CAHIERS DU BURO

P. CAZES
J. P. LECOUTRE

Étude de quelques problèmes de codage en analyse des correspondances

Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 27 (1977), p. 49-66

<http://www.numdam.org/item?id=BURO_1977__27__49_0>

ÉTUDE DE QUELQUES PROBLÈMES DE CODAGE EN ANALYSE DES CORRESPONDANCES

P. CAZES [1] et J.P. LECOUTRE [2]

---

(1) Maitre-Assistant à l'université Pierre et Marie Curie (Paris VI).

(2) Assistant à l'université de Paris II.

1 – INTRODUCTION

Notre propos est ici d'étudier l'influence des transformations d'un tableau de données brut en un tableau sur lequel on effectuera l'analyse des correspondances : dédoublement d'un tableau et application à l'analyse des préférences, transformation d'un tableau laissant invariants les facteurs de l'un des deux ensembles. Nous étudierons également l'influence du codage et des non réponses dans l'analyse d'un questionnaire. On verra en particulier qu'une question subdivisée en trop de modalités peut donner lieu à un facteur trivial opposant deux des modalités de cette question, et que les non réponses diminuent l'information (l'inertie) du tableau analysé. On verra également que si l'on ajoute une modalité "non réponse" pour chaque question et si les non réponses sont distribuées uniformément sur l'ensemble des individus et sur l'ensemble des questions, on obtient d'une part les facteurs que l'on aurait obtenus sans les non réponses, et d'autre part des facteurs de structure dûs à l'introduction des modalités de non réponse.

2 – ETUDE DU DEDOUBLEMENT EN ANALYSE DES CORRESPONDANCES.

2.1 Introduction. Notations.

Soit $(k_0)_{IJ} = \{k_0(i, j) \mid 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}$ un tableau de nombres positifs sur le produit de deux ensembles I et J de cardinaux respectivement égaux à n et p ; I correspond en général à un ensemble d'individus, et J à un ensemble de variables. Dans le cas où le tableau k_0 est un tableau de notes, pour analyser k_0, on effectue en général l'analyse factorielle des correspondances (AFC) de ce tableau. Dans ce cas, les éléments i de I n'ont pas en général le même poids. Pour donner même importance à chaque individu i, on dédouble l'ensemble J(1). On pose $J^+ = J$, $J' = J^+ \cup J^-$, avec

$$J^- = \{j^- \mid j \in J\},$$

j^- étant la colonne du tableau dédoublé où l'on inscrit le complément à a_j de la note dans la matière j, et l'on désigne par $k_{IJ'}$ le tableau dédoublé :

(1) L'analyse des correspondances d'un tableau de notes (ou de rangs) dédoublé semble d'un point de vue pratique donner des résultats plus clairs et plus facilement interprétables que l'analyse des correspondances du tableau initial.

$$\left.\begin{aligned} k\,(i\,,j^{+}) &= k_0\,(i\,,j) \\ k\,(i\,,j^{-}) &= a_j - k\,(i\,,j^{+}) = a_j - k_0\,(i\,,j) \end{aligned}\right\} \qquad (2.1.1)$$

où a_j peut être considéré comme la note maximale éventuellement affectée d'un coefficient de pondération, que l'on peut obtenir dans la matière j.

L'on posera également :

$$A = \sum_{j\in J} a_j\ ;\ b_j = a_j/A\ ;\ k\,(j^{+}) = \sum_i k\,(i\,,j^{+}) = \sum_i k_0\,(i\,,j) = k_0\,(j)$$

$$k\,(j^{-}) = \sum_i k\,(i\,,j^{-}) = n\,a_j - k_0(j)\ ;\ k\,(i) = \sum_{j\in J'} k\,(i\,,j) = A\ ;\ k = \sum_i k\,(i) = n\,A.$$

Du fait de la relation introduite par le dédoublement, il y a au plus p facteurs non triviaux dans l'AFC de $k_{IJ'}$.

2.2 Etude du nuage $N(J')$ des profils des colonnes du tableau $k_{IJ'}$.

Soit $\underset{(n\,,\,2\,p)}{Y} = (\underline{y}_{1+}\,, \ldots, \underline{y}_{p+}\,, \underline{y}_{1-}, \ldots, \underline{y}_{p-})$ le tableau associé au nuage $N(J')$ dans $\mathbf{R}^n$:

$$\underline{y}_{j+} = \frac{1}{k\,(j^{+})} \begin{pmatrix} k\,(1\,,j^{+}) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ k\,(n\,,j^{+}) \end{pmatrix} ;\ \underline{y}_{j-} = \frac{1}{k\,(j^{-})} \begin{pmatrix} k\,(1\,,j^{-}) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ k\,(n\,,j^{-}) \end{pmatrix}$$

le point j^{+} (resp. j^{-}) étant affecté de la masse $f_{j+} = k\,(j^{+})/k$ (resp. $f_{j-} = k\,(j^{-})/k$).

Le centre de gravité de $N(J')$ n'est autre que le vecteur $\underline{f}$ de composantes $f_i = k\,(i)/k = 1/n$.

De la seconde relation (2.1.1) qui peut encore s'écrire

$$f_{j+}\,\underline{y}_{j+} + f_{j-}\,\underline{y}_{j-} = (f_{j+} + f_{j-})\,\underline{f} \qquad (2.2.1)$$

l'on déduit que les trois points $\underline{y}_{j+}$, $\underline{y}_{j-}$ et $\underline{f}$ sont alignés, et si d^{α} désigne la $\alpha^{\text{ième}}$ composante principale ($\underline{d'}^{\alpha} = (d^{\alpha}_{1+}, \ldots, d^{\alpha}_{p+}, d^{\alpha}_{1-}, \ldots, d^{\alpha}_{p-})$), on a les p relations :

$$\forall\, j = 1, \ldots, p : f_{j+}\,d^{\alpha}_{j+} + f_{j-}\,d^{\alpha}_{j-} = 0. \qquad (2.2.2)$$

Toutes ces relations traduisent que le support du nuage $N(J')$ est au plus de dimension p.

2.3 Etude du nuage $N(I)$ des profils des lignes du tableau $k_{IJ'}$.

Soit $X = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ le tableau $(2p, n)$ associé au nuage $N(I)$:

$$\underline{x}_i = \frac{1}{k(i)} \begin{pmatrix} k(i, 1^+) \\ \cdot \\ \cdot \\ \cdot \\ k(i, p^-) \end{pmatrix} = \frac{1}{A} \begin{pmatrix} k(i, 1^+) \\ \cdot \\ \cdot \\ \cdot \\ k(i, p^-) \end{pmatrix} \in \mathbf{R}^{2p}.$$

Le tableau X n'est donc rien d'autre au facteur $1/A$ près que le tableau k' transposé de k.

Le point $\underline{x}_i$ est affecté de la masse $f_i = k(i)/k = 1/n$, et l'espace $\mathbf{R}^{2p}$ muni de la métrique du χ^2.

$$M = \text{diag}\,(1/f_{1+}, \ldots, 1/f_{p+}, 1/f_{1-}, \ldots, 1/f_{p-}).$$

Le carré de la distance du χ^2 entre $\underline{x}_i$ et $\underline{x}_{i'}$ peut s'écrire, en tenant compte de (2.1.1), sous la forme :

$$d^2(i, i') = \sum_{j=1}^{p} m_j \,(k_0(i, j) - k_0(i', j))^2$$

avec $$m_j = n^2 \, b_j / (k(j^+)\, k(j^-)).$$

Si l'on désigne par $M_0 = \text{diag}\,(m_1, \ldots, m_p)$ la métrique de $\mathbf{R}^p$ diagonale, de $j^{\text{ème}}$ terme diagonal m_j, et si l'on pose :

$$k_1(i, j) = \sqrt{m_j}\, k_0(i, j) \qquad (2.3.1)$$

les trois triplets (X, M, D_p), $(k_0', M_0, D_p)^{(1)}$, $(k_1', M_1, D_p)^{(1)}$ (où M_1 est la métrique usuelle de $\mathbf{R}^p$, et D_p la métrique diagonale des poids égale au facteur $\frac{1}{n}$ près à la métrique usuelle de $\mathbf{R}^n$) sont équivalents, en ce sens qu'ils ont mêmes composantes principales.

---

(1) k_0' et k_1' désignent respectivement les tableaux transposés de k_0 et k_1.

2.4 Aplications.

Des résultats précédents, il découle que l'on peut trouver toutes les caractéristiques associées à l'AFC de $k_{IJ'}$ à partir de l'ACP([1]) du tableau k_1, $\mathbf{R}^p$ étant muni de la métrique usuelle, ce qui revient à diagonaliser une matrice symétrique d'ordre p, au lieu d'une matrice d'ordre $2p$, ceci ne présentant bien sûr un intérêt que si $n > p$.

De façon précise, si $\underline{u}^\alpha$ désigne le $\alpha^{\text{ème}}$ vecteur propre de la matrice variance V_1 associée au tableau k_1 et correspondant à la valeur propre λ_α, et si $\underline{c}^\alpha$ désigne la composante principale associée, qui est aussi composante principale associée au nuage $N(I)$, on a :

$$c_i^\alpha = \sum_{j=1}^{p} \sqrt{m_j}\,(k_0(i,j) - k_0(j)/n)\,u_j^\alpha$$

c_i^α (resp. u_j^α) étant la $i^{\text{ème}}$ $(1 \leqslant i \leqslant n)$ (resp. $j^{\text{ème}}$ $(1 \leqslant j \leqslant p)$) composante de $\underline{c}^\alpha$ (resp. $\underline{u}^\alpha$), et l'on peut montrer (cf. [2]) que :

$$d_{j+}^\alpha = \left(\frac{k(j^-)}{k(j^+)}\,\frac{\lambda_\alpha}{b_j}\right)^{1/2} u_j^\alpha$$

avec $$d_{j-}^\alpha = -\,k(j^+)\,d_{j+}^\alpha/k(j^-).$$

Dans le cas particulier d'un tableau de 0 – 1 dédoublé

$$(a_j = 1\ ;\ A = \Sigma\, a_j = p\ ,\ b_j = a_j/A = 1/p\ ;\ m_j = n^2/(p\,k(j^+)\,k(j^-)))$$

la variance s_j^2 de la variable j (i.e. de $k_0(.,j)$) étant égale à $k(j^+)\,k(j^-)/n^2$, on a : $k_1(i,j) = k_0(i,j)/\sqrt{p}\,s_j$, ce qui revient au facteur $(1/\sqrt{p})$ près à réduire les variables initiales. Dans ce cas, l'AFC dédoublée (i.e. l'AFC de $k_{IJ'}$) se ramène à l'ACP sur matrice de corrélation de k_0. De façon précise, si e^α désigne la $\alpha^{\text{ème}}$ composante principale de l'ACP sur matrice de corrélation de k_0, et si $r_{j\alpha}$ désigne la corrélation entre $\underline{e}^\alpha$ et $k_0(.,j)$, on a (cf. [2]) :

$$c_i^\alpha = e_i^\alpha/\sqrt{p}$$

$$d_{j+}^\alpha = \sqrt{k(j^-)/k(j^+)}\,r_{j\alpha}$$

$$d_{j-}^\alpha = -\sqrt{k(j^+)/k(j^-)}\,r_{j\alpha}$$

e_i^α désignant la $i^{\text{ème}}$ $(1 \leqslant i \leqslant n)$ composante de $\underline{e}^\alpha$.

---

(1) Analyse en composantes principales.

3 – ANALYSE DES CORRESPONDANCES ET ANALYSE DES PREFERENCES.

Soit $k_0(i, j)$ un tableau de rangs : $k_0(i, j)$ est le rang (ou le classement) donné par l'individu i à l'objet j ($1 \leqslant i \leqslant n$; $1 \leqslant j \leqslant p$). Pour i fixé, l'ensemble des $\{k_0(i, j) \mid 1 \leqslant j \leqslant p\}$ constitue donc une permutation de $\{1, 2, \ldots, p\}$.

Pour analyser le tableau précédent, on peut soit :

a) faire l'ACP de k_0 sur matrice variance (analyse des préférences usuelles),

b) faire l'ACP de k_0 sur matrice de corrélation,

c) faire l'AFC de k_0,

d) faire l'AFC du tableau dédoublé $k_{IJ'}$, où

$$J' = J^+ \cup J^-,\ k(i, j^+) = k_0(i, j)\ ,\ k(i, j^-) = p + 1 - k_0(i, j)$$

($a_j = p + 1$ avec les notations du paragraphe 2).

Du point de vue de la représentation des individus, les quatre analyses précédentes sont respectivement équivalentes dans $\mathbf{R}^p$ muni de la métrique usuelle, chaque point i ayant masse $1/n$, à l'ACP des tableaux X, Y, Z, T de terme général respectivement donné par :

a) $x_i^j = (k_0(i, j) - g_j)$

b) $y_i^j = (k_0(i, j) - g_j)/s_j$,

c) $z_i^j = (k_0(i, j) - g_j)/\sqrt{p(p + 1)\, g_j/2}$

d) $t_i^j = (k_0(i, j) - g_j)/\sqrt{p(p + 1 - g_j)\, g_j}$

où $g_j = (1/n) \Sigma \{k_0(i, j) \mid i = 1, n\}$ désigne le rang moyen de l'objet j et s_j l'écart-type associé.

Les formules a) et b) sont évidentes : la formule c) résulte de la transformation usuelle en analyse des correspondances

$$z_i^j = \left(\frac{k_0(i, j)}{k_0(i)} - \frac{k_0(j)}{k_0}\right) / \sqrt{k_0(j)/k_0}$$

où $k_0(i)$, $k_0(j)$ et k_0 désignent dans le tableau $(k_0)_{IJ}$ respectivement la somme de la ligne i, de la colonne j, et de tous les éléments du tableau :

$$k_0(i) = p(p + 1)/2\ ,\ k_0(j) = n g_j\ ,\ k_0 = n p(p + 1)/2.$$

De plus $k_0(i)$ étant indépendant de i, chaque point i a même masse $1/n$.

Enfin la formule d) résulte après centrage (i.e. choix de l'origine au centre de gravité) de la formule (2.3.1) donnée dans l'étude du dédoublement en 2.3.

Interprétation des formules précédentes.

On voit que si tous les objets ont même rang moyen g ($g_j = g$, $\forall\ j \in J$), auquel cas $g = (p + 1)/2$, les analyses a) c) d) sont équivalentes. Si ce n'est pas le cas, l'analyse des préférences usuelles donne même pondération à chaque objet, tandis que l'AFC non dédoublée (analyse c)) fait jouer un rôle plus important aux objets bien classés (g_j faible) qu'aux objets mal classés (g_j élevé) ; l'AFC dédoublée (analyse d)) donnant un poids plus faible aux objets moyennement classés qu'aux objets bien ou mal classés ($(p + 1 - g_j)\, g_j$ est maximum pour $g_j = p + 1 - g_j = (p + 1)/2$). Enfin l'ACP sur matrice de corrélation (analyse b)) donne d'autant plus d'importance à un objet que les rangs qui lui sont attribués sont moins dispersés (s_j petit) et ceci indépendamment du classement moyen (i.e. de g_j) de cet objet.

4 – TRANSFORMATION D'UN TABLEAU LAISSANT INVARIANTS LES FACTEURS D'UN DES DEUX ENSEMBLES.

Soit $P = P_{IJ}$ un tableau de correspondances sur I x J (card $I = n$; card $J = p$, $\sum_{i,j}^{(n\,,\,p)} p_{ij} = 1$), p_I, p_J les lois marginales associées, D_{p_I} et D_{p_J} les métriques diagonales des poids respectivement associées dans $(\mathbf{R}^n)^*$ et $(\mathbf{R}^p)^*$ (où l'on a noté $(\mathbf{R}^n)^*$ (resp. $(\mathbf{R}^p)^*$) le dual de $\mathbf{R}^n$ (resp. $\mathbf{R}^p$)). Nous désignerons par $X = P' D_{1/p_I}$ et $Y = P D_{1/p_J}$ les tableaux respectivement associés aux lois conditionnelles des lignes et des colonnes de P, D_{1/p_I} (resp. D_{1/p_J}) étant l'inverse de D_{p_I} (resp. D_{p_J}).

Rappelons que les facteurs sur I (resp. J) sont vecteurs propres de $X' \circ Y'$ (resp. $Y' \circ X'$).

Soient K un ensemble de cardinal q, P_K une mesure sur K de masse totale 1, D_{pK} la métrique diagonale associée dans $(\mathbf{R}^q)^*$ dual de $\mathbf{R}^q$, et T' une application de $(\mathbf{R}^p)^*$ dans $(\mathbf{R}^q)^*$ telle que la norme induite par T' et D_{pK} sur $(\mathbf{R}^p)^*$ soit D_{pJ} :

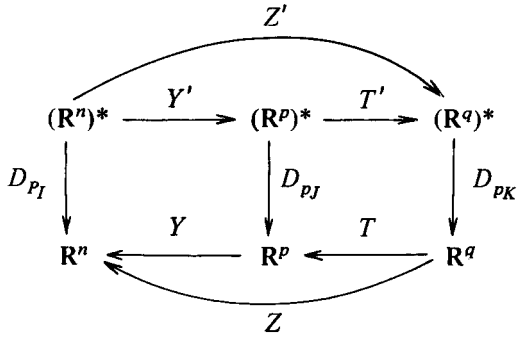
$$D_{pJ} = T\, D_{pK}\, T' \tag{4.1}$$

et telle que le vecteur constant $\underline{j}_p^{(1)}$ de $(\mathbf{R}^p)^*$ se transforme en le vecteur constant $\underline{j}_q^{(1)}$ de $(\mathbf{R}^q)^*$:

$$T' \underline{j}_p = \underline{j}_q. \tag{4.2}$$

Posons :

$$\underset{(n,q)}{Z} = \underset{(n,p)}{Y}\ \underset{(p,q)}{T}$$



Alors on a :

$$Z' \underline{j}_n = T'\ Y' \underline{j}_n = T' \underline{j}_p = \underline{j}_q.$$

Z peut donc être considéré comme un tableau de lois conditionnelles sur K, associé au tableau

$$Q = Z D_{P_K} = \underset{(n,p)}{Y}\ \underset{(p,q)}{T}\ \underset{(q,q)}{D_{P_K}}$$

Par ailleurs, on a :

$$Q \underline{j}_q = Y\,T\,D_{P_K} \underline{j}_q = Y\,T\,D_{P_K}\,T' \underline{j}_p = Y\,D_{P_J} \underline{j}_P = P\,\underline{j}_P$$

ce qui prouve que la somme des éléments de Q est égale à la somme des éléments de P, et que Q et P ont même loi marginale associée P_I sur I.

Si on fait l'AFC de Q, et si l'on pose $U = Q'\,D_{1/P_I}$, les facteurs associés sur I sont vecteurs propres de :

$$U' \circ Z' = D_{1/P_I}\,Y\,T\,D_{P_K}\,T'\,Y' = D_{1/P_I}\,Y\,D_{P_J}\,Y' = X' \circ Y'.$$

Q et P ont donc mêmes facteurs sur I. Il est immédiat de vérifier que les facteurs sur K de Q se déduisent des facteurs sur J de P par la transformation T'.

La transformation d'un tableau a été en particulier appliquée par F. Nakhlé dans sa thèse de 3e cycle pour l'étude d'un tableau de notes dédoublées $k_{IJ'}$ où $J' = J^+ \cup J^-$ (cf. 2e dont on conserve les notations),

(1) vecteur dont toutes les composantes valent 1.

On transforme tout couple (j^+, j^-) associé à une note en un nouveau couple noté (j_n^+, j_n^-) tel que si J_n^+ et J_n^- désignent respectivement l'ensemble des j_n^+ et des j_n^- et si $J'_n = J_n^+ \cup J_n^-$, le tableau transformé $k_{IJ'_n}$ soit tel que

$$\left.\begin{aligned} k_{IJ'_n}(i, j_n^+) &= c_j\, k_{IJ'}(i, j^+) + d_j = c_j\, k_{IJ'}(i, j^+) + \\ &\qquad + \frac{d_j}{a_j}\left(k_{IJ'}(i, j^+) + k_{IJ'}(i, j^-)\right) \\ k_{IJ'_n}(i, j_n^-) &= e_j - k_{IJ'_n}(i, j_n^+) = \left(\frac{e_j}{a_j} - \frac{d_j}{a_j} - c_j\right) k_{IJ'}(i, j^+) + \\ &\qquad + \left(\frac{e_j}{a_j} - \frac{d_j}{a_j}\right) k_{IJ'}(i, j^-) \end{aligned}\right\} \quad (4.3)$$

où c_j, d_j, e_j sont des constantes ne dépendant que de la matière j, et où l'on a tenu compte de la relation $k_{IJ'}(i, j^+) + k_{IJ'}(i, j^-) = a_j$.

Si P (resp. Q) désigne le tableau $(n, 2p)$ proportionnel à $k_{IJ'}$ (resp. $k_{IJ'_n}$) et dont la somme des éléments vaut 1, les relations précédentes peuvent se mettre avec les mêmes notations que plus haut où J est remplacé par J', K par J'_n, et p par $2p$, sous la forme :

$$\begin{aligned} Q = P\,R &= (P\,D_{1/pJ'})\,(D_{pJ'}\,R\,D_{1/pJ'_n})\,D_{pJ'_n} \\ &= Y\,.\,T\,.\,D_{pJ'_n} \end{aligned}$$

avec

$$T = D_{pJ'}\,R\,D_{1/pJ'_n}$$

L'ensemble I des individus aura la même représentation dans l'AFC de $k_{IJ'}$ et $k_{IJ'_n}$ si T' est une isométrie de $(\mathbf{R}^{2p})^*$ muni de la métrique $D_{pJ'}$ dans $(\mathbf{R}^{2p})^*$ muni de la métrique $D_{pJ'_n}$, isométrie laissant invariant le vecteur constant j_{2p}. L'intérêt d'une transformation comme (4.3) est par exemple de ramener l'intervalle théorique $0 - a$ de variation d'une note à l'intervalle réellement observé ; ainsi, si dans une matière les notes vont de 5 à 15, avec un coefficient égal à 4 (intervalle théorique 0 – 80 ; intervalle réel 20 – 60), il sera équivalent de ramener les notes à l'intervalle 0 – 20 avec un coefficient égal à 2.

5 – INFLUENCE DU CODAGE ET DES NON REPONSES DANS L'ANALYSE D'UN QUESTIONNAIRE.

5.1 Rappel – Equation des facteurs.

Soient $x_1, x_2, \ldots, x_r$, r variables qualitatives à respectivement $m_1, m_2, \ldots, m_r$ modalités mesurées sur n individus et X'_q $(1 \leqslant q \leqslant r)$ le tableau (n, m_q) des indicatrices des modalités de x_q. Nous désignerons par J_q l'ensemble des modalités de x_q, et par J l'union disjointe des J_q $(1 \leqslant q \leqslant r)$ et nous noterons :

$$\underset{(n, \Sigma m_i = m)}{X'} = (X'_1, X'_2, \ldots, X'_r)$$

le tableau disjonctif complet associé aux r variables précédentes ; nous poserons également :

$$B = XX' = \begin{pmatrix} P_{11} & \cdots & P_{1r} \\ \vdots & & \vdots \\ P_{r1} & \cdots & P_{rr} \end{pmatrix}$$

$$B_1 = \text{diag}\,(XX') = \begin{pmatrix} P_{11} & & 0 \\ & \ddots & \\ 0 & & P_{rr} \end{pmatrix}$$

où $P_{ij} = X_i X'_j$ est le tableau de contingence $m_i \times m_j$, croisant les variables x_i et x_j, le tableau P_{ii} n'étant autre que le tableau diagonal des effectifs des modalités de x_i. B n'est autre que le tableau de Burt associé aux variables x_i, i.e. le tableau croisant l'ensemble J des modalités de toutes les variables avec elles-mêmes.

Si $\underset{(m, 1)}{\underline{a}} = \begin{pmatrix} \underline{a}_1 \\ \vdots \\ \underline{a}_r \end{pmatrix}$ est un facteur, sur l'ensemble J des modalités des x_q, de l'analyse des correspondances de X', associé à la valeur propre λ, $\underline{a}_q$ étant la restriction de $\underline{a}$ aux modalités de x_q, on a :

$$B\,\underline{a} = \lambda\, r\, B_1\, \underline{a} \qquad (5.1.1)$$

soit encore

$$\Sigma \{P_{qq'} \underline{a}_{q'} \mid q' = 1, r\} = r \lambda P_{qq} \underline{a}_q$$

ou
$$\Sigma \{P_{qq'} \underline{a}_{q'} \mid q' = 1, r ; q' \neq q\} = (r\lambda - 1) P_{qq} \underline{a}_q . \qquad (5.1.2)$$

Rappelons que $\underline{a}$ est également un facteur de l'analyse des correspondances du tableau B, relatif à la valeur propre $\mu = \lambda^2$.

Si $\underline{b}$ est un facteur non trivial (i.e. de moyenne nulle $\sum_{i=1}^{n} b_i = 0$)[1] de X' sur l'ensemble I des n individus, et si A_q désigne dans $\mathbf{R}^n$ muni de la métrique $\frac{1}{n} U_n$ (U_n étant la métrique unité de $\mathbf{R}^n$) l'opérateur de projection sur le sous-espace vectoriel engendré par les indicatrices de x_q (i.e. les colonnes de X_q'), on a :

$$\Sigma \{A_q \underline{b} \mid q = 1, r\} = \lambda r \underline{b}$$

et l'on a, si $\underline{a}$ et $\underline{b}$ ont même variance :

$$X' \underline{a} = r \sqrt{\lambda}\, \underline{b}$$
$$X \underline{b} = \sqrt{\lambda}\, B_1 \underline{a},$$

cette dernière relation pouvant encore s'écrire :

$$\forall q \in (1, \ldots, r) : X_q \underline{b} = \sqrt{\lambda}\, P_{qq} \underline{a}_q .$$

Le nombre des facteurs non triviaux (i.e. centrés et relatifs à une valeur propre non nulle) de l'AFC de X' est au plus égal à $m - r$, chaque facteur non trivial $\underline{a}$ sur J étant centré sur chaque J_q i.e. étant tel que :

$$\underline{j}'_{mq} P_{qq} \underline{a}_q = 0, \qquad (5.1.3)$$

$\underline{j}_{mq}$ désignant le vecteur $(m_q, 1)$ de composantes toutes égales à 1.

Les r facteurs triviaux restant sont des facteurs constants sur chaque J_q, $\underline{a}_q = \alpha_q \underline{j}_{mq}$ et sont donc caractérisés par le vecteur $\underline{\alpha}$ de $\mathbf{R}^r$ de composantes α_q ($1 \leqslant q \leqslant r$). Munissant $\mathbf{R}^r$ de la métrique $(1/r)\, U_r$ où U_r est la métrique usuelle de $\mathbf{R}^r$, on obtient ces r facteurs à partir d'un système orthonormé de vecteurs $\{\underline{\alpha}_i \mid i = 1, \ldots, r\}$ de $\mathbf{R}^r$, dont le premier $\underline{\alpha}_1 = \underline{j}_r$ est le vecteur constant dont toutes les composantes valent 1, et qui correspond au facteur trivial $\underline{a} = \underline{j}_m$ associé à la valeur propre 1.

---

(1) b_i désignant la $i^{\text{ème}}$ composante de $\underline{b}$

5.2 Influence de la subdivision d'une modalité dans un tableau disjonctif complet.

Supposons qu'une modalité $z \in J$ soit divisée en deux sous-modalités z' et z'' de telle sorte que si $J_d = (J - \{z\}) \cup \{z', z''\}$, on ait, si B_d désigne le nouveau tableau de Burt associé à J_d :

$$\forall j, j' \in J - \{z\} : B_d(j, j') = B(j, j')$$

$$B_d(j, z') = \alpha B(j, z)$$

$$B_d(j, z'') = (1 - \alpha) B(j, z)$$

$$B_d(z', z') = \alpha B(z, z)$$

$$B_d(z'', z'') = (1 - \alpha) B(z, z),$$

avec bien sûr : $B_d(z', z'') = 0,$

B désignant toujours le tableau de Burt associé à J, et α un nombre compris entre 0 et 1.

Si l'on pose $J_0 = J - \{z\}$, on a les partitions suivantes de B et B_1

$$B = \begin{array}{c} \\ J_0 \\ z \end{array} \begin{array}{c} \begin{array}{cc} J_0 & z \end{array} \\ \begin{pmatrix} B_{00} & B_{0z} \\ B_{z0} & B(z, z) \end{pmatrix} \end{array} ; B_1 = \begin{array}{c} \\ J_0 \\ z \end{array} \begin{array}{c} \begin{array}{cc} J_0 & z \end{array} \\ \begin{pmatrix} (B_1)_{00} & 0 \\ 0 & B(z, z) \end{pmatrix} \end{array}.$$

Soit X'_d le tableau disjonctif complet associé à B_d, et désignons par $\underline{a}_d = \begin{pmatrix} \underline{a}_0 \\ a_{z'} \\ a_{z''} \end{pmatrix}$ un facteur de X'_d (qui est aussi facteur de B_d) relatif à la valeur propre λ_d ; on a d'après (5.1.1) où B et B_1 sont remplacés par B_d et diag (B_d) :

$$B_{00}\,\underline{a}_0 + (\alpha\, a_{z'} + (1 - \alpha)\, a_{z''})\, B_{0z} = r\,\lambda_d\,(B_1)_{00}\,\underline{a}_0$$

$$B_{z0}\,\underline{a}_0 + B(z, z)\, a_{z'} = r\,\lambda_d\, B(z, z)\, a_{z'}$$

$$B_{z0}\,\underline{a}_0 + B(z, z)\, a_{z''} = r\,\lambda_d\, B(z, z)\, a_{z''}$$

d'où l'on déduit que :

1) Si $\underline{a} = \begin{pmatrix} \underline{a}_0 \\ a_z \end{pmatrix}$ est facteur de X' (ou de B) associé à la valeur propre

λ , $\underline{a}_d = \begin{pmatrix} \underline{a}_0 \\ a_z \\ a_z \end{pmatrix}$ est facteur de X'_d (ou de B_d) associé à la même valeur propre.

2) $\underline{a}_d = \begin{pmatrix} \underline{0} \\ 1-\alpha \\ -\alpha \end{pmatrix}$ est facteur de X'_d relatif à la valeur propre $\lambda_d = 1/r$.

Ainsi la subdivision de la modalité z en deux sous-modalités z' et z'' entraine la création d'un facteur que l'on peut qualifier de trivial puisqu'il oppose ces deux sous-modalités, et qu'il est nul pour toutes les autres modalités.

5.3 Influence des non réponses sur les facteurs quand on rajoute pour chaque question une modalité associée à la non réponse.

$\forall\, q \in (1, \ldots, r)$, on posera $M_q = J_q \cup q_0$ où q_0 est la modalité "non réponse" à la question q, i.e. à x_q, et l'on désignera par M la réunion des M_q, par Y' le tableau disjonctif complet construit sur $I \times M$, et par $C = YY'$ le tableau de Burt associé.

Pour pouvoir étudier l'influence des non réponses, nous allons former un modèle permettant de déduire C du tableau de Burt $B = XX'$ que l'on aurait obtenu dans le cas où il n'y a pas de données manquantes :

$$\left.\begin{array}{l}
\forall\, q \in (1, \ldots, r),\ \forall\, j, j' \in J_q : \\
\qquad C(j, j') = (1-\gamma)\, B(j, j') \\
\forall\, q, q' \in (1, \ldots, r),\ \forall\, j \in J_q,\ \forall\, j' \in J_{q'} : \\
\qquad q' \neq q \Rightarrow C(j, j') = (1-\zeta)\, B(j, j') \\
\qquad \forall\, j \in J_q : \\
\qquad q' \neq q \Rightarrow C(j, q'_0) = \beta\, B(j, j) \\
\forall\, q \in (1, \ldots, r),\ \forall\, j \in J_q : \quad C(j, q_0) = 0 \\
\qquad\qquad\qquad\qquad\quad C(q_0, q_0) = \epsilon\, n \\
\forall\, q, q' \in (1, \ldots, r) : q \neq q' \ \Rightarrow\ C(q_0, q'_0) = \alpha\, n,
\end{array}\right\} \qquad (5.3.1)$$

C[1] peut encore s'écrire si $\underline{P}_q$ désigne le vecteur $(m_q, 1)$ de composantes $P_{qq}(j, j)$ (si $j \in J_q$), $\underline{P}$ le vecteur $(m, 1)$ tel que $\underline{P}' = (\underline{P}'_1, \underline{P}'_2, \ldots, \underline{P}'_r)$,

(1) Nous adopterons l'ordre $J_1, J_2, \ldots, J_r, l_0, \ldots, q_0, \ldots, r_0$ $(= J, l_0, \ldots, r_0)$ pour partitionner $\underline{c}$, C et C_1.

j_r le vecteur $(r, 1)$ dont toutes les composantes valent 1, et U_r la matrice unité d'ordre r.

$$C = \begin{array}{c} \overbrace{}^{J} \\ J_1 \quad \cdots \quad J_r \qquad 1_0 \quad \cdots \quad r_0 \\ \left(\begin{array}{c|c} (1-\zeta)\,B + (\zeta-\gamma)\,B_1 & \beta\,(\underline{P}\,\underline{j}'_r - R) \\ \hline \beta\,(\underline{j}_r\,\underline{P}' - R') & n\,(\alpha\,\underline{j}_r\,\underline{j}'_r + (\epsilon-\alpha)\,U_r) \end{array} \right) \begin{array}{l} J \\ 1_0 \\ \vdots \\ r_0 \end{array} \end{array} \qquad (5.3.2)$$

avec

$$R = \begin{array}{c} \\ J_1 \\ J_2 \\ \vdots \\ J_r \end{array} \begin{array}{c} \begin{array}{cccc} 1_0 & 2_0 & \cdots & r_0 \end{array} \\ \begin{pmatrix} \underline{P}_1 & & & \\ & \underline{P}_2 & & 0 \\ & & \ddots & \\ 0 & & & \underline{P}_r \end{pmatrix} \end{array}$$

Pour que C puisse être considéré comme un tableau de Burt les paramètres α, β, γ, ϵ, ζ doivent être liés par les relations :

$$\left.\begin{aligned} \beta + \gamma &= \zeta \\ \alpha + \beta &= \epsilon \\ \gamma &= \epsilon \end{aligned}\right\} \qquad (5.3.3)$$

que nous supposerons vérifiées.

Si l'on suppose de plus que les non réponses sont distribuées uniformément sur l'ensemble des sujets et des questions du questionnaire, les paramètres précédents sont liés par la relation supplémentaire $\alpha = \epsilon^2$; dans ce cas ϵ désigne la proportion des non réponses, et l'on a :

$$\beta = \epsilon\,(1-\epsilon)\ ;\ \gamma = \epsilon\ ;\ \zeta = \beta + \gamma = \epsilon\,(2-\epsilon).$$

Si $\underline{c}$ (1) est un facteur du tableau de Burt C (1) relatif à la valeur propre μ^2, il vérifie l'équation.

$$C\,\underline{c} = r\,\mu\,C_1\,\underline{c}, \qquad (5.3.4)$$

(1) cf. (1) en bas de la page précédente

équation analogue à (5.1.1) et où la matrice C_1 [1] est la matrice diagonale ayant mêmes éléments diagonaux que C :

$$C_1 = \begin{array}{c} \\ J \\ l_0 \\ \vdots \\ r_0 \end{array} \begin{array}{c} \begin{array}{cc} J & l_0 \quad \cdots \quad q_0 \quad \cdots \quad r_0 \end{array} \\ \left(\begin{array}{c|c} (1-\gamma)\, B_1 & 0 \\ \hline 0 & n \in U_r \end{array} \right) \end{array} \tag{5.3.5}$$

U_r désignant toujours la matrice unité d'ordre r.

Explicitant (5.3.4) sous une forme analogue à (5.1.2), on déduit que si $\underline{a}$ est un facteur non trivial de B (i.e. $\underline{a}$ vérifie (5.1.2) et (5.1.3)), relatif à la valeur propre λ^2, $\underline{c} = \begin{pmatrix} \underline{a} \\ \underline{0} \end{pmatrix}$ est facteur de C, i.e. vérifie (5.3.4), avec (cf. [4]) et [5])) :

$$\frac{1-\gamma}{1-\zeta}\,(r\,\mu - 1) = r\,\lambda - 1.$$

On obtient ainsi $m - r$ facteurs pour C. C étant d'ordre $m + r$, il reste donc $2r$ facteurs à expliciter.

Si $\underline{c} = \begin{pmatrix} \underline{b} \\ \underline{v}_1 \\ \vdots \\ v_r \end{pmatrix} = \begin{pmatrix} \underline{b} \\ \underline{v} \end{pmatrix}$ est l'un de ces facteurs, il doit être orthogonal pour la métrique C_1 aux $m - r$ facteurs $\begin{pmatrix} \underline{a} \\ \underline{0} \end{pmatrix}$ précédemment trouvés :

$$< \begin{pmatrix} \underline{a} \\ \underline{0} \end{pmatrix}, \begin{pmatrix} \underline{b} \\ \underline{v} \end{pmatrix} >_{C_1} = (1-\gamma) < \underline{a}\,,\, \underline{b} >_{B_1} = 0\,,$$

$\underline{b}$ devant être orthogonal (pour B_1) aux $m - r$ facteurs non triviaux de B est donc combinaison linéaire des r facteurs triviaux de B (y compris le facteur

(1) cf. (1) en bas de la page 61.

trivial constant) qui sont constants sur chaque J_q ; on a donc :

$$\underline{b} = \begin{pmatrix} u_1 \underline{j}_{m1} \\ \vdots \\ u_r \underline{j}_{m_r} \end{pmatrix} ; \quad \underline{c} = \begin{pmatrix} u_1 \underline{j}_{m1} \\ \vdots \\ u_r \underline{j}_{m_r} \\ v_1 \\ \vdots \\ v_r \end{pmatrix} \tag{5.3.6}$$

et l'on peut montrer que le rapport v_q/u_q est indépendant de q. Si $\underline{\psi}$ désigne un vecteur de $\mathbf{R}^r$, de composantes $\psi_1 \ldots \psi_r$, on posera donc :

$$u_q = u\ \psi_q \ ; v_q = v\ \psi_q .$$

Pour $\underline{\psi} = \underline{j}_r$ on obtient d'une part le facteur trivial constant ($u = v$) associé à la valeur propre 1, et d'autre part un facteur, associé à un couple (u, v) ($u \neq v$) que l'on peut déterminer directement en écrivant qu'il est de moyenne nulle (i.e. orthogonal pour la métrique C_1 au facteur trivial), et dont on désignera par μ_1 la valeur propre correspondante.

Les autres facteurs de la forme (5.3.6), étant orthogonaux aux deux facteurs que l'on vient d'obtenir, sont tels que $\Sigma\ \psi_q = 0$, i.e. $\underline{\psi}'\ \underline{j}_r = 0$.

Si donc $\underline{\psi}$ est tel que $\Sigma\ \psi_q = 0$, on obtient à nouveau deux facteurs, l'un trivial (car centré et constant sur chaque M_q) tel que $u = v$ et associé à la valeur propre nulle, l'autre associé à un couple (u, v) ($u \neq v$) que l'on peut déterminer directement en écrivant qu'il est de moyenne nulle, et correspondant à une valeur propre μ_2.

Faisant choix dans $\mathbf{R}^r$ d'un système de $r - 1$ vecteurs $\underline{\psi}$ orthogonaux 2 à 2 et orthogonaux à $\underline{j}_r$ (pour la métrique usuelle U_r de $\mathbf{R}^r$) on obtient ainsi les $(r - 1)$ facteurs triviaux de C, et $(r - 1)$ facteurs relatifs à la valeur propre μ_2 qui est de multiplicité $r - 1$.

Ainsi avec le modèle adopté, on obtient comme facteurs non triviaux les facteurs non triviaux que l'on aurait obtenus dans le cas où il n'y a pas de non réponses, et des facteurs de structure, en nombre égal à r, constants sur chaque J_q, et dûs à l'introduction des modalités de "non réponse".

Si $\alpha = \epsilon^2$ (cas de non réponses distribuées uniformément), alors $\mu_1 = \mu_2 = 1/r$ ce qui signifie que les r facteurs de structure précédents correspondent à une valeur propre multiple d'ordre r.

5.4 Influence sur l'inertie des non réponses dans un questionnaire binaire quand on code les non réponses par (1/2, 1/2).

Rappelons que dans un questionnaire quelconque, la contribution $CR\ (j)$ d'une modalité j $(j \in J_q)$ à la trace $CR\ (J)$ dans l'AFC du tableau disjonctif complet X' des réponses est :

$$CR\ (j) = (1 - p_j)/r$$

où $p_j = P_{J_q J_q}\ (j, j)/n$ désigne la fréquence des individus ayant fourni la réponse j à la question q ; la contribution $CR\ (q)$ de la question q vaut :

$$CR\ (q) = (m_q - 1)/r$$

tandis que la trace vaut :

$$CR\ (J) = (m - r)/r,$$

m_q désignant toujours le nombre des modalités de q, m la somme des m_q(i.e. le nombre total de modalités) et r le nombre de questions.

Dans le cas d'un questionnaire ne comportant que des questions binaires $(m_q = 2\ ,\ J_q = \{q^+\ ,\ q^-\})$, $\forall\, q \in \{1, \ldots, r\}$, on a :

$$CR\ (q^+) = (1 - p_{q+})/r = p_{q-}/r$$

$$CR\ (q^-) = p_{q+}/r$$

$$CR\ (q) = 1/r$$

$$CR\ (J) = 1.$$

Supposons maintenant, le questionnaire restant toujours binaire, qu'il y ait des non réponses. On codera $x_i^{q+} = x_i^{q-} = 1/2$ si l'individu i n'a pas répondu à la question q, et l'on désignera par p_{q0} la proportion des individus n'ayant pas répondu à la question q. Si l'on pose : $p_q = p_q^+ + p_{q0}/2$, les contributions précédemment calculées valent maintenant :

$$CR\ (q^+) = \left(1 - p_q - \frac{p_{q0}}{4\,p_q}\right)/r$$

$$CR\ (q^-) = \left(p_q - \frac{p_{q0}}{4\,(1 - p_q)}\right)/r$$

$$CR\ (q) = \left(1 - \frac{p_{q0}}{4\,p_q\,(1 - p_q)}\right)/r$$

$$CR\ (J) = 1 - \frac{1}{4\,r}\,\Sigma\left\{\frac{p_{q0}}{p_q\,(1 - p_q)} \mid q = 1\ ,\ r\right\}.$$

Pour p_{q_0} fixé, $\forall q = 1, \ldots, r$, $CR(J)$ sera maximum si $p_q = 1 - p_q = 1/2$, ce maximum valant $1 - \bar{p}_{q_0}$ si $\bar{p}_{q_0}$ désigne la moyenne des p_{q_0}.

Cette inertie maximale est donc plus faible que l'inertie $CR(J) = 1$ que l'on aurait obtenue s'il n'y avait pas eu de non réponses.

BIBLIOGRAPHIE

[1] J.P. BENZECRI – Sur l'analyse d'un tableau de notes dédoublées. Application aux épreuves du concours d'admission à l'école Polytechnique. Publications du Laboratoire de Statistique, juin 1975.

[2] P. CAZES – Etude du dédoublement d'un tableau en analyse des correspondances. Note du Laboratoire de Statistique, juin 1972.

[3] Fouad NAKHLE – Analyse d'un tableau dédoublé et transformé. Application à l'étude des épreuves d'un concours. Thèse de 3e cycle, Université Paris VI, juin 1973.

[4] J.P. LECOUTRE – Sur l'analyse des questionnaires mis sous forme disjonctive complète. Convergence et optimisation de certains estimateurs des densités de probabilité. Thèse de 3e cycle. Université Paris VI, novembre 1975.

[5] Annales du D.E.A. de statistique de M. BENZECRI (avril 1973, juin et septembre 1974).