsilon J\} \\ & \bar{x}_i & = \sum \{x_{ij}/j\epsilon J\}/p \\ & Ai & = \text{Max } \{M_i - \bar{x}_i, \bar{x}_i - m_i\} \end{array}\right\} \qquad (20)$$

III.4.2 Codage de x_j suivant deux modalités $j+$ et $j-$

Nous allons présenter trois codages qui sont les analogues des codages 1, 2 et 4 présentés au § III.1.1 mais où il faut intervertir les rôles joués par les indices i et j. Ces codages reviennent à définir un tableau de base k_{0IJ} dont le terme général $k_0(i,j)$ est de la forme $c_i + b_i x_{ij}$, et à le dédoubler par rapport à 1, c_i et b_i étant des coefficients dépendant du codage adopté.

L'analyse des correspondances du tableau dédoublé est donc équivalente à l'A.C.P. du tableau k_0, la métrique étant donnée par (3) (où $d_j = 1, D = p$), mais elle n'est plus équivalente à l'A.C.P. de X avec une métrique adéquate, comme c'était le cas au § III.1.

- 1^{er} *codage* :

Il est défini par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\left.\begin{array}{l} k_1(i,j+) = (x_{ij} - m_i)/(M_i - m_i) \\ k_1(i,j-) = 1 - k_1(i,j+) = (M_i - x_{ij})/(M_i - m_i) \end{array}\right\} \qquad (21)$$

Le codage $j+$ ramène l'intervalle de notation du sujet i à l'intervalle 0-1, les extrémités 0 et 1 étant atteintes, le codage $j-$ correspondant au dédoublement classique.

• $2^{ème}$ *codage* :

Il est défini par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\left.\begin{array}{l} k_2(i,j+) = (1 + (x_{ij} - \bar{x}_i)/Ai)/2 \\ k_2(i,j-) = 1 - k_2(i,j+) = (1 - (x_{ij} - \bar{x}_i)/Ai)/2 \end{array}\right\} \quad (22)$$

Ce codage a été utilisé comme on l'a déjà dit par A. DUBOIS dans sa thèse pour analyser les résultats d'une enquête relative à l'opinion des usagers sur les services offerts par les compagnies aériennes (chaque passager interviewé accordant une note de 1 (très médiocre) à 5 (très bien) à une dizaine de services).

• $3^{ème}$ *codage* :

Il est défini par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\left.\begin{array}{l} k_3(i,j+) = x_{ij}/M_i \\ k_3(i,j-) = 1 - x_{ij}/M_i \end{array}\right\} \quad (23)$$

C'est l'analogue du quatrième codage proposé au § III.1.1.

Remarque : On ne peut utiliser l'analogue du troisième codage donné au § III.1.1, car il peut conduire à des éléments de marge négatifs.

III.4.3 Codage de x_j suivant trois modalités $j-, j =, j+$

Ce codage dont on a déjà parlé (cf § III.2) semble être le plus utilisé actuellement. Il revient à recadrer la variable x_j entre -1 et +1, puis à éclater la variable ainsi recadrée et que nous appellerons y_j, en trois modalités $j-, j =$ et $j+$.

La variable y_j est définie par :

$\forall i \epsilon I, \forall j \epsilon J$:

$$\begin{array}{ll} y_{ij} & = (x_{ij} - \bar{x}_i)/(M_i - \bar{x}_i) \text{ si } x_{ij} \geq \bar{x}_i \\ & = (x_{ij} - \bar{x}_i)/(\bar{x}_i - m_i) \text{ sinon} \end{array}$$

Le codage se présente alors sous la forme suivante :

$$\left.\begin{array}{lll} \text{Si } x_{ij} < \bar{x}_i : & k(i,j+) & = 0 \\ & k(i,j=) & = 1 + y_{ij} = (x_{ij} - m_i)/(\bar{x}_i - m_i) \\ & k(i,j-) & = 1 - k(i,j=) = -y_{ij} = (\bar{x}_i - x_{ij})/(\bar{x}_i - m_i) \\ & & \\ \text{Si } x_{ij} \geq \bar{x}_i : & k(i,j+) & = y_{ij} = (x_{ij} - \bar{x}_i)/(M_i - \bar{x}_i) \\ & k(i,j=) & = 1 - y_{ij} = (M_i - x_{ij})/(M_i - \bar{x}_i) \\ & k(i,j-) & = 0 \end{array}\right\} \quad (24)$$

Ce codage vérifie les relations (18) (où il faut remplacer m_j, $\bar{x}_j$, et M_j par m_i, $\bar{x}_i$ et M_i respectivement), relations traduisant, rappelons-le, le fait que le barycentre de m_i, $\bar{x}_i$ et M_i affectés respectivement des masses $k(i,j-)$, $k(i,j=)$et $k(i,j+)$ est x_{ij}.

Ce codage a notamment été employé pour traiter des données relatives à la notation de 0 à 10 d'hommes politiques grecs (cf [3]), pour étudier un questionnaire ergonomique relatif à l'appréciation des réglages d'un poste de travail (cf [12]) ainsi que pour analyser les réponses (sous forme de notes) d'étudiants à un questionnaire relatif à leur mémoire de recherche (cf [13]).

Remarques : 1) On peut à partir de la variable recadrée entre -1 et 1 y_j l'éclater en trois modalités suivant le codage défini au § III.2, moyenne, minimum et maximum étant calculés sur y_j (et non plus sur x_j); ce qui correspond après avoir recadré x_j en ligne à la recadrer en colonne, et à éclater le résultat en trois modalités que nous noterons $j <, j \approx$et $j >$ pour les distinguer de celles définies par (24). L'intérêt de ce double recadrage est de produire des notes $j >$ et $j <$ équilibrées qu'il s'agisse d'une matière j généralement bien ou mal notée (alors que la modalité $j-$ (resp $j+$) définie par (24) est d'autant plus lourde que j est mal (resp. bien) notée). Ce double recadrage a en particulier été utilisé, conjointement avec le recadrage simple, dans l'enquête sur les mémoires de recherche mentionnée ci-dessus.

2) Des programmes permettant d'effectuer facilement le codage (24) ainsi que le codage barycentrique, et donc tous les codages présentés précédemment sont donnés dans BENZECRI (cf [5], [6]).

III.5 Codage à partir des rangs

III.5.1 Codage variable par variable

Le codage par rangs est souvent utilisé quand on veut obtenir un traitement robuste des données non perturbé en particulier par des valeurs extrêmes qui peuvent être aberrantes ou avoir trop de poids dans les analyses. Rappelons que ce codage revient à remplacer la valeur x_{ij} de la variable x_j pour l'individu i par son rang r_{ij} dans la suite ordonnée par valeurs croissantes des x_{ij} $(i = 1, n)$.

Les r_{ij} $(1 \leq i \leq n)$ constituent alors une permutation des entiers de 1 à n. On peut appliquer à partir des r_{ij} soit les codages des §§ III.1 et III.2, soit les codages du § III.4 en considérant alors les rangs comme des notes. Dans le premier cas, les formules utilisées aux §§ III.1 et III.2 se simplifient. En effet, gardant des notations analogues à (8), mais où x_{ij} est remplacé par r_{ij}, on a :

$$\begin{aligned}
m_j &= \text{Min } \{r_{ij}/i = 1, n\} = 1\\
M_j &= \text{Max } \{r_{ij}/i = 1, n\} = n\\
\bar{r}_j &= \sum \{r_{ij}/i = 1, n\}/n = (n+1)/2\\
\sigma_j^2 &= \sum \{(r_{ij} - \bar{r}_j)^2/i = 1, n\}/n = (n^2-1)/12\\
A_j &= M_j - \bar{r}_j = \bar{r}_j - m_j = \frac{n-1}{2}
\end{aligned}$$

Toutes ces quantités étant indépendantes de j, il en est de même des coefficients de métrique q_j associés à l'un des quatre codages définis au § III.1.1. L'analyse des correspondances du tableau $k_{mIJ'}(m = 1, 4)$associé à l'un de ces quatre codages est donc équivalente à l'A.C.P. sur matrice variance du tableau R des rangs r_{ij}. Les quatre codages sont donc équivalents.

III.5.2 Cas de l'analyse des préférences

Dans ce cas, chaque individu i classe les variables qui sont des critères de 1 à p. Chaque ligne x_{ij} $(j = 1, p)$constitue donc une permutation des entiers 1 à p. Les quantités m_i, $\bar{x}_i, M_i$ et A_i sont indépendantes de $i(m_i = 1, M_i = p, \bar{x}_i = (p+1)/2, A_i = (p-1)/2)$. Les transformations liées aux trois codages (21), (22) et (23) sont donc indépendantes de i, ce qui implique que dans ce cas, l'analyse des correspondances du tableau associé à l'un de ces codages est équivalente à l'A.C.P. du tableau X avec une métrique appropriée (en effet, on a pour chacun de ces codages une transformation qui est donnée par (1) et (4) avec $d_j = 1$, les coefficients b_j et c_j ne dépendant pas de j, ce qui permet d'appliquer les résultats du § II).

De façon précise, les deux codages (21) et (22) sont identiques (puisque A_i est équidistant de m_i et M_i) le coefficient de métrique q_j associé s'écrivant :

$$q_j = (p(\bar{x}_j - 1)(p - \bar{x}_j))^{-1} \tag{25}$$

tandis que le coefficient de métrique associé au codage (23) est donné par :

$$q_j = (p\bar{x}_j(p - \bar{x}_j))^{-1} \tag{26}$$

Ces coefficients sont à comparer avec ceux donnés au § III.1.1 (formules (10), (12) et (15)) quand on adopte les codages définis dans ce paragraphe. Comparons par exemple (10) et (25). On voit que (10) est identique à (25) si les valeurs m_j et M_j atteignent leur borne, à savoir 1 et p, ce qui signifie qu'il existe au moins deux individus i et i' tels que i classe le critère j en premier, i' le classant en dernier. S'il en est ainsi, les codages (9), (11), (21) et (22) sont identiques. De même (15) et (26) sont identiques, si $M_j = p$, i.e. s'il existe au moins un individu i classant j en dernier, auquel cas les codages (14) et (23) sont identiques.

On peut noter que le codage (21) revient en fait à faire aller les rangs de 0 à $p-1$, puis à effectuer un dédoublement à partir de la valeur maximale $p-1$, tandis qu'avec le codage (23) on considère les rangs de 1 à p, et on dédouble par rapport à p. Usuellement on considère les rangs de 1 à p, et on dédouble par rapport à $p+1$, ce qui correspond à un coefficient de métrique q_j égal à $(p\bar{x}_j(p+1-\bar{x}_j))^{-1}$ (cf [2b], [2c], [7]).

IV Codage flou associé à un codage disjonctif complet

Il arrive souvent qu'on divise une variable quantitative x_j en un nombre de classes suffisant pour que cela ait un sens, et ceci même si on a un nombre d'individus relativement faible. Pour que l'analyse ne soit pas trop sensible à ce découpage et à ce petit nombre d'individus, on peut à partir du codage disjonctif complet correspondant définir un codage flou. On ramène, ainsi qu'il est expliqué ci-après, un codage disjonctif complet à $2r-1$ (ou $2r-2$) modalités à un codage flou à r modalités, ce qui permet de stabiliser les analyses où interviennent la variable x_j.

Désignons par $j_1, j_2, ..., j_s$ ($s = 2r-1$ ou $2r-2$) les modalités (ordonnées) associées au codage disjonctif complet de x_j et par $k\,(i, j_m)$ la valeur de ce codage pour l'individu i et la modalité j_m $(1 \leq m \leq s)$ $(k(i,j_m) = 1$ si i tombe dans la tranche j_m ,0 sinon).

Désignons de même par $j'_1, j'_2, ..., j'_r$ les modalités du codage flou associé et par $k'(i,j'_m)(1 \leq m \leq r)$ la valeur de ce codage pour l'individu i et la modalité j'_m. Ce codage est défini de la façon suivante :

$$\forall m : 1 \leq m \leq r \qquad : k(i, j_{2m-1}) \ = 1 \Longrightarrow \ \begin{array}{ll} k'(i,j'_m) & = 1 \\ k'(i,j'_q) & = 0 \text{ si } q \neq m \end{array}$$

$$\forall m : 1 \leq m \leq r-1 \quad : k(i, j_{2m}) \ = 1 \Longrightarrow \ \begin{array}{ll} k'(i,j'_m) & = k'(i,j'_{m+1}) = 1/2 \\ k'(i,j'_q) & = 0 \text{ si } q \neq m \text{ ou } m+1 \end{array}$$

La correspondance entre les deux codages est donnée dans le tableau 1 dans le cas particulier d'une variable découpée en $s = 7$ tranches, auquel cas on a $r = 4$ modalités pour le codage flou.

TABLEAU 1

	j_1	j_2	j_3	j_4	j_5	j_6	j_7
j'_1	1	1/2	0	0	0	0	0
j'_2	0	1/2	1	1/2	0	0	0
j'_3	0	0	0	1/2	1	1/2	0
j'_4	0	0	0	0	0	1/2	1

Ce tableau se lit de la façon suivante : si un individu i tombe dans la tranche j_t, le codage flou associé est donné dans la colonne correspondant à j_t.

Ce codage qui est maintenant très usité a été utilisé pour la première fois par S. CHAIEB (cf [8]) dans une étude pharmacocinétique. Une variante de ce codage, plus précise et utilisant l'histogramme de la variable à coder, a été réalisée par A. SKALLI (cf [14]) dans une étude portant sur les propriétés olfactives des deux enanthiomorphes de l'α ionone.

De façon plus générale, on peut (cf [3bis]) ramener un codage disjonctif complet à $s = u(r-1)+1$ modalités à un codage flou à r modalités, u étant un entier supérieur ou égal à 2.

Gardant des notations analogues à celles qui viennent d'être utilisées, ce codage est défini de la façon suivante :

$$\forall m : 1 \leq m \leq r-1, \forall t : 1 \leq t \leq u+1 :$$

$$k(i, j_{u(m-1)+t}) = 1 \Longrightarrow \begin{array}{ll} k'(i, j'_m) & = 1 - \dfrac{t-1}{u} \\ k'(i, j'_{m+1}) & = \dfrac{t-1}{u} \\ k'(i, j'_q) & = 0 \text{ si } q \neq m \text{ ou } m+1 \end{array}$$

Le tableau 2, analogue au tableau 1, donne la correspondance entre les deux codages dans le cas où $s = 7$, $u = r = 3$.

TABLEAU 2

	j_1	j_2	j_3	j_4	j_5	j_6	j_7
j'_1	1	2/3	1/3	0	0	0	0
j'_2	0	1/3	2/3	1	2/3	1/3	0
j'_3	0	0	0	0	1/3	2/3	1

Le codage précédent est en fait un codage barycentrique (cf § III 3) ayant pour pivots : $t_1 = 1$, $t_2 = 1+u, \ldots, t_m = 1+(m-1)u, \ldots, t_r = 1+(r-1)$ u=s; codage barycentrique effectué non sur la variable initiale x_j,mais sur la variable z_j prenant la valeur m ($1 \leq$ m $\leq$ s) pour une observation tombant dans la $m^{ème}$ tranche j_m de x_j. Cette façon de voir permet de ramener un codage disjonctif complet à s modalités ordonnées à un codage flou à r modalités avec r et s quelconques ($2 \leq$ r $\leq$ s); il suffit d'effectuer un codage barycentrique à r modalités sur la variable z_j considérée ci-dessus, avec pour pivots : $t_1 = 1$, $t_2 = 1 + (s-1)/(r-1), \ldots,$ $t_m = 1 + (m-1)(s-1)/(r-1), \ldots, t_r = s$.

V Fenêtre glissante

Quand on découpe une variable quantitative x_j en classes pour effectuer un codage disjonctif complet, on se heurte au problème suivant :

- ou bien on effectue, pour ne pas perdre trop d'information un codage fin comportant beaucoup de classes de faible amplitude, mais on risque d'avoir des résultats instables et peu reproductibles du fait de classes d'effectif très faible.
- ou bien on effectue un découpage grossier en un nombre de tranches relativement restreint, ces tranches étant de grande amplitude auquel cas on perd beaucoup d'information.

Pour éviter ces écueils, on peut effectuer plusieurs découpages sur cette variable, en changeant les limites des classes d'un découpage à l'autre, le nombre de classes, pas trop grand, étant en général le même pour chaque découpage. On peut prendre des classes de même amplitude (à part les classes extrêmes bien sûr) que l'on décale d'une même quantité d'un découpage au suivant. Toutes ces classes, à part les extrêmes, constituent les positions successives d'une fenêtre glissante de longueur fixée. On peut aussi prendre une fenêtre glissante d'effectif constant (i.e. contenant un nombre donné d'individus) et définir le décalage de celle-ci par le nombre d'individus qu'elle perd à gauche et celui qu'elle gagne à droite.

Si on a s découpages et si J_q $(1 \leq q \leq s)$ désigne l'ensemble des classes associé au q^{ieme} découpage, la variable x_j est caractérisée par s codages disjonctifs complets sur l'ensemble $J = U\,\{J_q/q = 1, s\}$.

Cette technique a en particulier été utilisée par C. ARBACHE (cf [1]) dans une étude où il s'agissait d'expliquer le prix de vente d'un bien (un appartement par exemple) en fonction d'un certain nombre de caractéristiques de ce bien.

Pour effectuer cette étude C. ARBACHE a classiquement découpé en tranches toutes les variables et construit le tableau croisant variable à expliquer et variables explicatives. C'est au niveau de la variable à expliquer qu'il a appliqué la technique de la fenêtre glissante, croisant ainsi l'ensemble des modalités de tous les découpages de cette variable (5 découpages en 7 tranches, soit 35 modalités en l'occurence) avec l'ensemble des modalités des variables explicatives.

Remerciements

L'auteur remercie le Professeur J.P. BENZECRI et G. CELEUX pour avoir relu son article et effectué des suggestions pour l'améliorer.

Bibliographie

[1] ARBACHE, Ch. (1985) : L'évaluation du prix de biens durables d'une espèce donnée [BIEN DURABLE], *C.A.D.*, Vol. X n° 4, pp. 401-411.

[2] BASTIN, C., BENZECRI, J.P., BOURGARIT, C., CAZES, P. (1980) : Pratique de l'analyse des données, Vol. 2 : *Abrégé théorique - Etudes de cas modèle* :

a) V n° 1 et 2 : Analyse et recodage d'un tableau de notes dédoublé.

b) V n° 7 : Codage et analyse des préférences.

c) V n° 8 : Représentation géométrique des préférences et tableaux de correspondance.

[3] BEHRAKIS, T., NICOLACOPOULOS, I. (1988) : Analyse des réponses de 2 000 électeurs à un thermomètre de sympathie vis à vis de personnalités politiques grecques [POLIT. GREC.], *C.A.D.*, Vol XIII n° 2, pp. 233-238.

[3bis] BENJELLOUN, A. (1990) : Thèse de doctorat, Un. de Paris VI, *à paraître*.

[4] BENZECRI, J.P. (1989) : Essai d'analyse des notes attribuées par un ensemble de sujets aux mots d'une liste [NOTES MOTS], *C.A.D.*, Vol. XIV n° 1, pp. 73-98.

[5] BENZECRI, J.P. et F. (1989) : Le codage linéaire par morceaux : réalisation et applications [CODAGE LIN.], *C.A.D.*, Vol XIV n° 2, pp. 203-210.

[6] BENZECRI, J.P. et F.(1989) : Codage linéaire par morceaux et équation personnelle [EQ. PERS.], *C.A.D.*, Vol XIV n° 3, pp. 331-336.

[7] CAZES, P. (1972) : Etude du dédoublement en analyse des correspondances. *Note du Laboratoire de Statistique*, Un. de Paris VI.

[8] CHAIEB, S. (1984) : Variation de la concentration plasmatique d'une substance au cours d'une perfusion et après celle-ci : cas du dinitrate d'isosorbide, [CONCENTRATION DNIS], *C.A.D.*, Vol IX n° 1, pp. 43-58.

[9] DUBOIS, A. (1984) : Applications de l'analyse factorielle des correspondances à l'étude d'une enquête relative au transport aérien, *Thèse de 3ème cycle*, Un. de Paris VI.

[10] ESCOFIER, B. (1979) : Traitement simultané de variables qualitatives et quantitatives en analyse factorielle [QUALITATIVES ET QUANTITATIVES], *C.A.D.*, Vol IV n° 2, pp. 137-146.

[11] GALLEGO, F.J. (1982) : Codage flou en analyse des correspondances [COD. FLOU], *C.A.D.*, Vol VII n° 4, pp. 413-430.

[12] LOSLEVER, P., GUERRAT, T.M., ROGER, D. (1988) : Analyse des questionnaires en ergonomie : l'appréciation des réglages d'un poste de travail [ERGO. REGLAGES], *C.A.D.* Vol XIII n° 2, pp. 175-196.

[13] MC GIBBON TAYLOR, B., LEDUC, P., TIBEIRO, J.S. (1989) : Analyse des réponses des étudiants à un questionnaire relatif au mémoire de recherche de la maitrise en administration des affaires [QUEST. MEM. RECH.], *C.A.D.*, Vol XIV n° 3, pp. 337-346.

[14] SKALLI, A. (1986) : Etude des propriétés olfactives des deux enanthiomorphes de l'α - ionone, [α - IONONE], *C.A.D.*, Vol XI n° 1, pp. 67-80.