LES CAHIERS DE L'ANALYSE DES DONNÉES

P. CAZES

Note sur les éléments supplémentaires en analyse des correspondances I. Pratique et utilisation

Les cahiers de l'analyse des données, tome 7, n° 1 (1982), p. 9-23

<http://www.numdam.org/item?id=CAD_1982__7_1_9_0>

Les Cahiers de l'Analyse des Données
Vol. VII - 1982 - n° 1 - p. 9-23

NOTE SUR LES ÉLÉMENTS SUPPLÉMENTAIRES EN ANALYSE DES CORRESPONDANCES

I. PRATIQUE ET UTILISATION

[EL. SUPP. I]

(à suivre)

par P. Cazes (1)

0 *Introduction*

Les éléments supplémentaires présentent un grand intérêt en analyse des données, soit pour ne pas perdre d'information sur un cas dont on ne désire pas qu'il participe activement aux analyses (cas nouveau, observation aberrante ou douteuse, élément de nature différente de ceux analysés, etc.), soit pour interpréter plus facilement ces analyses (à l'aide de variables illustratives par exemple, comme l'appartenance à une C.S.P. dans une analyse portant sur des questions d'opinion), soit pour effectuer une régression, etc. . Rappelons que ces éléments sont représentés sur les axes issus d'une analyse factorielle, ou affectés aux classes déterminées à partir d'une classification automatique, pour s'en tenir à ces deux techniques, sans qu'ils aient contribué à la formation de ces axes ou de ces classes.

Nous étudierons la pratique des éléments supplémentaires en analyse des correspondances. L'article est divisé en deux parties : dans la première partie, publiée ici, nous faisons des rappels sur l'utilisation (§ 1), le principe et les propriétés (§ 2) des éléments supplémentaires en analyse factorielle, puis nous examinons le cas fort important en pratique de la représentation d'un groupe d'individus (ou de variables) en tant que ligne ou colonne supplémentaire (§ 3).

Dans la seconde partie, qui sera publiée ultérieurement , nous montrons le rôle et l'intérêt des éléments supplémentaires dans l'analyse des tableaux ternaires (§ 4) puis dans l'analyse des correspondances multiples (§ 5). Nous terminons ces deux articles en montrant comment les éléments supplémentaires permettent de comparer les facteurs issus de deux analyses différentes.

1 *Utilisation des éléments supplémentaires en analyse des correspondances*

Nous distinguerons plusieurs cas d'utilisation des éléments supplémentaires en analyse des correspondances, dont certains ont été brièvement rappelés dans l'introduction.

1.1 *Utilisation des éléments supplémentaires pour représenter* :

- soit une observation relevée dans des conditions douteuses (ou différentes des autres observations) ou encore une variable sur laquelle

(1) *Professeur de statistique à l'Université de Paris Dauphine.*

la précision est moindre que sur les autres variables mesurées

- soit un élément aberrant, ou ayant perturbé une analyse préliminaire

- soit un cas nouveau

- soit des éléments de nature différente de ceux analysés; par exemple dans l'analyse d'un tableau échantillon × (élément majeur + élément trace) (cf. P. Cazes, thèse 3° cycle, Paris 6, 1970) on mettra les éléments trace en éléments supplémentaires, car ils ne sont pas de même nature que les éléments majeurs (bien que mesurés en parties par millions, les éléments trace correspondent plutôt à des variables qualitatives, soit à 2 modalités, présence-absence, soit à plusieurs modalités correspondant à des abondances différentes, tandis que les majeurs, mesurés en pourcentages sont des variables continues). De même dans un questionnaire d'opinions, on analysera par exemple le signalétique, en mettant les questions d'opinion en éléments supplémentaires.

Remarque : L'usage des éléments supplémentaires en analyse des correspondances a débuté il y a une dizaine d'années avec les cas que l'on vient d'examiner dans ce § 1.1.

1.2 *Utilisation des éléments supplémentaires pour représenter un groupe d'individus ou de variables*

Cette utilisation des éléments supplémentaires est d'une grande importance pratique ; le groupe d'individus que l'on veut représenter sur les axes factoriels par un point caractéristique de ce groupe, 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 différents C.S.P., aux tranches de revenus etc., exemple donné dans Lebart et coll. Techniques de la description statistique, Dunod (1977)).

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 (cf. M. Jambu & M. Sadaka ; 1974 ; M. Jambu, J. Robert, M. Roux, C.A.D. Vol I pp 61-30 , 1976 ; M. Jambu, C.A.D. Vol I pp 77-86 , 1976; cf. aussi [*AID. INT. CLASS.*] & [*AID. CAH.FACOR.*], *Cahiers*, Vol V n°1, 1980).

Signalons également que dans l'analyse d'un tableau issu d'une analyse de correspondance multiple (tableau disjonctif complet, tableau de Burt, ou sous tableau d'un tableau de Burt) il peut être intéressant pour l'interprétation de représenter les groupes d'individus qui fournissent les mêmes réponses à 2 ou 3 questions différentes. Pour un exemple d'une telle utilisation des éléments supplémentaires, on pourra consulter [*ORIENTATION* 3°], *Cahiers*, Vol VI n° 1, pp 19-38 (1981).

La représentation de groupes d'individus qui joue un grand

rôle dans l'étude des tableaux ternaires (cf. §§ 1.3 et 4) sera étudiée d'un point de vue théorique au § 3, où l'on donnera trois méthodes de représentation possibles.

1.3 *Utilisation des éléments supplémentaires dans le cas de tableaux ternaires*

Soit k_{IJT} un tableau ternaire, k_{IJ} , k_{IT} et k_{TJ} les tableaux marges associés. k_{IJT} peut être considéré de trois manières différentes comme un tableau binaire k_{AB} où A est le produit de deux des ensembles intervenant, et B le dernier ensemble (par exemple A = I×T,B=J) et on pourra pour étudier k_{IJT} analyser un ou plusieurs des tableaux précédents, en mettant en éléments supplémentaires les tableaux marge.

On pourra aussi analyser un tableau marge (ou deux tableaux marges accolés) en mettant les autres tableaux marges,et les deux tableaux de type k_{AB} possibles en élément supplémentaire (cf. figure 3, cas 3 , où le tableau marge analysé est le tableau k_{IJ}).

Citons brièvement quelques exemples d'analyses de tableaux ternaires :

- dans sa thèse (Paris 6, 1977) M. Jabbour étudie le tableau k_{IJT} où I est l'ensemble des départements français, J l'ensemble de 4 attitudes de vote (oui-non-blanc-abstention) et T l'ensemble des 6 référendums qui se sont déroulés en France entre 1958 et 1972. Pour faire cette étude, M. Jabbour fait l'analyse des correspondances du tableau k_{AB} où A = I, B = J × T, et il indique, comment à l'aide des éléments supplémentaires, on peut représenter chaque couple (i , t) associé à un département pour un référendum donné. La technique employée par Jabbour pour représenter ces couples est développée au § 4 (cf. en particulier figure 3, cas n° 2) où l'on étudie en détail et de façon théorique l'analyse de ces tableaux ternaires.

- Jabbour étudie également une suite de tableaux $\{k_{IJ_t} | t \in T\}$ où T désigne l'ensemble des 3 élections du président de la république française en 1965, 1969, 1974, I l'ensemble des circonscriptions électorales, et J_t l'ensemble des attitudes de vote pour l'élection t (1-er et 2-ème tours)). Pour faire cette étude, Jabbour analyse chaque tableau k_{IJ_t} avant de faire une analyse générale en effectuant l'analyse des correspondances du tableau $k_{I \times JT}$ superposition des k_{IJ_t} ($J = \cup\{J_t | t \in T\}$). L'étude d'une telle suite de tableaux est également effectuée au § 4 (à ceci près qu'on considère des tableaux de la forme k_{I_tJ} et non k_{IJ_t}).

- dans les études les plus fréquemment rencontrées de tableaux ternaires, I est un ensemble d'industries (ou de produits), J un ensemble de pays (ou de régions), T un ensemble d'époques, k(i,j, t) désignant les échanges pour le produit i , à l'instant t en provenance (ou à destination) du pays j . On pourra consulter par exemple les articles suivants parus dans les *Cahiers* : [*BRESIL II*],C.A.D. Vol III n° 3, pp 307-342 (1978) ; [*MULTINAT.*], C.A.D. Vol V n° 1,pp 17-43 (1980) ; [*EXPORT INDE*], C.A.D, Vol V n° 4 , pp 407-442 (1980) ; [*AGRI. SYRIE*], ce cahier, pp 67-91 , et pour une généralisation au cas quaternaire [*OPEP O.C.D.E.*] à paraître.

1.4 *Utilisation des éléments supplémentaires dans la régression après analyse factorielle*

Si l'on veut expliquer une variable y en fonction d'une série de variables explicatives $\{x_j | j \in J\}$, toutes ces variables étant mesurées sur un n-échantillon I, on peut, pour essayer de s'affranchir des limitations de la régression usuelle utiliser les procédures que l'on va décrire ci-dessous, procédures où les éléments supplémentaires jouent un rôle important.

Soit $x_{IJ} = \{x_{ij} | i \in I, j \in J\}$ le tableau des valeurs des variables explicatives $\{x_j | j \in J\}$ mesurées sur I, et $y_I = \{y_i | i \in I\}$ l'ensemble des valeurs de y sur I.

Nous supposerons dans un premier temps, d'une part que x_{IJ} est un tableau homogène de nombres positifs sur lequel on peut effectuer l'analyse des correspondances et que d'autre part les y_i sont positifs ou nuls (les $\{x_{Ij} | j \in J\}$ peuvent correspondre par exemple à une série de courbes, qui peuvent être des lois de probabilité, des courbes granulométriques, etc., x_{ij} désignant l'ordonnée de la courbe j au point d'abscisse i) et nous allons décrire 3 procédures dont la dernière s'affranchit des hypothèses restrictives énoncées ci-dessus.

1.4.1 *Première procédure : analyse des correspondances du tableau* x_{IJ}

Dans cette procédure, on effectue l'analyse des correspondances du tableau x_{IJ} en y adjoignant y_I comme élément supplémentaire. Cette façon de procéder revient à faire la régression (visualisée) de y sur les facteurs F_α issus du tableau x_{IJ}, facteurs qui sont non corrélés. C'est donc une régression sur variables orthogonales, et on s'affranchit du problème de colinéarité des variables explicatives initiales en ne gardant que les premiers facteurs F_α associés à un pourcentage d'inertie suffisant. On peut ensuite le cas échéant revenir aux variables initiales.

Pour des exemples de cette façon de procéder, on pourra consulter [*PHOTOMULTIPLICATEURS*], *Cahiers*, Vol III n° 4, pp 393-417 (1978) où y_I est une loi de probabilité combinaison linéaire (avec des coefficients inconnus à estimer) de lois de probabilité x_{Ij}. On notera que dans cet exemple, les coefficients de régression à estimer doivent être positifs pour avoir un sens.

Un autre exemple (cf. [*PALEOCLIMATS*], *Cahiers*, Vol IV n°1, pp 61-79 (1979), cf. aussi la thèse de docteur-ingénieur de A. Abdel Shahid, Paris 6, 1981, ainsi que [*REGR. CLIMATS*], ce cahier pp 93-111) est relatif à l'estimation des paléoclimats d'après l'écologie des foraminifères : y est la température, les variables $\{x_j | j \in J\}$ correspondent aux abondances de 30 foraminifères, I est un ensemble de relevés récents ; y et les x_j sont connus sur l'échantillon I des relevés récents et le problème est de prévoir y sur un ensemble I' de relevés anciens pour

lesquels seuls les x_j sont connus.

L'analyse des correspondances du tableau x_{IJ} relatif aux observations récentes avec y_I en supplémentaire permet d'avoir une formule de régression de y sur les facteurs F_α issus de x_{IJ} ; l'application de cette formule de régression aux facteurs obtenus en adjoignant le tableau $x_{I'J}$ relatif aux observations anciennes en supplémentaire de x_{IJ} permet de prévoir y sur I'. En fait la position de l'ensemble I et de l'ensemble supplémentaire I' sur les axes factoriels a amené à calculer la formule de régression sur un sous ensemble I_1 de I, correspondant aux relevés de I proches de I', les relevés de I' étant très concentrés dans le plan 1-2, alors que les relevés de I sont étalés dans tout le plan ; cette façon de procéder permet d'améliorer la qualité de la régression sur l'ensemble I_1, et donc *a priori* la qualité de la prévision sur I'.

1.4.2 *Deuxième procédure : découpage en classes de la variable à expliquer*

Dans cette procédure, valable même si la variable à expliquer y peut prendre des valeurs négatives, on découpe l'intervalle de variation de cette variable en tranches, et on considère si C désigne l'ensemble des classes ainsi défini pour y le tableau x_{CJ} défini par:

$$\forall\, c \in C : x(c,j) = \Sigma\{x_{ij} \mid i \in c\}$$

en bref, la ligne c du tableau x_{CJ} est la somme des lignes du tableau x_{IJ} associées aux individus tombant dans la classe c de la variable y.

On fait alors l'analyse factorielle du tableau x_{CJ} en lui adjoignant en élément supplémentaire le tableau x_{IJ}, puis on effectue la régression de y sur les facteurs F_α^I ainsi obtenus. On peut noter qu'ici les facteurs F_α^I ne sont pas non corrélés, comme c'était le cas au § 1.4.1, puisque l'ensemble I est projeté en supplémentaire sur les axes factoriels issus du tableau x_{CJ}, axes factoriels qui sont déterminés par les centres de gravité c ($c \in C$) des éléments i appartenant à la classe c. Pour un exemple d'utilisation d'une telle procédure, on pourra consulter la thèse de A. Abdel Shahid, déjà citée sur l'étude des paléoclimats, ou [*REGR. CLIMATS*].

Remarques :

1) Si les variables $\{x_j \mid j \in J\}$ et y sont des variables quantitatives non homogènes, prenant des valeurs aussi bien positives que négatives, on peut encore appliquer la procédure du § 1.4.1, ou celle exposée ici, à condition de remplacer l'analyse des correspondances par l'analyse en composantes principales normée (A.C.P. normée). On effectuera donc dans le cas de la première procédure l'A.C.P. normée du tableau x_{IJ} avec y_I en élément supplémentaire, tandis que dans le cas de la seconde procédure on effectuera l'A.C.P. normée du tableau des

centres de gravité associés à chaque classe c de y, c étant affecté d'une masse égale à la somme des masses des individus i appartenant à cette classe.

2) Si on veut étudier les liaisons non linéaires entre y et les x_j, il vaut mieux appliquer la procédure du § suivant, qui correspond aux procédures déjà décrites, mais dans le cadre des correspondances multiples, ces procédures étant valables quel que soit le type de variables intervenant ; variables de signes quelconques avec des unités différentes, qualitatives, quantitatives ; mélange des variables précédentes.

3) Quand on effectue l'analyse du tableau x_{IJ} (cf. § 1.4.1), on a toujours intérêt à diviser y en classes, et à projeter le tableau x_{CJ} associé en élément supplémentaire, ce qui permet de visualiser les liaisons non linéaires entre y et les x_j.

1.4.3 *Troisième procédure : découpage en classes de toutes les variables*

On divise toutes les variables en classes, et on désignera ici par J l'ensemble des modalités de toutes les variables explicatives, C désignant toujours l'ensemble des classes de la variable à expliquer y.

Soit k_{IJ} le tableau disjonctif complet associé aux variables explicatives et k_{CJ} le tableau déduit de k_{IJ}, comme x_{CJ} est déduit de x_{IJ}. On pourra alors appliquer une des deux procédures précédentes où le tableau x_{IJ} est remplacé par k_{IJ}, à savoir :

- analyse de k_{IJ} (avec k_{CJ} en supplémentaire pour visualiser les liaisons de y avec les variables explicatives)
- analyse de k_{CJ} avec k_{IJ} en supplémentaire.

Dans chacun des deux cas, on obtient des facteurs F_α sur I, et on pourra expliquer la variable y (avant découpage en classes) en fonction des F_α.

Du fait du découpage en classes, cette procédure reste valable comme on l'a déjà signalé (cf. remarque 2 du § 1.4.2) même si on a un mélange de variables explicatives qualitatives et quantitatives. Elle a été appliquée en particulier dans une étude géologique où il s'agissait d'expliquer la teneur en kérogène d'un certain nombre de roches en fonction d'un grand nombre de variables explicatives aussi hétérogènes que la couleur de la roche, le diamètre des grains, la teneur en calcaire, l'épaisseur de la roche, la texture en surface polie, etc. (cf. Cazes, R.S.A., Vol 24, n° 4, pp 5-22, (1976)).

1.5 *Applications diverses des éléments supplémentaires*

Nous nous contenterons de donner trois applications. La première est relative à un tableau ternaire k_{IJT} où l'on désire voir le rôle des interactions d'ordre supérieur à 2. Pour cela on construit le tableau h_{IJT} qui a mêmes marges d'ordre inférieur ou égal à 2 que k_{IJT}, mais dont les interactions d'ordre 3 sont nulles. On analyse alors le

tableau k_{IJT} (avec h_{IJT} en supplémentaire) ou h_{IJT} (avec k_{IJT} en supplémentaire) suivant les techniques des tableaux ternaires (cf §§ 1.3 et 4). Pour un exemple d'application relatif à des données linguistiques on pourra consulter la thèse de 3-ème cycle de A. Bener (Paris 6, 1981). D'un point de vue théorique , on pourra consulter [*INTER.CORR. MULT.*], ce cahier, pp 25 - 32 .

Le deuxième exemple est relatif à un tableau de Burt B_{JJ} croisant l'ensemble $J = \cup \{J_q | q \in Q\}$ des modalités des questions d'un questionnaire avec lui-même. On peut avoir intérêt après avoir analysé B_{JJ}, à effectuer l'analyse des correspondances d'un sous tableau (une bande) B_{JJ_q} de B_{JJ}, le reste du tableau étant en élément supplémentaire. (Eventuellement si les dimensions du tableau $J \times J$ ne permettent pas de l'analyser sans une dépense excessive, on se bornera à des analyses partielles de sous tableaux en bandes). L'analyse d'une bande semble en particulier s'imposer si une question q sort avec un fort effet Guttman sur le plan de deux axes factoriels issus de B_{JJ}. On a en effet l'habitude dans ce cas d'interpréter ce plan en référence à cette variable, et il semble donc logique d'analyser la bande B_{JJ_q} (avec le reste du tableau B_{JJ} en supplémentaire) ce qui permet de représenter dans R_J le nuage $N(J)$ des profils des colonnes de B_{JJ} sur des axes déterminés uniquement par le sous nuage $N(J_q)$ associé à la question q. Notons que dans les deux analyses précédentes (analyse du tableau entier $J \times J$, ou de la bande $J \times J_q$), du fait de la structure en blocs du tableau B_{JJ}, l'espace R_J est muni de la même métrique, et l'on peut dans la seconde analyse calculer facilement à partir du listage, l'inertie du nuage $N(J)$ sur les axes factoriels issus de B_{JJq} (cf. § 5.2 *in fine*, ainsi que [*BANDES BURT*], ce cahier, pp 33 - 43).

On peut trouver un exemple d'application de cette façon de procéder dans une étude socio-démographique de l'agglomération de Dijon (cf M. Essadaoui, thèse 3-ème cycle, Paris 6, (1981), et [*DIJON*], à paraître dans les *Cahiers*). Dans cette étude, l'on possède des caractéristiques liées à l'habitat et à la population, et dans l'analyse factorielle sur l'habitat, certaines variables comme la date de construction du logement, ou la possession, ou non, du téléphone ressortaient fortement sur certains axes ou certains plans factoriels. La projection du nuage entier sur les axes (ou sur l'axe dans le cas d'une variable à 2 modalités, comme le téléphone) déterminés par chacune des variables qui ressortaient dans l'analyse globale, a permis de confirmer et d'affiner les interprétations obtenues. Un dépouillement d'enquête fondé exclusivement sur l'analyse de tableaux en bandes, se trouve dans la thèse de Mme Ivarson (W. Kubów) ; cf. [*ENQUETE ECOLE*] ce cahier, pp 45-65.

La dernière application est relative aux données manquantes. Supposons que dans un tableau k_{IJ}, il existe une donnée manquante relative à la case (i_s , j_s). On peut pour reconstituer $k(i_s , j_s)$ analyser le sous tableau $k_{I'J'}$ (avec $I' = I - \{i_s\}$, $J' = J - \{j_s\}$) en lui adjoignant la ligne i_s et la colonne j_s en éléments supplémentaires, puis appliquer la formule de reconstitution (cf. Prat. de l'A. des données, Vol 2, § III n° 7). On peut aussi analyser $k_{I'J}$ (resp. $k_{IJ'}$) avec i_s

(resp. j_s) en supplémentaire, puis appliquer la formule de reconstitution (cf. Burtschy, Nora, Vercken, actes du colloque I.R.I.A., analyses des données et informatique, pp 589-600 (1977)) .

2 *Rappels sur la technique des éléments supplémentaires en analyse des correspondances*

2.1 *Principe des éléments supplémentaires*

Considérons la figure 0 , où sont représentés trois tableaux :

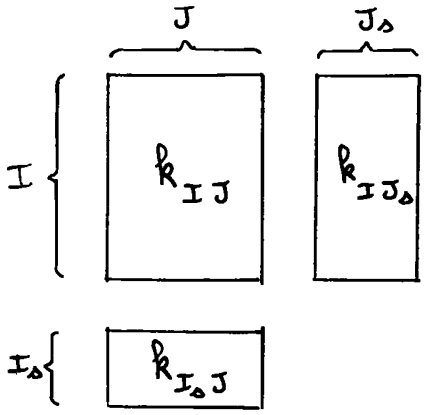

Figure 0

k_{IJ} , k_{I_sJ} , k_{IJ_s} .

On dit que les tableaux k_{I_sJ} et k_{IJ_s} sont adjoints en tableaux supplémentaires à k_{IJ} , si l'on effectue l'analyse des correspondances du tableau k_{IJ} , et si l'on projette sur les axes factoriels ainsi trouvés les ensembles I_s et J_s . Toute ligne de k_{I_sJ} est dite ligne supplémentaire, tandis que toute colonne de k_{IJ_s} est dite colonne supplémentaire. On dit que I_s est un ensemble de lignes supplémentaires adjointes au tableau k_{IJ} et J_s un ensemble de colonnes supplémentaires.

Notons que I_s (resp. J_s) peut se réduire à un élément, ou être vide ; dans ce dernier cas, on n'a pas de ligne (resp. colonne) supplémentaire.

2.2 *Projections de I_δ et J_δ sur le α-ème axe factoriel issu de k_{IJ}: facteurs sur I_δ et J_δ*

Soit i_s une ligne supplémentaire. Pour visualiser i_s sur le α-ème axe factoriel issu de k_{IJ}, on projette 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 cette projection s'écrit :

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

Dans cette formule, $G_\alpha^J = \{G_\alpha(j) \mid j \in J\}$ désigne le α-ème facteur sur J (i.e. l'ensemble des abscisses des projections des éléments de J sur l'axe α), $k(i_s)$ le total des éléments de la ligne i_s ($k(i_s) = \Sigma\{k(i_s,j) \mid j \in J\}$) et λ_α l'inertie de l'axe α.

Comme pour un élément principal (i.e. un élément de I) on peut calculer le cosinus (ou mieux le cosinus carré noté COR sur les listings) entre le profil de i_s et l'axe α, ce qui permet de juger de la qualité de la représentation de i_s sur cet axe. En cumulant les cosinus carrés de i_s avec deux (ou plusieurs) axes on obtient de même la qualité de la représentation de i_s sur le plan associé à ces deux axes (ou sur l'espace engendré par les axes considérés). Si l'on considère un ensemble A de facteurs, la reconstitution approchée de $k(i_s,j)$ à l'aide de cet ensemble de facteurs s'écrit :

$$\hat{k}(i_s,j) = (k(i_s)\ k(j)/k)(1 + \Sigma\{F_\alpha(i_s)\ G_\alpha(j)/\sqrt{\lambda_\alpha} \mid \alpha \in A\}) \qquad (2)$$

formule qui est identique à la formule de reconstitution que l'on aurait écrite pour un élément principal i de I, et où $k(j)$ désigne le total de la colonne j du tableau k_{IJ}, et k le total de tous les éléments de ce tableau. Si l'on prend pour A l'ensemble des facteurs non triviaux issus de k_{IJ}, on n'obtient pas en général (contrairement au cas d'un élément principal i de I) une reconstitution exacte de $k(i_s,j)$, puisque l'on projette le profil de i_s sur le sous espace engendré par les profils des éléments i de I, et que i_s n'appartient pas en général à ce sous espace.

Notons que la formule (2) peut s'interpréter comme une formule de régression, où l'on cherche à expliquer le vecteur r^J des $r^j = k(i_s,j)\ k/(k(i_s)\ k(j))$ en fonction des facteurs normés $\varphi_\alpha^J = G_\alpha^J/\sqrt{\lambda_\alpha}$ ($\alpha \in A$) et du facteur trivial constant. Placer un élément supplémentaire dans une analyse factorielle revient donc à faire une régression sur les facteurs issus de cette analyse, résultat évident géométriquement, puisque l'on peut toujours interpréter une régression comme une projection.

Si l'on considère maintenant une colonne supplémentaire j_s, on peut développer les mêmes considérations que précédemment. En particulier, l'abscisse $G_\alpha(j_s)$ de la projection du profil de j_s sur l'axe α s'écrit, si $F_\alpha^I = \{F_\alpha(i) \mid i \in I\}$ désigne le α-ème facteur sur I (de variance λ_α) et $k(j_s)$ la somme des éléments de la colonne j_s ($k(j_s) = \Sigma\{k(i,j_s) \mid i \in I\}$) :

$$G_\alpha(j_s) = \Sigma\{k(i,j_s)\ F_\alpha(i) \mid i \in I\}/(k(j_s)\sqrt{\lambda_\alpha}) \qquad (3)$$

Remarque

Considérons un élément supplémentaire, disons une ligne supplémentaire i_s pour fixer les idées. Outre le carré du cosinus $COR_\alpha(i_s)$

de i_s avec l'axe α, les listings comportent la quantité $CTR_\alpha(i_s) = (k(i_s)/k)\ F_\alpha^2(i_s)/\lambda_\alpha$ que l'on peut interpréter comme l'inertie (rapportée à λ_α) du point i_s par rapport à l'origine (i.e. le centre de gravité des i de I). Il faut bien noter que CTR_α n'est pas une contribution, puisque i_s ne sert pas à déterminer l'axe α (il a un poids nul dans l'analyse de k_{IJ}). Par contre, cette quantité peut servir pour effectuer des calculs d'inertie relatifs à l'ensemble I_s des éléments supplémentaires (cf. en particulier le § 3.4) ou pour tester si l'on peut considérer que sur l'axe α, le point i_s ne s'écarte pas significativement de l'origine (cf. Lebart et coll., Techniques de la description statistique, *op. cit.*, § IV 4.5).

3 *Utilisation des éléments supplémentaires pour représenter un groupe d'individus ou de variables. Application au cas d'un tableau de correspondance où l'un des ensembles est muni d'une partition*

Nous supposerons pour fixer les idées que l'on veuille représenter un sous ensemble d'individus I_1 de I. Nous proposons trois méthodes pour représenter I_1 sur les axes factoriels issus d'un tableau k_{IJ}, la dernière, qui revient à projeter sur les axes factoriels le centre de gravité du groupe d'individus concerné I_1 nous semblant la meilleure d'un point de vue pratique.

Après avoir présenté ces trois méthodes, nous considérons, comme application importante, le cas où l'ensemble I est muni d'une partition, cas dont les tableaux de correspondance ternaire et multiple sont des cas particuliers traités par la suite.

3.1 *Première méthode*

Pour représenter I_1, on considère la colonne supplémentaire notée s_I et définie par :

$$\forall\ i \in I_1 : s_i = 1$$

$$\forall\ i \in I - I_1 : s_i = 0$$

L'abscisse de la projection $G_\alpha(s)$ de s sur l'axe α s'écrit alors d'après (3) (où $k(i,j_s)$ est remplacé par s_i, et compte tenu de ce que le total de la colonne s_I est égal à Card I_1) :

$$G_\alpha(s) = \Sigma\{F_\alpha(i) \mid i \in I_1\}/(\text{Card } I_1 \sqrt{\lambda_\alpha}) \qquad (4)$$

Au coefficient $(1/\sqrt{\lambda_\alpha})$ près $G_\alpha(s)$ est la moyenne (ordinaire) des $F_\alpha(i)$ pour i dans I_1.

Cette représentation a le désavantage de donner égale importance à tous les individus quel qu'en soit le poids : en particulier, dans le cas où I_1 est l'ensemble I tout entier, s ne sera généralement pas projeté à l'origine ; car F_α^I est de moyenne nulle quand on utilise

la moyenne pondérée des $F_\alpha(i)$ par k(i) total de la ligne i de k_{IJ} , mais non (en général) si les poids sont tous égaux ; de façon précise, la moyenne $G_\alpha(s)$ se trouve décalée du côté de l'axe α où les individus légers se trouvent en plus grand nombre. En particulier , si s_I correspond à une colonne d'un tableau disjonctif complet placé en élément supplémentaire de k_{IJ} , et si cette colonne correspond à une modalité choisie par tous les individus, cette colonne , bien que n'apportant aucune information différenciant les individus , n'est pas projetée à l'origine, d'où la deuxième méthode proposée ci-après, où l'on pondère s_i par le total k(i) de la ligne i (total calculé sur J).

3.2 *Deuxième méthode*

On considère la colonne supplémentaire notée s'_I et définie par :

$$\forall\, i \in I : s'_i = k(i)s_i$$

soit :

$$\forall\, i \in I_1 : s'_i = k(i)$$

$$\forall\, i \in I-I_1 : s'_i = 0$$

L'abscisse de la projection $G_\alpha(s')$ de s' sur l'axe factoriel α s'écrit alors, en posant :

$$k(I_1) = \Sigma\{k(i) \mid i \in I_1\} :$$

$$G_\alpha(s') = \Sigma\{k(i)F_\alpha(i) \mid i \in I_1\}/(\sqrt{\lambda_\alpha}\ k(I_1)) \qquad (5)$$

En particulier, si $I_1 = I$, le point s' se projette bien à l'origine sur l'axe α .

On peut noter que si k(i) est constant sur I_1 , on obtient le même résultat que dans la première méthode.

On retiendra donc que si on met un tableau disjonctif complet T_{IL} en élément supplémentaire du tableau k_{IJ} , il faut pondérer chaque ligne i du tableau T par le total k(i) de la ligne i du tableau k.

3.3 *Troisième méthode*

On considère la ligne supplémentaire notée t_J et définie par :

$$\forall\, j \in J : t_j = \Sigma\{k(i,j) \mid i \in I_1\} ;$$

t_J est la somme des lignes i de k_{IJ} pour i appartenant à I_1. Le profil de t_J est alors le centre de gravité des profils des lignes i de I_1 (i étant affecté du poids k(i)). L'abscisse $F_\alpha(t)$ de la projection de t sur l'axe factoriel α s'écrit donc :

$$F_\alpha(t) = \Sigma\{k(i)\ F_\alpha(i) \mid i \in I_1\}/k(I_1) \qquad (6)$$

d'où l'on déduit :

$$F_\alpha(t) = \sqrt{\lambda_\alpha}\ G_\alpha(s') \qquad (7)$$

relation qui est classique dans le cas particulier où I_1 ne comporte qu'un élément.

On voit sur cette formule l'intérêt de cette troisième méthode relativement à la deuxième : elle donne une suite de coordonnées F_α décroissant plus rapidement avec α que les G_α ; donc des valeurs plus élevées pour les premiers COR (COR_1 , COR_2 ?,...) et en général une représentation de meilleure qualité sur les premiers axes. En revanche les coordonnées F_α sont plus faibles que les G_α , ce qui fait que sur un graphique, les points tendent à être peu lisibles, perdus au voisinage de l'origine.

3.4 *Cas d'un tableau k_{IJ} dont l'ensemble I est muni d'une partition*

Cette situation qui se rencontre fréquemment en pratique est en particulier réalisée comme on l'a déjà dit si on étudie un tableau ternaire (cf. § 4) ou de façon plus générale un tableau n-aire, ce qui est en particulier le cas des correspondances multiples (cf. § 5).

Soit C l'ensemble des classes de la partition de I , et

$$I = \cup \{I_c \mid c \in C\}$$

la partition associée.

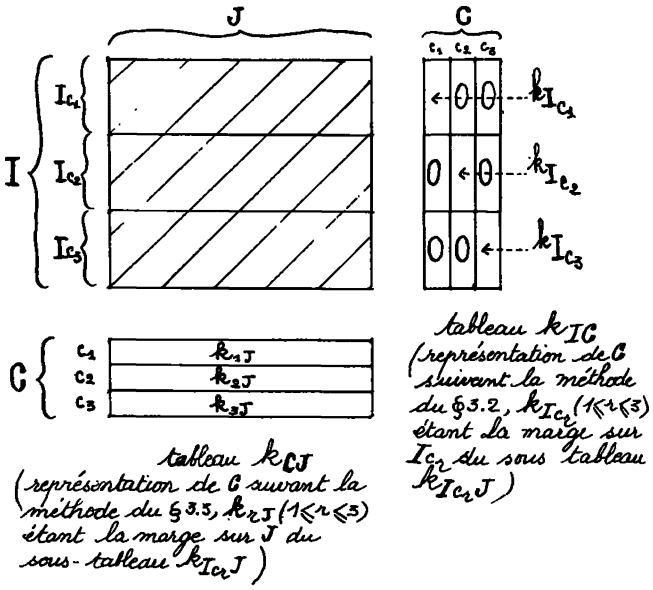

Figure 1 : Représentations de C
(on a hachuré le tableau principal analysé k_{IJ})

Pour représenter chaque classe c de C, on emploiera soit la deuxième, soit la troisième méthode préconisée précédemment, d'où les deux tableaux (cf. figure 1) k_{CJ} et k_{IC} définis par :

$$\forall\, j \in J,\ \forall\, c \in C : k(c,j) = \Sigma\{k(i,j) \mid i \in I_c\}$$

$$\forall\, i \in I,\ \forall\, c \in C : k(i,c) = \Sigma\{k(i,j) \mid j \in J\} = k(i) \text{ si } i \in I_c$$

$$= 0 \text{ sinon}$$

L'adjonction à k_{IJ} du tableau supplémentaire k_{CJ} est particulièrement intéressante si l'on veut effectuer des calculs d'inertie et de contributions pour voir l'importance de chaque classe I_c dans l'analyse factorielle de k_{IJ}, ces calculs se faisant très simplement, comme on va le voir, à partir des listings des résultats du programme TABET.

Soit $CR_\alpha(i)$ la contribution du point i à l'inertie λ_α du α-ième facteur issu de k_{IJ} (sur le listing, on a la quantité $CTR_\alpha(i) = CR_\alpha(i)/\lambda_\alpha$). La contribution de la classe I_c à λ_α s'écrit :

$$CR_\alpha(I_c) = \Sigma\{CR_\alpha(i) \mid i \in I_c\}$$

Par ailleurs si c_J désigne le profil de la ligne c du tableau k_{CJ}, $F_\alpha(c_J)$ l'abscisse de la projection de c_J sur l'axe α (i. e. la valeur du facteur α pour c_J), $f_c = \Sigma\{k(i) \mid i \in I_c\}/k$ (k désignant le total du tableau k_{IJ}) la masse de c_J, l'inertie du point c_J sur l'axe α s'écrit :

$$CR_\alpha(c_J) = f_c\, F_\alpha^2(c_J)$$

tandis que l'inertie de la classe c sur l'axe α vaut :

$$In_\alpha(I_c) = CR_\alpha(I_c) - CR_\alpha(c_J)$$

$In_\alpha(I_c)$ est la contribution de la classe c à l'inertie intra-classe $In_\alpha(I - C)$ du facteur α relativement à la partition C :

$$In_\alpha(I - C) = \Sigma\{In_\alpha(I_c) \mid c \in C\}$$

L'inertie interclasse du facteur α s'écrit :

$$In_\alpha(C) = \Sigma\{CR_\alpha(c_J) \mid c \in C\}$$

avec : $In_\alpha(C) + In_\alpha(I - C) = \lambda_\alpha$

relation classique de décomposition de l'inertie, que l'on retrouve aisément, à partir des égalités précédentes.

Le rapport $In_\alpha(C)/\lambda_\alpha$ permet de juger si en projection sur l'axe α, le nuage N(I) des individus peut être schématisé par le nuage N(C) des centres des classes, ou si les dispersions à l'intérieur des classes interviennent dans l'explication du facteur α (*).

(*) *On pourra consulter [CONTRIB. INTERCL.], in Prat. de l'A. des données, Vol 2, § III n° 6, pour un exemple de calcul d'inertie interclasse sur un axe, exemple relatif à une enquête sur les attitudes des paysans iraniens.*

Sur le listing on trouve outre les $CTR_\alpha(i)$ déjà mentionnés, les $CTR_\alpha(c_J) = CR_\alpha(c_J)/\lambda_\alpha$, ainsi que le total des $CTR_\alpha(c_J)$, i.e. le rapport $IN_\alpha(C)/\lambda_\alpha$ dont on vient de voir l'intérêt. On peut ainsi effectuer sans difficultés tous les calculs de contributions et d'inerties précédents, l'inertie de l'axe α ayant été normaliséeà 1.

Tous les calculs précédents restent valables si au lieu de considérer l'inertie en projection sur un axe α, on considère l'inertie dans l'espace R_J tout entier. Il suffit dans les formules données d'enlever l'indice α, et de remplacer λ_α par l'inertie totale égale à la somme des valeurs propres issues de l'analyse de k_{IJ}, inertie qu'on notera Trace(I) et qu'il est immédiat de calculer à partir d'une valeur propre λ_α et du pourcentage d'inertie associé $\tau_\alpha = 100\,\lambda_\alpha/\text{Trace(I)}$. Sur le listing, il faut alors considérer les quantités INR au lieu des CTR_α ($INR(i) = CR(i)/\text{Trace(I)}$, CR(i) étant l'inertie du point i, ou contribution de i à l'inertie totale, $INR(c_J) = CR(c_J)/\text{Trace(I)}$, $CR(c_J)$ étant l'inertie de c_J etc.).

Remarques

1) Analyse de k_{IJ} avec k_{IC} en supplémentaire.

Supposons que l'on effectue l'analyse de k_{IJ}, avec k_{IC} adjoint en supplémentaire (mais sans placer k_{CJ} en supplémentaire). A partir du listing de cette analyse, on peut effectuer sans difficultés les calculs précédents d'inerties interclasse et intraclasse sur un axe α (en effet, d'après (7), on a, avec des notations évidentes :

$G_\alpha(c_I) = F_\alpha(c_J)/\sqrt{\lambda_\alpha}$, $CTR_\alpha(c_I) = CTR_\alpha(c_J)/\lambda_\alpha$) ; par contre, on ne peut plus calculer facilement les inerties interclasse et intraclasse dans l'espace R_J tout entier.

2) Analyse de k_{CJ} avec k_{IJ} en supplémentaire.

Au lieu d'analyser k_{IJ}, on analyse souvent k_{CJ}, k_{IJ} étant adjoint en tableau supplémentaire à k_{CJ} (cf. §§ 1.4.2, 1.4.3 dans le cas de la régression ; §§ 1.3, 4, ainsi que l'exemple d'application à la fin du § 6, dans le cas d'un tableau ternaire dont on analyse les marges d'ordre 2 ; § 5 dans le cas des correspondances multiples).

Dans ce cas, on peut encore effectuer tous les calculs précédents et en particulier calculer l'inertie (totale-interclasse-intraclasse) de I sur un axe ou dans l'espace R_J tout entier. On a en particulier :

$$In(I) = \text{Trace(C)}\,(\Sigma\{INR(i) \mid i \in I\})$$

$$In_\alpha(I) = \lambda_\alpha\ (\Sigma\{CTR_\alpha(i) \mid i \in I\})$$

In(I) désignant l'inertie totale de I*,Trace (C) celle de C (i.e. l'inertie interclasse de I), $In_\alpha(I)$ l'inertie de I sur l'axe α issu de k_{CJ}, λ_α la valeur propre correspondante (i.e. l'inertie de C sur cet axe).

Ce type d'analyse, ainsi que ses cas particuliers (comme le cas où la partition de I est constituée d'une part par une classe I_1 de I et d'autre part par les Card($I - I_1$) classes associées à chaque élément de $I - I_1$; ou encore le cas où l'on analyse une bande dans un tableau de Burt (cf. 2-ème exemple du § 1.5)) sont traités en détail dans [*BANDES BURT*] déjà cité.

3) Outre les calculs d'inertie précédents, on peut pour comparer les résultats de l'analyse de k_{IJ}, avec ceux de k_{CJ}, faire des calculs de corrélation entre les facteurs issus de ces deux analyses (cf.§ 6).

---

* *inertie notée précédemment Trace(I)*