La revue de Modulad

QUELQUES CONSIDERATIONS SUR L'UTILISATION DES ELEMENTS SUPPLEMENTAIRES EN ANALYSE FACTORIELLE

P. CAZES

Université Paris 9 Dauphine

Résumé :

Dans cet article, on fait le point sur l'utilisation des éléments supplémentaires en Analyse Factorielle (Composantes Principales et Correspondances) et sur un certain nombre de pratiques associées. Après des rappels, on examine l'intérêt des éléments supplémentaires quand l'ensemble des observations est muni d'une partition ou quand on a un tableau ternaire. L'application des éléments supplémentaires aux méthodes de régression ou de discrimination après analyse factorielle est également détaillée.

Mots-clés : Analyse en Composantes Principales, Analyse Factorielle des Correspondances, Discrimination, Eléments Supplémentaires, Partition, Régression, Tableaux Ternaires.

1. INTRODUCTION

Les éléments supplémentaires jouent un rôle très important en analyse des données. Ils permettent en particulier de ne pas perdre une information de nature un peu différente de celles qui sont analysées (observation ou variable effectivement de nature différente ; observation douteuse, donnée aberrante, etc.). Ils sont également très utiles pour faciliter l'interprétation d'une analyse factorielle (variables illustratives - projection du centre de gravité d'un groupe d'observations, etc.) et interviennent dans certains enchaînements d'analyses (par exemple dans certaines méthodes de régression après analyse factorielle, ou en analyse factorielle discriminante) ainsi que dans certaines techniques comme l'analyse des tableaux ternaires.

Nous nous proposons ici de faire le point sur l'utilisation des éléments supplémentaires en Analyse Factorielle (Composantes Principales et Correspondances) et sur un certain nombre de pratiques associées

Après des rappels sur les éléments supplémentaires en Analyse en Composantes Principales (A.C.P.) et sur leur utilisation, nous traitons du cas où l'ensemble des individus est muni d'une partition, puis nous examinons le cas des tableaux ternaires. Les résultats obtenus sont alors transposés au cas où l'on remplace l'A.C.P. par l'Analyse des Correspondances, puis nous étudions l'intérêt des éléments supplémentaires dans les méthodes de régression ou de discrimination après analyse factorielle.

2. RAPPELS SUR LES ELEMENTS SUPPLEMENTAIRES EN ANALYSE EN COMPOSANTES PRINCIPALES

Soit $X = \{x_{ij} \mid i \in I, j \in J\}$ un tableau individus x variables, x_{ij} désignant la valeur de la variable x_j pour l'individu i ($1 \leq i \leq n$, $1 \leq j \leq p$). Nous supposerons que chaque individu i est muni d'une masse p_i ($\Sigma p_i = 1$, en général $p_i = 1/n$) et que les variables sont centrées ($\forall j \in J : \Sigma \{p_i x_{ij} \mid i \in I\} = 0$). Nous supposerons également l'espace $\mathbb{R}^p$ muni de la métrique usuelle, auquel cas l'Analyse en Composantes Principales (A.C.P.) de X correspond à l'A.C.P. normée (ou sur matrice de corrélation) si les variables x_j sont réduites ($\Sigma \{p_i (x_{ij})^2 \mid i \in I\} = 1$) et à l'A.C.P. sur matrice variance sinon. Nous désignerons par $F_\alpha(i)$ la projection de l'individu i sur le $\alpha^{ème}$ axe factoriel $\Delta \mathbf{u}_\alpha$ (de vecteur unitaire $\mathbf{u}_\alpha$) issu de l'A.C.P. de X et par $G_\alpha(j)$ ($= u_{j\alpha} \sqrt{\lambda_\alpha}$, $u_{j\alpha}$ étant la $j^{ème}$ composante de $\mathbf{u}_\alpha$ et λ_α la valeur propre associée à cet axe) la projection de la variable j sur la composante principale normée $F_\alpha / \sqrt{\lambda_\alpha}$ ($G_\alpha(j)$ n'est autre que la covariance entre x_j et $F_\alpha / \sqrt{\lambda_\alpha}$, et la corrélation entre F_α et x_j si x_j est réduite)

Une ligne supplémentaire i_s rajoutée au tableau X sera représentée sur l'axe $\Delta \mathbf{u}_\alpha$ par sa projection qui est le point d'abscisse $F_\alpha(i_s)$ définie par :

$$F_\alpha(i_s) = \Sigma \{u_{j\alpha} x_{i_s j} \mid j \in J\} = \Sigma \{G_\alpha(j) x_{i_s j} \mid j \in J\} / \sqrt{\lambda_\alpha} \qquad (1)$$

$x_{i_s j}$ désignant la valeur de x_j pour l'observation supplémentaire i_s.

Représenter i_s en même temps que les n éléments de I sur les axes factoriels issus de l'A.C.P. du tableau X revient à faire l'A.C.P. du tableau associé à ces n + 1 points, en donnant à i_s un poids nul

Comme pour un élément principal i de I, on peut chiffrer la qualité de la représentation de i_s sur un axe (ou sur un sous-espace factoriel) à l'aide du carré du cosinus de l'angle entre i_s et cet axe (ou ce sous-espace).

De même, soit j_s une colonne supplémentaire (i.e. une variable x_{j_s} mesurée sur les n individus, mais n'ayant pas participé à l'A.C.P. de X). On pourra représenter j_s sur la composante principale normée $F_\alpha / \sqrt{\lambda_\alpha}$ par sa projection ($\mathbf{R}^n$ étant muni de la métrique diagonale des poids p_i) qui est le point d'abscisse $G_\alpha (j_s)$ donnée par :

$$G_\alpha (j_s) = \sum \{p_i F_\alpha (i) x_{ij_s} \mid i \in I\} / \sqrt{\lambda_\alpha} \qquad (2)$$

x_{ij_s} étant la valeur de x_{j_s} pour l'individu i.

Si on a effectué l'A.C.P. normée, il faut supposer que la variable x_{j_s} a été centrée et réduite au préalable, $G_\alpha (j_s)$ représentant alors la corrélation entre x_{j_s} et la composante principale (qu'on appelera aussi facteur par la suite) F_α.

On pourra, comme pour une variable principale x_j, juger de la qualité de la représentation de x_{j_s} sur le facteur α en calculant la corrélation (ou son carré) entre x_{j_s} et ce facteur.

De façon similaire au cas d'une ligne supplémentaire, représenter $x_1, \ldots, x_p, x_{j_s}$ sur les facteurs issus de l'A.C.P. de X, revient à faire l'A.C.P. de ces p + 1 variables, l'espace $\mathbf{R}^{p+1}$ étant muni de la pseudométrique dont la matrice associée est diagonale, tous les éléments diagonaux valant 1, sauf le dernier qui est nul.

Remarque : On représente souvent une variable x_j sur un facteur α par sa corrélation avec ce facteur, corrélation qu'on notera $G'_\alpha (j)$ et qui est égale à $G_\alpha (j) / \sigma_j$, σ_j étant l'écart type de x_j (égal à 1 en A.C.P. normée). Dans ce cas on représentera la variable supplémentaire x_{j_s} par sa corrélation $G'_\alpha (j_s) = G_\alpha (j_s) / \sigma_{j_s}$ avec F_α, σ_{j_s} étant l'écart type de x_{j_s} (égal à 1 en A.C.P. normée). La représentation des variables (principales et supplémentaires) dans le plan associé à deux facteurs α et β se fait alors à l'intérieur du traditionnel cercle des corrélations, représentation que l'on n'obtenait en utilisant la formule (2) que dans le cas de l'A.C.P. normée.

3. UTILISATION DES ELEMENTS SUPPLEMENTAIRES EN ANALYSE EN COMPOSANTES PRINCIPALES

On peut se servir des éléments supplémentaires pour représenter une observation relevée dans des conditions douteuses (ou différentes des autres observations) ou encore une variable sur laquelle la précision est moindre que sur les autres variables mesurées. On peut aussi utiliser cette technique pour représenter un élément aberrant, ou un élément (variable ou observation) ayant perturbé une analyse préliminaire, sans perdre complètement l'information associée à cet élément.

Les éléments supplémentaires permettent également de placer un cas nouveau sur les axes factoriels issus d'une analyse antérieure. Ce cas est fréquent en médecine où après avoir étudié un certain nombre de patients durant une période de temps donnée, il se présente de nouveaux malades qu'il est intéressant de situer par rapport aux patients déjà analysés. Ce cas intervient également quand l'ensemble des observations

représente le temps et qu'ayant effectué une A.C.P., on désire regarder où se placent les instants ultérieurs.

Dans le cas où l'on a des éléments (en général des variables) de nature différente, il est d'usage d'effectuer l'A.C.P. sur un groupe de variables relativement homogènes et de placer le reste des variables en éléments supplémentaires. Par exemple, si on a un ensemble de sédiments (ou de roches) pour lesquels on connait la composition en éléments majeurs (exprimés en pourcentages) et sur lesquels on a mesuré des éléments trace (exprimés en p.p.m. : partie par million) on fera l'A.C.P. du tableau associé aux éléments majeurs en considérant les éléments trace comme des éléments supplémentaires.

Une autre application des éléments supplémentaires réside dans la visualisation sur les plans factoriels d'un groupe d'individus, ce groupe étant représenté par son centre de gravité, et donc sur un plan factoriel par la projection de ce point.

Cette application présente une importance pratique considérable ; le groupe d'individus que l'on veut visualiser peut être soit une classe issue d'une classification automatique, soit les individus ayant pris la même modalité d'une variable qualitative. En particulier, si le nombre d'individus est très élevé, on projettera sur les plans issus d'une analyse factorielle faite sur ces individus, non ces individus, mais des groupes d'individus : ces groupes peuvent correspondre aux classes obtenues par des méthodes comme la méthode des nuées dynamiques par exemple ; ils peuvent aussi correspondre aux individus associés aux différentes modalités de variables illustratives (pour un exemple, on pourra consulter l'étude des dépenses de 841 ménages où l'on projette les classes d'individus correspondant aux différentes C.S.P., aux tranches de revenus etc., exemple donné dans Lebart et coll. (1979)).

Même si le nombre d'individus n'est pas très élevé, on a intérêt à projeter des groupes d'individus sur les axes factoriels, soit pour faciliter l'interprétation (cas des modalités de variables illustratives) soit quand les groupes d'individus correspondent aux classes d'une classification automatique pour comparer les résultats de cette classification avec ceux de l'analyse factorielle

La représentation d'un groupe d'individus joue un rôle important quand l'ensemble des individus est muni d'une partition (cf §4.) ou quand l'on a un tableau ternaire (cf §5.).

4. CAS OU L'ENSEMBLE DES INDIVIDUS I EST MUNI D'UNE PARTITION

On suppose ici que I est muni d'une partition en r classes, classes indicées par un ensemble Q :

$$I = \cup \{I_q \mid q \in Q\} \qquad (3)$$

On désignera par $G = \{g_{qj} \mid q \in Q, j \in J\}$ le tableau r x p des centres de gravité :

$$g_{qj} = \Sigma \{p_i x_{ij} \mid i \in I_q\} / p_q$$

avec

$$p_q = \Sigma \{p_i \mid i \in I_q\}$$

p_i désignant, rappelons-le, la masse de l'observation i et donc p_q la masse de la classe q.

On considèrera également le tableau Y défini par :

$$\forall i \in I_q \subset I, \forall j \in J : y_{ij} = x_{ij} - g_{qj}$$

Dans ce tableau, on a centré chaque observation i non sur le centre de gravité global (qui est l'origine par hypothèse (cf §2.)) mais sur le centre de gravité de la classe à laquelle elle appartient.

On peut alors

- soit faire l'A.C.P. du tableau X (Analyse globale)
- soit faire l'A.C.P. du tableau G (Analyse interclasse)
- soit faire l'A.C.P. du tableau Y (Analyse intraclasse)

Dans la première A.C.P. on rajoutera le tableau G en lignes supplémentaires au tableau X, ce qui revient à caractériser chaque classe I_q par son centre de gravité comme indiqué au paragraphe précédent. On rajoutera également en lignes et en colonnes supplémentaires le tableau Y (on désignera par iy (resp. jy) la $i^{ème}$ ligne (resp. $j^{ème}$ colonne) du tableau Y pour la différencier de la ligne i (resp. colonne j) du tableau X). Les projections sur l'axe factoriel α issu de l'A.C.P. de X des individus i, iy ($i \in I$) et des centres de gravité des classes q ($q \in Q$) aident souvent le praticien pour interpréter cet axe. Ces projections permettent en particulier de juger si sur cet axe α le nuage des individus i peut être schématisé par le nuage des centres des classes, ou si les dispersions à l'intérieur des classes interviennent dans l'explication du facteur α. Des calculs d'inertie interclasse (ou intraclasse) permettent de traduire numériquement l'interprétation précédente ; en particulier le rapport de corrélation de l'axe α qui est, rappelons le, le rapport entre la variance interclasse de l'axe et sa variance totale (égale à la valeur propre λ_α) est un indice très utile. Si ce rapport vaut 80 %, cela signifie que l'axe est expliqué à 80 % par la partition, la dispersion des individus à l'intérieur de chaque classe n'intervenant que pour 20 %.

Ce rapport de corrélation que nous noterons η^2_α peut s'écrire sous l'une des deux formes suivantes :

$$\eta^2_\alpha = \Sigma \{p_q F^2_\alpha (q) \mid q \in Q\} / \lambda_\alpha = 1 - (1/\lambda_\alpha) \Sigma \{p_i F^2_\alpha (iy) \mid i \in I\}$$

$F_\alpha (q)$ et $F_\alpha (iy)$ désignant respectivement les abscisses des projections de la classe q et de iy sur l'axe α.

L'A.C.P. du tableau G avec le tableau X en supplémentaire permet de placer les individus i de I sur des axes déterminés uniquement par les centres des classes. Si par exemple l'ensemble Q correspond au temps, les axes factoriels issus de l'A.C.P. de G correspondent à des axes d'évolution et il est intéressant de situer l'ensemble I sur ces axes. On peut comme précédemment calculer le rapport de corrélation associé au $\alpha^{ème}$ axe factoriel issu de l'A.C.P. de G (la valeur propre étant ici égale à la variance interclasse de l'axe). On peut aussi faire des calculs de corrélation entre les représentations de I sur les axes issus des A.C.P. de X et de G respectivement. Tous ces calculs permettent d'affiner l'interprétation de ces axes et de mieux comprendre la structure du tableau X compte tenu de la partition Q.

L'A.C.P. du tableau Y (avec X en lignes et en colonnes supplémentaires) permet de décomposer l'inertie intraclasse de I et donc de déterminer des axes dus uniquement à la dispersion à l'intérieur des classes. Il s'agit d'une analyse qui élimine l'effet de structure dû à la partition Q, ce qui peut être intéressant si cet effet est important. On peut ainsi découvrir des relations intéressantes entre les variables, relations qui étaient cachées par cet effet de structure.

Si on reprend le cas où Q correspond au temps, il peut exister un effet de taille très important dû à l'évolution (par exemple si X correspond à des données économiques en période de croissance). L'A.C.P. de X fournit alors des résultats très voisins de l'A.C.P. de G et si l'on veut avoir des résultats plus fins et moins triviaux, il faut faire l'A.C.P. de Y pour s'affranchir de cet effet taille.

Remarques :

1) Les analyses précédentes reviennent à considérer la décomposition suivante du tableau X :

$$X = Y + Z \qquad (4)$$

Z étant le tableau de terme général $z_{ij} = g_{qj}$ si $i \in I_q$, i.e. le tableau où chaque individu est caractérisé par le centre de gravité de sa classe.
On peut noter que l'A.C.P. de G (chaque classe q étant munie de sa masse p_q) est équivalente à l'A.C.P. de Z.

Si W désigne le sous espace de R^n engendré par les indicatrices associées à la partition Q de I, P le projecteur sur W (R^n étant muni de la métrique des poids), R = Id - P (où Id est la matrice unité d'ordre n) le projecteur sur le sous espace $W^{\perp}$ orthogonal à W, on a :

$$Z = P X \; ; \; Y = R X$$

Effectuer l'A.C.P. de Y ou de Z revient suivant la terminologie de Sabatier (1987) à faire une A.C.P. en tenant compte de la structure associée à X (ici la partition). On parle d'analyse des données structurées. L'A.C.P. de Y s'appelle souvent A.C.P. conditionnelle (à la structure sous jacente, i.e. à la partition).

En pratique, dans la plupart des cas, l'A.C.P. de X (avec G et Y en supplémentaire) suffit pour avoir une bonne idée des liaisons entre les variables, et des ressemblances entre les individus, compte tenu de la partition de I.

2) Il est inutile dans l'A.C.P. du tableau G (resp. Y) de placer Y (resp. G) en supplémentaire. En effet, d'après la remarque précédente, les lignes iy de Y (resp. classes q de Q) se projettent à l'origine des axes factoriels non triviaux (i.e. associés à une valeur propre non nulle) issus de cette A.C.P.

5. CAS D'UN TABLEAU TERNAIRE

On considère ici une suite de tableau $X_t = \{x_{ijt} \mid i \in I_t, j \in J_t\}$ indicée par t, t décrivant un ensemble fini T (Card T = r), et on fait l'hypothèse que :

- soit l'ensemble des individus I_t est indépendant de t, auquel cas on posera $I_t = I$
- soit l'ensemble des variables J_t est indépendant de t, auquel cas on posera $J_t = J$
- soit les ensembles I_t et J_t ne dépendent pas de t, auquel cas on a $I_t = I$ et $J_t = J$

On désignera par IT (resp. JT) l'union directe des I_t (resp. J_t), ce qui signifie en particulier que si $I_t = I$ (resp. $J_t = J$), IT (resp. JT) est la superposition de r fois l'ensemble I (resp. J).

5.1 CAS OU $J_t = J$

On désignera dans ce cas par X le tableau obtenu en empilant les tableaux X_t les uns en dessous des autres, et par x(it, j) le terme général de ce tableau :

$$\forall i \in I_t, \forall j \in J, \forall t \in T : x(it, j) = x_{ijt}$$

L'ensemble IT des lignes du tableau X étant muni de la partition définie par les I_t, on se trouve dans un cas particulier du §4. On pourra donc effectuer soit l'A.C.P. du tableau X (A.C.P. globale) en plaçant les centres de gravité des tableaux X_t en lignes supplémentaires, soit l'A.C.P. du tableau G de ces centres de gravité (avec X en supplémentaire), soit une A.C.P. intraclasse.

Si le nombre r de tableaux X_t est petit (r = 2 ou 3), il peut aussi être intéressant de faire l'A.C.P. (en général normée) de chacun de ces tableaux en mettant les autres en supplémentaire. Cette façon d'opérer revient à placer l'ensemble IT des individus sur les axes factoriels déterminés par le sous ensemble I_t de ces individus. Dans ce cas, on centre (et le cas échéant on réduit) les variables au niveau de chaque sous-tableau, alors que dans l'analyse globale, on centre (et le cas échéant on réduit) les variables sur l'ensemble IT de tous les individus.

5.2 CAS OU $I_t = I$

Ce cas est relativement plus simple (quoique moins fréquent si T représente le temps) que le précédent. On peut :

- soit faire l'A.C.P. du tableau X obtenu en accolant les tableaux X_t (le terme général x(i, jt) du tableau X valant x_{ijt})
- soit faire des analyses partielles des tableaux X_t en mettant le reste des variables JT - J_t en supplémentaire.

C'est cette dernière façon d'opérer que l'on utilise en général si les variables associées à des sous-ensembles J_t distincts sont de nature différente.

5.3 CAS OU $I_t = I$ ET $J_t = J$

Dans ce cas, on peut bien sur effectuer les analyses préconisées aux §§5.1 ET 5.2.

On peut aussi tenir compte du fait que tous les tableaux X_t ont mêmes dimensions, et définir un tableau moyen $U = \frac{1}{r} \sum \{X_t \mid t \in T\}$. On peut alors effectuer l'A.C.P. du tableau U en mettant chaque tableau X_t à la fois en lignes et en colonnes supplémentaires (cf. Figure 1). On obtient ainsi, pour chaque individu i, r + 1 représentations : r représentations de i associées à chacun des r tableaux X_t, et une représentation moyenne qui est le centre de gravité des r représentations précédentes. Si T représente le temps, on visualise ainsi sur chaque plan factoriel l'évolution de chaque individu. Outre I, on peut aussi représenter l'ensemble T à partir du tableau G des centres de gravité associés à chaque tableau X_t placés en lignes supplémentaires au tableau U.

En ce qui concerne l'ensemble J on obtient de la même façon r + 1 représentations de J ; mais la représentation d'une variable principale j (i.e. de la $j^{ème}$ colonne du tableau U) ne peut plus s'interpréter comme le centre de gravité des représentations de j associées aux $j^{èmes}$ colonnes des tableaux X_t ($t \in T$), sauf dans le cas de l'A.C.P. sur matrice variance, les variables étant représentées à l'aide de leur covariance (cf. formule (2)) et non de leur corrélation avec les composantes principales.

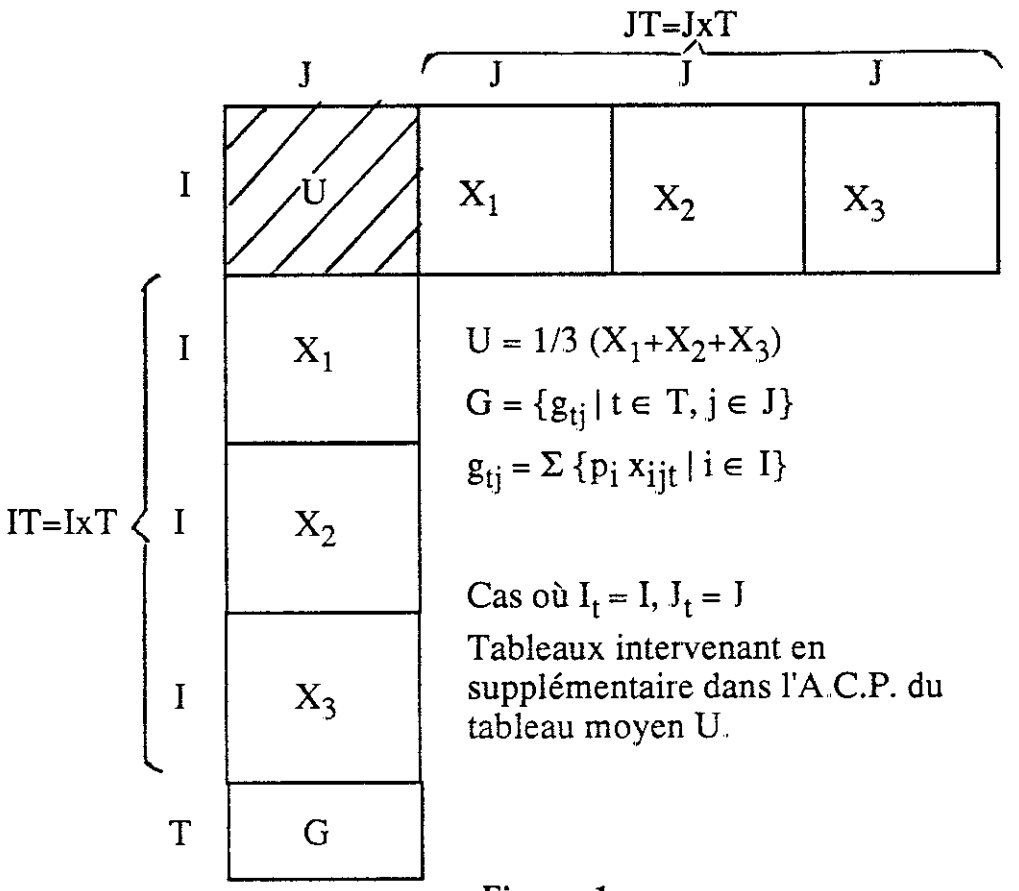

Figure 1

De façon précise, soit

- G'_α (j) la corrélation entre la $\alpha^{ème}$ composante principale issue du tableau moyen U et la $j^{ème}$ colonne de ce tableau.
- G'_α (jt) la corrélation entre cette composante principale et la $j^{ème}$ colonne du tableau X_t, qu'on notera j_t.

Ces corrélations qui permettent d'avoir les r + 1 représentations de j sur la composante principale précédente sont liées par la relation :

$$G'_\alpha (j) = \Sigma \{(\sigma_{jt} / \sigma_j) G'_\alpha (jt) \mid t \in T\} / r \qquad (5)$$

σ_j (resp. σ_{jt}) étant l'écart type de la $j^{ème}$ colonne du tableau U (resp. X_t).

On n'obtient pas, comme annoncé, une relation barycentrique. Si on avait voulu que j soit au centre de gravité des jt (affectés des masses 1/r), il aurait fallu, comme on l'a dit

ci-dessus, représenter les variables non par leurs corrélations avec les composantes principales, mais par leurs covariances.

Remarque : En général les variables associées à chaque tableau X_t sont centrées réduites (ce qui revient à considérer qu'on accole ces tableaux et non qu'on les empile) auquel cas dans la relation (5) σ_{jt} vaut 1. On effectue alors l'A.C.P. sur matrice variance du tableau U.

6. CAS DE L'ANALYSE DES CORRESPONDANCES

Dans ce cas, soit $k_{IJ} = \{k(i, j) \mid i \in I, j \in J\}$ le tableau étudié, qui peut être un tableau de contingence croisant deux variables qualitatives dont I et J constituent les ensembles de modalités ; ce peut être aussi un tableau homogène comme le tableau dont on a déjà parlé au §3. et donnant, exprimées en francs, les dépenses de 841 ménages pour un certain nombre de postes ; ce peut être encore un tableau intervenant en analyse des correspondances multiples, comme un tableau disjonctif complet, ou un tableau croisant deux ensembles de variables qualitatives.

On désignera par F_α (i) (resp. G_α (j)) l'abscisse de la projection de la ligne i (resp. colonne j) (i.e. de son profil) sur le $\alpha^{ème}$ axe factoriel issu de l'Analyse Factorielle des Correspondances (A.F.C.) du tableau k_{IJ}, et par λ_α la valeur propre associée.

Pour visualiser une ligne supplémentaire $\{k(i_s, j) \mid j \in J\}$ sur ce $\alpha^{ème}$ axe factoriel, on projettera le profil de i_s (i.e. la ligne i_s divisée par son total) sur cet axe, l'abscisse $F_\alpha (i_s)$ de la projection correspondante s'écrivant :

$$F_\alpha (i_s) = \Sigma \{k(i_s, j)\, G_\alpha (j) \mid j \in J\} / ((\sqrt{\lambda_\alpha})\, k(i_s, \ ))$$

$k(i_s, \ )$ désignant le total de la ligne i_s :

$$k(i_s, \ ) = \Sigma \{k(i_s, j) \mid j \in J\}$$

De façon symétrique, une colonne supplémentaire $\{k(i, j_s) \mid i \in I\}$ sera représentée sur le $\alpha^{ème}$ axe factoriel issu de l'A.F.C. de k_{IJ} par le point d'abscisse $G_\alpha (j_s)$ donné par :

$$G_\alpha (j_s) = \Sigma \{k(i, j_s)\, F_\alpha (i) \mid i \in I\} / ((\sqrt{\lambda_\alpha})\, k(\ , j_s))$$

$k(\ , j_s)$ désignant le total de la colonne j_s.
La plupart des considérations effectuées dans le cas de l'A.C.P. au niveau des individus supplémentaires restent valables ici aussi bien pour les lignes que pour les colonnes supplémentaires du fait de la symétrie des rôles joués par les lignes et les colonnes en A.F.C.

Dans le cas où I est muni d'une partition $(I = \cup \{I_q \mid q \in Q\})$ il suffit, si on veut faire une analyse interclasse de faire l'A.F.C. du tableau k_{QJ} dont la $q^{ème}$ ligne s'obtient par cumul des lignes i de I_q du tableau k_{IJ} $(\forall j \in J : k(q, j) = \Sigma \{k(i, j) \mid i \in I_q\})$. En effet, si l'on raisonne dans $\mathbf{R}_J$, on obtient la même métrique, que l'on raisonne dans l'A.F.C. de k_{QJ} ou dans celle de k_{IJ}. De plus, du fait des pondérations intervenant en A.F.C., le profil de chaque ligne q du tableau k_{QJ} n'est autre que le centre de gravité des profils des lignes i de I_q du tableau initial k_{IJ}, q ayant pour masse la somme des masses des i de I_q

Si l'on veut effectuer une analyse intraclasse, il suffit (cf. Benzecri (1983), Escofier (1983)) de faire l'A.F.C. du tableau k'_{IJ} dont le terme général est donné par :

$$\forall i \in I_q, \forall j \in J : k'(i, j) = k(i,j) - \frac{k(i, .)\, k(q, j)}{k(q, .)} + \frac{k(i, .)\, k(., j)}{k}$$

$k(i, .)$ et $k(., j)$ désignant respectivement les totaux de la ligne i et de la colonne j du tableau k_{IJ} tandis que k est le total général de ce tableau, $k(q, .)$ étant le total du bloc $I_q \times J$ de k_{IJ} ($k(q, .) = \sum \{k(i, .) \mid i \in I_q\}$).

Remarques :

1) Le tableau k'_{IJ} peut comporter des termes négatifs ; par contre il a mêmes marges que le tableau k_{IJ}, ce qui permet de définir sans problème les éléments (profils, métriques, poids) qui interviennent en A.F.C. et donc d'employer un programme usuel d'A.F.C. pour analyser k'_{IJ}

2) Considérons le tableau k''_{IJ} de terme général défini par :

$$\forall i \in I_q, \forall j \in J : k''(i, j) = k(i, .)\, k(q, j) / k(q, .)$$

tableau qui a même marges que k_{IJ}.

Compte tenu de ce que toutes les lignes de I_q du tableau k''_{IJ} sont proportionnelles, l'A.F.C de k''_{IJ} est d'après le principe d'équivalence distributionnelle équivalente à l'A.F.C. du tableau obtenu en additionnant pour chaque q de Q toutes les lignes de I_q, tableau qui n'est autre que le tableau k_{QJ}. On a :

$$k'(i, j) + k''(i, j) = k(i, j) + k(i, .)\, k(., j) / k$$

décomposition qui est analogue à (4), à ceci près qu'on a rajouté le terme $k(i, .)\, k(., j) / k$ de façon à éviter que le tableau k' ait des marges nulles, et de façon à pouvoir utiliser un programme usuel d'analyse des correspondances pour effectuer l'analyse intraclasse.

Si au lieu d'avoir une partition sur I, on a une partition sur J, tous les résultats précédents s'appliquent à condition d'intervertir les rôles de I et J. Si on a une partition sur I et J simultanément, on peut également faire une analyse interclasse et une analyse intraclasse simultanées (cette dernière analyse ayant été appelée analyse interne dans Cazes et alt (1988) qu'on pourra consulter pour plus de détails).

Dans le cas d'un tableau ternaire, les considérations du §5. se transposent aisément ici, à ceci près que les tableaux U et G de la figure 1 correspondraient ici à des cumuls (et non à des centres de gravité, les centres de gravité correspondant aux profils des lignes ou des colonnes de ces tableaux de cumuls).

La technique des éléments supplémentaires joue également un grand rôle en analyse des correspondances multiples. On pourra à ce sujet consulter Cazes (1982 b)) ainsi que Cazes (1982 a)) où l'étude des éléments supplémentaires en A.F.C. est exposée en détail.

7. APPLICATIONS DES ELEMENTS SUPPLEMENTAIRES A LA REGRESSION APRES ANALYSE FACTORIELLE

On suppose ici que l'on veut expliquer une variable quantitative y en fonction de p variables quantitatives $x_1, x_2, ..., x_p$, toutes ces variables ayant été mesurées sur un échantillon de taille n. Nous désignerons par X le tableau n x p associé aux variables explicatives et nous utiliserons les notations employées au §2. Nous supposerons de plus qu'on a divisé y en r tranches, ce qui définit une partition de l'ensemble I des individus :

$$I = \cup \{I_q \mid q \in Q\}$$

I_q désignant l'ensemble des individus rentrant dans la $q^{ème}$ tranche de y, et Q l'ensemble des r tranches de y. On peut, pour s'affranchir des limitations de la régression usuelle utiliser plusieurs procédures. Nous en proposons trois ci-dessous, la division en classes de y permettant d'enrichir la première et étant à la base des deux autres.

7.1 PREMIERE METHODE : A.C.P. DU TABLEAU X

On effectue l'A.C.P. du tableau X (en général l'A.C.P. sur matrice de corrélation) en plaçant y en élément supplémentaire, et en projetant également sur les axes factoriels les centres de gravité de chaque classe I_q, ce qui permet en particulier de voir si un facteur est lié (linéairement à l'intérieur du cercle de corrélation, et non linéairement sur les plans factoriels du fait du découpage en classes de y) à y. On effectue ensuite la régression usuelle de y sur les premiers facteurs issus de cette analyse, en ne conservant le cas échéant que ceux qui sont significativement liés à y. L'avantage de cette façon de procéder réside dans les trois points suivants :

a) avant de faire la régression, on visualise la structure des variables explicatives, ainsi que leurs liaisons avec y, ce qui nous semble fort important en pratique.

b) on élimine le bruit dû aux fluctuations d'échantillonnage en ne conservant que les premiers facteurs associés à un pourcentage d'inertie suffisant.

c) la régression se fait sur les facteurs qui sont non corrélés, ce qui permet de rajouter sans difficultés un ou plusieurs facteurs dans la régression si on le juge nécessaire.

On peut après avoir obtenu la formule de régression sur les facteurs, revenir aux variables initiales.
De façon précise, soit

$$y^*(i) = \Sigma \{c_\alpha F_\alpha (i) \mid \alpha \in A\} \qquad (6)$$

la formule de régression sur les facteurs, A désignant l'ensemble des facteurs conservés, F_α (i) l'abscisse de la projection de l'individu i sur le $\alpha^{ème}$ axe factoriel issu de l'A.C.P. de X et $y^*(i)$ la valeur approchée de y pour i. Le coefficient de régression c_α est donné par :

$$c_\alpha = \Sigma \{p_i F_\alpha (i) y(i) \mid i \in I\} / \lambda_\alpha \qquad (7)$$

p_i étant le poids de l'individu i, y(i) la valeur de y pour i et λ_α la valeur propre associée à l'axe α.

Compte tenu de ce que

$$F_\alpha (i) = \Sigma \{u_{j\alpha} x_{ij} \mid j \in J\} \qquad (8)$$

$u_{j\alpha}$ désignant la $j^{ème}$ composante du $\alpha^{ème}$ vecteur axial factoriel, la formule (6) peut se mettre sous la forme :

$$y^*(i) = \Sigma \{b_j x_{ij} \mid j \in J\} \qquad (9)$$

avec :

$$b_j = \Sigma \{c_\alpha u_{j\alpha} \mid \alpha \in A\} \qquad (10)$$

On peut noter la contribution additive de chacun des facteurs conservés dans la formule (10), formule qui redonne les coefficients de régression usuels si on garde tous les facteurs. Il peut être intéressant, en utilisant cette formule, d'effectuer une régression pas à pas sur les facteurs, le pas à pas se faisant soit par valeurs propres décroissantes, soit par corrélations (en valeur absolue) décroissantes de ces facteurs avec y.

Remarques :

1) Pour prévoir la valeur de y pour un individu s à partir de la connaissance des valeurs des x_j pour s, on peut, au lieu d'utiliser la régression sur facteurs, employer la régression par voisinage (ou par boule) en considérant dans l'espace des premiers axes factoriels où s a été projeté en élément supplémentaire, les voisins de s (parmi les n individus initiaux) et en faisant la moyenne des valeurs de y pour ces voisins

2) Si les données s'y prêtent, on peut, au lieu de faire l'A.C.P. de X, effectuer l'A.F.C. préalablement à la régression. Des exemples d'application de cette méthodologie sont donnés dans Abdel Shahid (1982), Cazes (1978) et Roux (1979).

7.2 DEUXIEME METHODE : ANALYSE INTERCLASSE

Dans ce cas, on effectue l'A.C.P. du tableau G des centres de gravité des classes associées à chaque tranche de y, en plaçant chacune des lignes du tableau X en élément supplémentaire. Les facteurs F_α obtenus par projection de l'ensemble I des individus sur les axes factoriels issus de l'A.C.P. de G sont alors par construction liés à y, et on peut effectuer la régression de y sur ces facteurs qui en général ne sont plus non corrélés (sur I). Les formules (6), (8) à (10) restent valables, $u_{j\alpha}$ désignant ici la $j^{ème}$ composante du $\alpha^{ème}$ vecteur axial factoriel issu de l'A.C.P. de G. Par contre, du fait de la corrélation des facteurs, la formule (7) n'est plus valable, et si l'on décide de rajouter un facteur dans la régression, tous les coefficients c_α changent, ce qui complique les calculs si on décide d'effectuer une méthode de pas à pas sur les facteurs, basée sur (10), la contribution des facteurs n'étant plus additive, quand leur nombre n'est pas fixé.

L'avantage de cette méthode est de "forcer" la liaison entre y et les x_j, si cette liaison est faible ; par contre la méthode est biaisée dans la mesure où elle a tendance à surestimer cette liaison. Il est donc indispensable de contrôler les résultats obtenus à l'aide d'un échantillon test.

On peut bien sûr, comme dans la première méthode, utiliser la régression par boule ou effectuer, si les données s'y prêtent l'A.F.C. (non plus du tableau G, mais du tableau obtenu à partir de X par cumul des lignes i de chaque classe I_q) préalablement à la régression. On trouve un exemple intéressant de cette façon d'opérer dans Abdel Shahid (1982).

7.3 TROISIEME METHODE : ANALYSE DES CORRESPONDANCES MULTIPLES

Dans ce cas, on suppose qu'outre y, toutes les variables explicatives x_j ($1 \leq j \leq p$) sont divisées en classes, et l'on caractérise ces variables explicatives non plus par le tableau X, mais par le tableau disjonctif complet k_{IJ_X} croisant l'ensemble I des individus avec l'ensemble J_X des tranches de ces p variables. Le tableau G précédemment considéré est ici remplacé par le tableau k_{QJ_X} croisant Q avec J_X et obtenu en cumulant dans k_{IJ_X} les lignes i appartenant à une même tranche q de la variable à expliquer y. Toutes les considérations effectuées lors des deux méthodes précédentes restent valables à condition de remplacer X par k_{IJ_X}, G par k_{QJ_X} et l'A.C.P. par l'A.F.C.

L'avantage de cette façon d'opérer est, du fait du découpage en classes de toutes les variables, de pouvoir visualiser les liaisons non linéaires entre y et les x_j. Elle permet également de traiter immédiatement le cas où les variables explicatives sont qualitatives ou un mélange de variables qualitatives et quantitatives.

On trouve un exemple d'application de cette méthodologie dans Cazes (1976) où l'on effectue l'A.F.C. de k_{QJ_X} avec k_{IJ_X} en supplémentaire, avant d'effectuer la régression usuelle et la régression par boule dans une étude géologique où l'on devait expliquer le pourcentage de matière organique dans une roche en fonction d'un grand nombre de variables géologiques qualitatives et quantitatives.

8. CAS DE LA DISCRIMINATION

Dans ce cas où la variable à expliquer y est qualitative, les méthodes du § précédent se transposent aisément. La deuxième de ces méthodes correspond du reste exactement à l'Analyse Factorielle Discriminante si on munit l'espace $\mathbf{R}^p$ non de la métrique usuelle, comme c'était le cas au §7.2, mais de la métrique de Mahalanobis.

Dans la première méthode, on effectue la discrimination sur les facteurs issus de l'A.C.P. de X, et on peut, comme au §7.1 repasser d'une combinaison linéaire discriminante $\sum_\alpha c_\alpha F_\alpha$ sur les facteurs à une combinaison linéaire $\sum_j b_j x_j$ sur les variables initiales à l'aide de la formule (10). Par contre (sauf dans le cas où y ne comporte que 2 modalités), le fait de rajouter un facteur, change les coefficients c_α des facteurs déjà présents dans la combinaison linéaire discriminante, contrairement à ce qui se passait en régression.

Dans le cas où toutes les variables explicatives sont qualitatives (ou rendues qualitatives par découpage en tranches) la discrimination sur les facteurs issus de l'A.F.C. du tableau disjonctif complet k_{IJ_X} associé correspond à la méthode DISQUAL (Saporta (1976)). Il nous semble plus intéressant dans ce dernier cas, de faire l'A.F.C. du tableau k_{QJ_X} croisant l'ensemble Q des modalités de y avec l'ensemble J_X de toutes les modalités explicatives puis d'effectuer l'analyse discriminante sur les facteurs obtenus en plaçant chaque ligne i du tableau k_{IJ_X} en ligne supplémentaire.

Comme dans le cas de la régression, cette façon de procéder "force" la discrimination, en particulier si la liaison de y avec les variables explicatives est faible. Il faudra bien sur valider obligatoirement les résultats de la discrimination obtenue en utilisant un échantillon test. On peut trouver des exemples de cette méthodologie (qu'on appelle usuellement discrimination barycentrique, et qui est exposée dans Bastin et alt. (1980) dans le cas particulier où y ne comporte que deux classes) dans Bergougnan et alt. (1982) ainsi que dans Nakache et alt. (1977).

BIBLIOGRAPHIE

[1] ABDEL SHAHID, A. (1982) : "Application de la régression d'après un tableau de correspondance : L'estimation des paléoclimats d'après l'écologie des foraminifères" ; C.A.D., Vol. VII n° 1, pp. 93-111

[2] BASTIN, Ch., BENZECRI, J.P., BOURGARIT, Ch., CAZES, P. (1980) : "Dicrimination entre deux classes par affectation barycentrique", in Pratique de l'Analyse des Données, Tome 2, pp. 102, 104, Dunod.

[3] BENZECRI, J.P. (1983) : "Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondance", C.A.D., Vol. VIII n° 3, pp. 341-354

[4] BERGOUGNAN, D., COURAUD, C. (1982) : "Pratique de la discrimination barycentrique", C.A.D., Vol. VII n° 3, pp. 351-358

[5] CAZES, P. (1976) : "Régression par boule et par l'analyse des correspondances", R.S.A., Vol. XXIV n° 4, pp. 5-22

[6] CAZES, P. (1978) : "Estimation de la statistique de multiplication d'un photomultiplicateur à dynodes", C.A.D., Vol. III n° 4, pp. 393-417

[7] CAZES, P. (1982) : "Note sur les éléments supplémentaires en analyse des correspondances" :
a) Pratique et Utilisation, C.A.D., Vol. VII n° 1, pp. 9-23
b) Tableau Multiples, C.A.D., Vol. VII n° 2, pp. 133-154

[8] CAZES, P., CHESSEL, D., DOLEDEC, S. (1988) : "L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie", R.S.A., Vol. XXXVI n° 1, pp. 39-54

[9] ESCOFIER, B., DROUET, D. (1983) : "Analyse des différences entre plusieurs tableaux de fréquences", C.A.D., Vol. VIII n° 4, pp. 491-499

[10] LEBART, L., MORINEAU, A., TABARD, N. (1977) : "Techniques de la description statistique", Dunod

[11] LEBART, L., MORINEAU, A., FENELON, J.P. (1979) : "Traitement des données statistiques", Dunod

[12] NAKACHE, J.P., LORENTE, P., BENZECRI, J.P., CHASTANG, J.F. (1977) : "Aspects pronostiques et thérapeutiques de l'infarctus myocardique aigu compliqué d'une défaillance sévère de la pompe cardiaque. Application des méthodes de discrimination", C.A.D., Vol. II n° 4, pp. 415-434

[13] ROUX, M. (1979) : "Estimation des paléoclimats d'après l'écologie des foraminifères", C.A.D., Vol. IV n° 1, pp. 61-79

[14] SABATIER, R. (1987) : "Méthodes factorielles en analyse des données. Approximation et prise en compte de variables concomitantes, thèse de doctorat ès sciences", Un. de Montpellier

[15] SAPORTA, G. (1976) : "Discriminant analysis when all the variables are nominal : a stepwise method", note de travail COREF n° 8