

REVUE DE STATISTIQUE APPLIQUÉE

J. PAGÈS

Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes

Revue de statistique appliquée, tome 50, n° 4 (2002), p. 5-37

http://www.numdam.org/item?id=RSA_2002__50_4_5_0

© Société française de statistique, 2002, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ANALYSE FACTORIELLE MULTIPLE APPLIQUÉE AUX VARIABLES QUALITATIVES ET AUX DONNÉES MIXTES

J. PAGÈS

*Laboratoire de mathématiques appliquées
Pôle d'Enseignement Supérieur et de Recherche Agronomique, 35042 Rennes cedex
email : pages@agrorennes.educagri.fr*

RÉSUMÉ

L'Analyse Factorielle Multiple (AFM) s'applique à des tableaux dans lesquels un ensemble d'individus est décrit par plusieurs ensembles de variables. Le cas des variables quantitatives est bien connu et les nombreuses applications réalisées en font une méthode éprouvée. Le principe de l'extension de la méthode au cas des variables qualitatives et à celui incluant les deux types de groupes de variables a déjà été présenté mais sans envisager tous les aspects techniques impliqués par ces extensions. Il est montré ici de façon approfondie comment l'AFM s'applique à ces deux cas. Un exemple d'application sur des données de petite taille illustre les propriétés de la méthode.

Mots-clés : Analyse Factorielle Multiple, Analyse des Correspondances Multiples, Analyse en Composantes Principales.

ABSTRACT

Multiple Factor Analysis (MFA) applies to tables in which a set of individuals is described by several sets of variables. The case of the quantitative variables is well-known and many applications have been carried out. The principle of the extension of MFA to the case of qualitative variables and to that including the two types of groups of variables was already presented but without considering all the technical aspects implied by these extensions. It is shown here thoroughly how MFA applies to these two cases. An example of application on a small data set illustrates the properties of the method.

Keywords : Multiple Factor Analysis, Multiple Correspondence Analysis, Principal Components Analysis.

1. Introduction

L'analyse exploratoire de tableaux dans lesquels un ensemble d'individus est décrit par un ensemble de variables qualitatives est classiquement réalisée par l'analyse des correspondances multiples (Lebart *et al.* 1977). Un exemple typique de ces données est fourni par les enquêtes; les variables sont alors des questions

comportant plusieurs modalités de réponse parmi lesquelles chaque enquêté en choisit une et une seulement.

En pratique, les questionnaires sont toujours structurés en thèmes qui définissent autant de groupes de variables, par exemple un ensemble d'opinions, un ensemble de comportements, un ensemble de repères sociaux, etc. Une première façon d'analyser les données consiste à focaliser l'attention sur un thème, en introduisant dans l'analyse les variables correspondantes en actif, les autres étant introduites en supplémentaire. Cette méthodologie est excellente et éprouvée (Cazes 1982, Lebart 1989, Morineau 1993). Si l'on souhaite analyser plusieurs thèmes simultanément en actif, plusieurs problèmes apparaissent, le premier d'entre eux étant l'équilibre entre les groupes de variables. En effet, que dire d'une analyse réalisée à partir de plusieurs thèmes mais dont par exemple les deux premiers facteurs ne seraient dus qu'à l'un d'entre eux ?

Une solution à ce problème est fournie par l'Analyse Factorielle Multiple (AFM) (Escofier et Pagès 1983, 1998). Cette méthode a été développée initialement pour les variables quantitatives. Elle s'applique aussi aux variables qualitatives, ce qui est montré ici de façon détaillée. La même méthode s'appliquant à la fois aux variables quantitatives et qualitatives, il est naturel de l'appliquer à des données dites mixtes, dans lesquelles un ensemble d'individus est décrit à la fois par des groupes de variables qualitatives et des groupes de variables quantitatives. Les propriétés de l'AFM dans le traitement de telles données sont décrites ici.

Dans un premier temps, les notations sont introduites en rappelant quelques relations de l'Analyse Factorielle Générale (AFG), terme désignant la recherche des axes principaux d'inertie des deux ensembles duaux de points pondérés associés aux lignes et aux colonnes d'un tableau numérique (Lebart *et al.* 1999). Les méthodes usuelles, Analyse en Composantes Principales (ACP), Analyse Factorielle des Correspondances (AFC) et Analyse des Correspondances Multiples (ACM) sont alors présentées comme cas particulier de l'AFG.

Dans un second temps, l'ACM pondérée (ACMP), introduite initialement par (Cazes 1980), est détaillée en présentant des propriétés inédites de cette méthode.

La troisième partie est consacrée à l'AFM de variables qualitatives. Les principales caractéristiques de la méthode sont rappelées et transposées au cas de variables qualitatives.

La quatrième partie présente les propriétés de la méthode appliquées à des tableaux mixtes. Enfin, un exemple numérique mixte illustre la méthode.

2. Analyse factorielle générale

2.1. Notations; principales relations de l'analyse factorielle générale

2.1.1. Deux nuages

Soit X le tableau analysé, ayant I lignes et K colonnes. Son terme général est x_{ik} . Pour fixer les idées, en ACP, les lignes sont des individus et les colonnes des variables numériques.

On affecte :

- le poids p_i à la ligne i ; ces pondérations sont regroupées dans une matrice diagonale de dimension I notée D ; $D(i, i) = p_i$; par commodité : $\sum_i p_i = 1$;
- la pondération m_k à la colonne k ; ces poids sont regroupés dans une matrice diagonale de dimension K notée M ; $M(k, k) = m_k$;
- l'ensemble des lignes est représenté par un ensemble de points, noté N_I , dans l'espace à K dimensions, noté R^K , dont chaque dimension est associée à une colonne. La matrice M induit une métrique euclidienne dans R^K ;
- l'ensemble des colonnes est représenté par un ensemble de points, noté N_K , dans l'espace à I dimensions, noté R^I , dont chaque dimension est associée à une ligne. La matrice D induit une métrique euclidienne dans R^I .

L'analyse factorielle générale fournit les axes principaux d'inertie des ensembles N_I et N_K .

2.1.2. Analyse du nuage des lignes

Dans R^K , le vecteur unitaire de l'axe principal de rang s de N_I , noté u_s , vérifie :

$$X'DXM u_s = \lambda_s u_s \quad \|u_s\|_M^2 = u_s' M u_s = 1$$

avec λ_s valeur propre de rang s de $X'DXM$.

Les coordonnées des lignes (points de N_I) le long de u_s sont regroupées dans F_s de dimension $(I, 1)$: $F_s = XM u_s$; F_s est dit «facteur sur I de rang s » (en ACP, F_s est la composante principale de rang s). F_s est un vecteur de R^I de norme :

$$\|F_s\|_D^2 = F_s' D F_s = \lambda_s$$

2.1.3. Analyse du nuage des colonnes

Dans R^I , le vecteur unitaire de l'axe principal de rang s de N_K , noté v_s , vérifie :

$$v_s = \frac{F_s}{\|F_s\|} = \frac{1}{\sqrt{\lambda_s}} F_s$$

Les coordonnées des colonnes (points de N_K) le long de v_s sont regroupées dans G_s de dimension $(K, 1)$: $G_s = X' D v_s$; G_s est dit «facteur sur K de rang s ».

G_s est un vecteur de R^K de norme :

$$\|G_s\|_M^2 = G_s' M G_s = \lambda_s$$

Le vecteur v_s vérifie aussi :

$$X M X' D v_s = \lambda_s v_s$$

La matrice $X M X'$ rassemble les produits scalaires entre individus. Elle est usuellement notée W .

En outre, les directions principales de N_I sont liées aux projections de N_K par :

$$u_s = \frac{G_s}{\|G_s\|} = \frac{1}{\sqrt{\lambda_s}} G_s$$

2.1.4. Relations de transition

Les coordonnées des lignes (F_s) et celle des colonnes (G_s) sont liées par les relations dites de transition :

$$F_s = \frac{1}{\sqrt{\lambda_s}} X M G_s$$

Soit, pour la ligne i :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} m_k G_s(k)$$

En outre :

$$G_s = \frac{1}{\sqrt{\lambda_s}} X' D F_s$$

Soit, pour la colonne k :

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i x_{ik} p_i F_s(i)$$

Il est ainsi possible d'interpréter la position d'une ligne (resp. colonne) en fonction de la position de l'ensemble des colonnes (resp. lignes). Ce couple de relations est connu sous cette forme surtout dans le cadre de l'AFC.

2.2. Trois cas particuliers

2.2.1. Analyse en composantes principales

En ACP le tableau X est toujours centré. En ACP normée, X est centré et réduit.

Les poids p_i sont définis par l'utilisateur. En général : $p_i = 1/I \quad \forall i$.

Les logiciels usuels imposent $m_k = 1/\forall k$.

2.2.2. Analyse des correspondances

Bien que présentée presque toujours à partir des profils et de la métrique du χ^2 (Benzécri 1973), l'analyse des correspondances peut être introduite de multiples façons (Cazes *et al.* 1977; Cazes *et al.* 1988).

Les données sont rassemblées dans un tableau dont les lignes représentent les I modalités d'une variable Q_1 et les colonnes les K modalités d'une variable Q_2 . On considère le tableau de fréquence, noté F , dont le terme général, noté f_{ik} , est la fréquence relative observée des individus possédant à la fois la modalité i de Q_1 et k de Q_2 . En outre :

$$\sum_{ik} f_{ik} = 1; \quad \sum_k f_{ik} = f_{i.}; \quad \sum_i f_{ik} = f_{.k}$$

L'analyse des correspondances peut être vue comme une AFG telle que :

$$x_{ik} = \frac{f_{ik}}{f_{i.}f_{.k}} - 1; \quad p_i = f_{i.}; \quad m_k = f_{.k}$$

Le tableau de terme général x_{ik} est centré par ligne et par colonne ce qui assure un rôle symétrique aux lignes et aux colonnes.

On peut aussi réaliser l'AFG suivante :

$$x_{ik} = \frac{f_{ik}}{f_{i.}f_{.k}}; \quad p_i = f_{i.}; \quad m_k = f_{.k}$$

Cette seconde analyse fournit exactement les résultats de la première à condition d'écarter le facteur associé à la première valeur propre qui vaut toujours 1. Son tableau X étant plus simple, nous utilisons par la suite cette seconde présentation de l'AFG.

En notant D et M les matrices diagonales rassemblant respectivement les poids des lignes ($f_{i.}$) et des colonnes ($f_{.k}$), la matrice X de cette seconde AFG s'écrit : $D^{-1}FM^{-1}$.

2.2.3. Analyse des correspondances multiples

L'ACM peut aussi être présentée de multiples façons à tel point que, selon les pays, elle porte un nom différent (Tenenhaus et Young 1985). Le point de vue adopté ici est celui introduit initialement par (Benzécri 1973), (Lebart *et al.* 1977).

I individus sont décrits par J variables qualitatives totalisant K modalités. La variable j possède K_j modalités.

Les données sont regroupées dans un tableau disjonctif complet, noté Y , dont le terme général y_{ik} vaut 1 si l'individu i possède la modalité k et 0 sinon. On a :

$$\forall i \in I : \sum_k y_{ik} = J \quad \sum_i y_{ik} = I_k$$

en notant I_k le nombre d'individus possédant la modalité k .

L'ACM est, en France, usuellement présentée comme une AFC sur le tableau disjonctif complet, soit comme l'AFG suivante :

$$\begin{aligned} x_{ik} &= \frac{y_{ik}I}{I_k} & X &= IY E^{-1} \\ p_i &= \frac{1}{I} & D &= \frac{1}{I} Id_I & m_k &= \frac{I_k}{IJ} & M &= \frac{E}{IJ} \end{aligned}$$

en appelant :

- E la matrice diagonale de dimension K regroupant les effectifs des modalités : $E(k, k) = I_k$.
- Id_I la matrice identité de dimension I .

2.3. L'ACM vue comme une ACP particulière

Les résultats de l'ACM peuvent aussi être obtenus en appliquant une ACP à un Tableau Disjonctif Complet (TDC).

En ACM la distance entre deux individus s'écrit :

$$d^2(i, l) = \frac{1}{J} \sum_k \frac{I}{I_k} (y_{ik} - y_{lk})^2$$

On obtient la même distance en considérant le TDC comme un tableau de variables quantitatives non réduites et en affectant la pondération $I/(JI_k)$ à l'indicatrice k . De même, on obtient la même distance en centrant et en réduisant le TDC à condition d'affecter la pondération $(I - I_k)/IJ$ à l'indicatrice k . En effet :

$$d^2(i, l) = \frac{1}{J} \sum_k \frac{(I - I_k)}{I} \frac{I^2 (y_{ik} - y_{lk})^2}{(I - I_k) I_k}$$

Rappel : la variance de l'indicatrice k vaut : $(I - I_k)I_k/I^2$

Ainsi, les résultats d'une ACM peuvent être obtenus à l'aide d'une ACP pondérée du TDC. Cette équivalence étant exploitée en AFM, il est utile d'apporter quelques précisions quant à la pratique de cette ACP.

En ce qui concerne les modalités, une telle ACP conduit à la représentation des indicatrices par leurs coefficients de corrélation avec les facteurs. Bien que cette représentation soit interprétable, il est préférable de construire plutôt celle des centres de gravité des individus qui possèdent une même modalité, homothétique axe par axe de la représentation usuelle de l'ACM. Soit g_k le centre de gravité des individus qui possèdent la modalité k . Les relations de transition de l'ACM deviennent, d'une part :

$$F_s(g_k) = \frac{1}{I_k} \sum_i y_{ik} F_s(i) = \sqrt{\lambda_s} G_s(k)$$

ce qui indique qu'une modalité apparaît au barycentre des individus qui la possèdent (par définition), et d'autre part :

$$F_s(i) = \frac{1}{\lambda_s} \frac{1}{J} \sum_k y_{ik} F_s(g_k)$$

ce qui indique que le long de l'axe de rang s , un individu apparaît au barycentre des modalités qu'il possède, au coefficient λ_s près (et non sa racine comme en ACM usuelle).

Remarque sur les logiciels. Dans les programmes d'AFM inclus dans les logiciels SPAD (2000) et LADDAD (1997), les indicatrices sont initialement centrées, réduites et munies de la pondération décrite ci-dessus.

3. ACM pondérée

3.1. Définition

Dans une ACM, si l'on duplique une variable, on lui attribue un rôle a priori deux fois plus important. Il est donc aisée de concevoir une ACM dans laquelle on affecte une pondération aux variables. La possibilité de pondérer des variables en ACM a déjà été exploitée par (Cazes 1980). Concrètement, on peut mettre en œuvre une telle analyse en appliquant un programme d'AFC à un TDC pondéré dans lequel y_{ik} ($= 0$ ou 1), valeur du TDC, est remplacé par $y_{ik}c_j$ (avec k une indicatrice appartenant à la variable j et c_j la pondération de la variable j). Ce TDC pondéré a une marge colonne constante et égale à $c = \sum c_j$

D'après 2.2.2, l'AFC d'un tel tableau revient à réaliser l'AFG suivante :

$$x_{ik} = \frac{y_{ik}c_j(Ic)}{I_k c_j c} = \frac{y_{ik}I}{I_k} \quad p_i = 1/I \quad m_k = \frac{I_k}{I} \frac{c_j}{c}$$

L'ACM pondérée (ACMP) ne diffère de l'ACM usuelle telle qu'elle est présentée en 2.2.3, que par la pondération des modalités. La pondération de la modalité k , appartenant à la variable j peut être vue comme le produit de la pondération de la modalité dans la variable j (I_k/I) par la pondération de la variable j dans l'ensemble des variables ($1/J$ dans le cas de l'ACM usuelle, c_j/c dans le cas de l'ACMP).

Lorsque c_j est un nombre entier, il revient au même de réaliser l'ACMP ou l'ACM usuelle sur le tableau disjonctif complet dont les modalités de la variable j ont été dupliquées c_j fois.

3.2. Nuage des modalités en ACM pondérée

En ACM, le nuage des modalités possède des propriétés remarquables (Lebart *et al.* 1977). Les principales caractéristiques de ce nuage en ACM et en ACM pondérée sont regroupées dans le tableau 1, dans lequel l'ACM (resp. l'ACM pondérée) est vue comme une AFC appliquée au TDC (resp. TDC pondéré).

TABLEAU 1
Quelques propriétés comparées de l'ACM et de l'ACM pondérée

	ACM	ACM pondérée
Terme général du tableau brut	$y_{ik}(= 0 \text{ ou } 1)$	$y_{ik}c_j$
Poids de la modalité k (appartenant à la variable j)	$\frac{I_k}{I} \frac{1}{J}$	$\frac{I_k}{I} \frac{c_j}{c}$
Poids de l'ensemble des modalités de la variable j	$\frac{1}{J}$	$\frac{c_j}{c}$
Barycentre des modalités de la variable j	$\frac{1}{J}$	$\frac{1}{J}$
Inertie de la modalité k (appartenant à la variable j)	$\frac{1}{J} \left(1 - \frac{I_k}{I}\right)$	$\frac{c_j}{c} \left(1 - \frac{I_k}{I}\right)$
Inertie de l'ensemble des modalités de la variable j	$\frac{1}{J}(K_j - 1)$	$\frac{c_j}{c}(K_j - 1)$
Inertie totale du nuage des modalités	$\sum_j \frac{1}{J} K_j - 1$	$\sum_j \frac{c_j}{c} K_j - 1$
Carré de la distance entre deux modalités k et h	$\sum_i I \left(\frac{y_{ik}}{I_k} - \frac{y_{ih}}{I_h}\right)^2$	$\sum_i I \left(\frac{y_{ik}}{I_k} - \frac{y_{ih}}{I_h}\right)^2$
Carré de la distance entre deux individus i et l	$\sum_j \frac{1}{J} \sum_{k \in K_j} \frac{I}{I_k} (y_{ik} - y_{lk})^2$	$\sum_j \frac{c_j}{c} \sum_{k \in K_j} \frac{I}{I_k} (y_{ik} - y_{lk})^2$

Ces relations montrent que l'on passe de l'ACM à l'ACM pondérée en remplaçant la pondération de la variable j (égale à $1/J$ en ACM) par c_j/c . Remarquons que, bien qu'ayant introduit initialement la pondération c_j , la pondération d'une variable vaut finalement c_j/c : en ACM pondérée (comme en ACM), la somme des pondérations des variables vaut automatiquement 1.

3.3. Analyse du TDC pondéré et du tableau de Burt pondéré

En ACM, l'équivalence entre l'analyse du TDC et celle du tableau de Burt (Lebart *et al.* 1977) joue un rôle important aussi bien au niveau de la justification de la méthode que de l'interprétation des résultats.

Le tableau de Burt, noté B , croise l'ensemble des modalités avec lui-même. Son terme général, à l'intersection de la ligne h (appartenant à la variable l) et de la colonne k (appartenant à la variable j) vaut : $B(h, k) = I_{hk}$ en notant I_{hk} le nombre d'individus possédant à la fois la modalité h (de la variable l) et la modalité k (de la variable j). Matriciellement : $B = Y'Y$. Nous examinons ce que devient cette équivalence dans le cas pondéré.

L'ACM pondérée peut aussi être vue comme une AFC du tableau de Burt pondéré. Dans ce tableau, noté B_p , on trouve, à l'intersection de la ligne h (appartenant à la variable l) et de la colonne k (appartenant à la variable j) $B_p(h, k) = I_{kh}c_jc_l$. Ce tableau s'obtient en multipliant à gauche le TDC pondéré par son transposé. Il dérive du tableau de Burt usuel en multipliant chaque terme du sous-tableau croisant les variables j et l par c_jc_l .

Les deux marges de B_p sont identiques et ont pour k^e terme (la modalité k appartenant à la variable j) : I_kc_jc ; la somme de tous les termes du tableau vaut : Ic^2 .

Considérons les AFC du TDC pondéré (TDCP) et du tableau de Burt pondéré (B_p). De façon strictement analogue à l'ACM usuelle, on vérifie les propriétés suivantes :

- les profils-lignes de B_p sont dans le même espace euclidien que les profils lignes du TDCP;
- le profil-ligne k de B_p est le barycentre des profils-lignes du TDCP qui possèdent la modalité k ;
- ces deux AFC conduisent aux mêmes facteurs sur les modalités, chaque valeur propre de la première étant égale au carré de la valeur propre de même rang de la seconde.

L'équivalence entre ces deux AFC apparaît au tableau 2, qui explicite l'analyse factorielle générale (les pondérations des lignes – respectivement des colonnes – constituent la métrique dans l'espace des colonnes – respectivement des lignes; cf. § 2.1) dans le cas de ces deux analyses.

Ce tableau montre que l'ACM pondérée possède la propriété très importante de l'ACM usuelle selon laquelle il revient au même d'analyser le nuage des individus et celui de l'ensemble des centres de gravité chacun associé à une modalité.

TABLEAU 2
Comparaison entre l'analyse du TDC et celle du T. de Burt
dans le cas usuel et dans le cas pondéré

	AFC du TDC	AFC du TB	AFC du TDCP	AFC du TBP
Tableau analysé (X)	IYE^{-1}	$IE^{-1}Y'YE^{-1}$	IYE^{-1}	$IE^{-1}Y'YE^{-1}$
Pondération des colonnes (M)	$\frac{E}{I} \frac{1}{J}$	$\frac{E}{I} \frac{1}{J}$	$\frac{E}{I} \frac{C}{C}$	$\frac{E}{I} \frac{C}{C}$
Poids des lignes (D)	$\frac{I_d}{I}$	$\frac{E}{I} \frac{1}{J}$	$\frac{I_d}{I}$	$\frac{E}{I} \frac{C}{C}$
Diagonalisation ($X'DXM$)	$\frac{1}{J}E^{-1}Y'Y$	$\frac{1}{J^2}E^{-1}Y'YE^{-1}Y'Y$	$E^{-1}Y'Y\frac{C}{C}$	$E^{-1}Y'Y\frac{C}{C}E^{-1}Y'Y\frac{C}{C}$
V. pr. rang s	λ_s	$(\lambda_s)^2$	μ_s	$(\mu_s)^2$

La première colonne rappelle les notations et les formules de l'analyse factorielle générale (Cf. § 2.1). Y : tableau disjonctif complet; E : matrice diagonale contenant les effectifs des modalités; I_d : matrice identité de dimension I ; C : matrice diagonale (K, K) contenant les pondérations des variables affectées à chaque modalité $C(k, k) = c_j$ si $k \in$ variable j .

TABLEAU 2bis
Relations de transition associées aux analyses du tableau 2

	AFC du TDC	AFC du TB	AFC du TDCP	AFC du TBP
$F_s = \frac{1}{\sqrt{\lambda_s}} XMG_s$	$F_s = \frac{1}{\sqrt{\lambda_s}} \frac{1}{J} YG_s$	$F_s = \frac{1}{\lambda_s} \frac{1}{J} E^{-1}Y'Y G_s$	$F_s = \frac{1}{\sqrt{\mu_s}} Y\frac{C}{C} G_s$	$F_s = \frac{1}{\mu_s} E^{-1}Y'Y\frac{C}{C} G_s$
$G_s = \frac{1}{\sqrt{\lambda_s}} X'DF_s$	$G_s = \frac{1}{\sqrt{\lambda_s}} E^{-1}Y'F_s$	$G_s = \frac{1}{\lambda_s} \frac{1}{J} E^{-1}Y'Y F_s$	$G_s = \frac{1}{\sqrt{\mu_s}} E^{-1}Y'F_s$	$G_s = \frac{1}{\mu_s} E^{-1}Y'Y\frac{C}{C} F_s$

3.4. Relations de transition en ACM pondérée

Ces relations, sous leur forme matricielle, sont rassemblées dans le tableau 2 bis.

3.4.1. Relations usuelles découlant de l'analyse du TDC

Pour appliquer directement les relations de l'AFG (§ 2.1), λ_s , F_s et G_s désignent ici alternativement les résultats d'une ACM et ceux d'une ACM pondérée.

En ACM et en ACM pondérée, une modalité est (à un coefficient près) au barycentre des individus qui la possèdent :

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{I_k} \sum_i y_{ik} F_s(i) = \frac{1}{\sqrt{\lambda_s}} F_s(g_k)$$

En ACM un individu est (à un coefficient près) au barycentre des modalités (affectées chacune de la même pondération) qu'il possède :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{J} \sum_{k \in K} y_{ik} G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{1}{J} \sum_{k \in K_j} y_{ik} G_s(k)$$

En ACM pondérée, un individu est (à un coefficient près) au barycentre des modalités (affectées chacune de la pondération de la variable à laquelle elle appartient) qu'il possède :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{c_j}{c} \sum_{k \in K_j} y_{ik} G_s(k) = \frac{1}{\lambda_s} \sum_j \frac{c_j}{c} \sum_{k \in K_j} y_{ik} F_s(g_k)$$

3.4.2. Relation découlant de l'analyse du tableau de BURT

On peut utiliser la relation de transition issue de l'AFC de B.

En ACM usuelle, cette relation est trop peu souvent citée. Soit, en notant $F_s(g_k)$ la coordonnée de la modalité k sur l'axe de rang s et I_{kl} le nombre d'individus possédant à la fois les modalités k et l :

$$F_s(g_k) = \frac{1}{\lambda_s} \frac{1}{I_k J} \sum_l I_{kl} F_s(g_l) = \sqrt{\lambda_s} G_s(k)$$

Au coefficient λ_s près, la modalité k apparaît au barycentre de l'ensemble des modalités, chaque modalité l étant affectée de la pondération I_{kl} ($\sum_l I_{kl} = I_k J$).

Ainsi, une modalité apparaît du côté des modalités auxquelles elle s'associe le plus, propriété exacte plus facile à exploiter, lors d'une interprétation, que les projections de distances entre indicatrices induites par l'analyse du TDC.

Dans l'ACM pondérée cette relation devient :

$$F_s(g_k) = \frac{1}{\lambda_s} \frac{1}{I_k} \sum_{j \in J} \frac{c_j}{c} \sum_{l \in K_j} I_{kl} F_s(g_l) = \sqrt{\lambda_s} G_s(k)$$

Cette relation est fondamentalement la même que la précédente, en affectant à la variable j non plus la pondération $1/J$ mais c_j/c .

3.5. ACM pondérée et ACP

De même que l'ACM usuelle, l'ACM pondérée peut être obtenue en appliquant une ACP sur le TDC, l'indicatrice k (appartenant à la variable j) ayant la pondération : $\frac{I}{I_k} \frac{c_j}{c}$ si l'ACP est non normée; $\frac{I - I_k}{I} \frac{c_j}{c}$ si l'ACP est normée.

Ce résultat découle directement de la distance entre deux individus donnée tableau 1.

Si l'on réalise l'ACMP à l'aide d'un programme d'AFC (appliquée au TDC pondéré), les pondérations des variables ne sont prises en compte qu'au travers du rapport c_j/c puisque, dans l'AFC, les termes du tableau étudié sont «d'abord» divisés par leur somme. Autrement dit, la somme des pondérations qui interviennent effectivement vaut automatiquement 1.

Si l'on réalise une ACM pondérée à l'aide d'un programme d'ACP, alors il est possible d'introduire des pondérations dont la somme c est différente de 1, ce qui revient à faire l'analyse avec les poids c_j et non c_j/c .

Soient $\{F_s, G_s, \lambda_s\}$ les résultats de cette ACM avec les poids c_j et $\{F_s^r, G_s^r, \lambda_s^r\}$ les résultats de l'ACM pondérée de référence, c'est-à-dire avec les poids c_j/c . Dans l'ACM avec les pondérations c_j , par rapport à l'ACM de référence :

- le nuage des indicatrices est inchangé : $G_s = G_s^r$; les pondérations de ces indicatrices étant multipliées par c , on a ; $\lambda_s = c\lambda_s^r$;
- le changement de métrique dans R^K peut être vu comme une homothétie du nuage des individus de rapport \sqrt{c} ; d'où $F_s = \sqrt{c}F_s^r$, ce qui conduit également à $\lambda_s = c\lambda_s^r$;
- cette homothétie du nuage d'individus rejailit sur la représentation des modalités en tant que centre de gravité des individus : $F_s(g_k) = \sqrt{c}F_s^r(g_k)$;
- la relation liant la coordonnée d'un individu et celle des modalités (en tant que centres de gravité) peut s'exprimer comme dans l'ACM de référence (i.e. avec λ_s^r et c_j/c) :

$$F_s(i) = \frac{1}{\lambda_s} \sum_j c_j \sum_{k \in K_j} y_{ik} F_s(g_k) = \frac{1}{\lambda_s^r} \sum_j \frac{c_j}{c} \sum_{k \in K_j} y_{ik} F_s(g_k)$$

Ainsi, si l'on s'intéresse aux modalités en tant que barycentres d'individus (et non en tant qu'indicatrices), les deux relations de dualité entre ces deux types de points, individus et modalités, et donc leurs positions relatives, sont inchangées que l'on considère les pondérations c_j ou c_j/c . Cette propriété sera utilisée en AFM, en introduisant des pondérations c_j ayant de bonnes propriétés tout en conservant les relations de l'ACM de référence.

4. AFM de variables qualitatives

4.1. Notations

Au sein du TDC, dont le terme général est toujours noté y_{ik} , les indicatrices sont maintenant rassemblées en groupes qui incluent plusieurs variables. Les indicatrices d'une même variable appartiennent toujours à un même groupe : on parle indifféremment de groupe d'indicatrices ou de groupe de variables. Comme dans toute AFM, nous réservons la lettre j aux groupes (groupe j ; J : nombre de groupes); la lettre q est réservée aux variables (variable q ; Q : nombre de variables).

En résumé :

K_j : nombre de modalités du groupe j ; K : nombre total de modalités ($K = \sum K_j$);

Q_j : nombre de variables du groupe j ; Q : nombre total de variables ($Q = \sum Q_j$).

4.2. AFM et analyse factorielle

4.2.1. Principe de la pondération des groupes de variables

La problématique qui découle de la prise en compte de groupes dans un ensemble de variables est identique que les variables soient quantitatives ou qualitatives : équilibrer l'influence a priori des différents groupes, étant entendu qu'un groupe présentant de nombreuses dimensions influencera, dans une analyse globale, plus d'axes qu'un autre présentant peu de dimensions (voire une seule). La solution choisie par l'AFM – harmoniser les inerties axiales maximum des nuages associés aux différents groupes – présentée classiquement dans le cadre de variables quantitatives, n'est pas spécifique de ces variables et s'applique aux variables qualitatives.

4.2.2. L'AFM de variables qualitatives repose sur une ACM pondérée

L'analyse factorielle usuelle d'un ensemble de variables qualitatives étant l'ACM, il est naturel de faire reposer l'AFM de J groupes de variables qualitatives sur une ACM pondérée. Dans cette ACM pondérée, les pondérations des variables, induites par l'égalisation (à 1) des inerties maximums des sous-nuages, découlent du raisonnement suivant.

Dans le groupe j avant la pondération (i.e. dans l'ACM de ce groupe), la distance entre deux individus i et l s'écrit :

$$d^2(i^j, l^j) = \frac{1}{Q_j} \sum_{k \in K_j} \frac{I}{I_k} (y_{ik} - y_{lk})^2$$

en notant i^j l'individu i considéré du point de vue des variables du seul groupe j (appelé «individu partiel» en AFM).

Pour rendre égale à 1 l'inertie maximum du nuage associé au groupe j , on divise la pondération de chaque modalité du groupe j par λ_1^j (première valeur propre de l'ACM du groupe j). A l'issue de cette harmonisation, la distance entre i^j et l^j s'écrit :

$$d^2(i^j, l^j) = \frac{1}{Q_j \lambda_1^j} \sum_{k \in K_j} \frac{I}{I_k} (y_{ik} - y_{lk})^2$$

Ces distances induisent la distance suivante entre les points i et l de N_I :

$$\begin{aligned} d^2(i, l) &= \sum_j d^2(i^j, l^j) \\ &= \sum_j \frac{1}{Q_j \lambda_1^j} \sum_{k \in K_j} \frac{I}{I_k} (y_{ik} - y_{lk})^2 = \sum_j \sum_{k \in K_j} \frac{1}{Q_j \lambda_1^j} \frac{I}{I_k} (y_{ik} - y_{lk})^2 \end{aligned}$$

Cette relation montre que l'AFM d'un ensemble de variables qualitatives structurées en groupes doit reposer sur une ACM pondérée dans laquelle la pondération des variables du groupe j est $1/Q_j \lambda_1^j$.

Le poids $1/Q_j \lambda_1^j$ de la variable V du groupe j provient de la succession suivante :

- considérer que la variable V appartient au groupe j auquel cas sa pondération vaut $1/Q_j$ dans l'ACM de ce groupe;
- rendre égale à 1 l'inertie axiale maximum des nuages de l'ACM du groupe j en divisant par λ_1^j (première valeur propre de cette ACM) la pondération des variables précédente.

Cette ACM pondérée fournit des représentations des individus et des modalités conformes à la problématique, en ce sens que le rôle a priori des différents groupes est équilibré. La discussion des conditions de cet équilibre – harmoniser les inerties axiales maximums plutôt que, par exemple, les inerties totales – peut être faite dans les mêmes termes que pour les variables quantitatives.

Remarques

Lorsque chaque groupe est composé d'une seule variable $\forall_j Q_j = 1$ et $\lambda_1^j = 1$; l'AFM est alors confondue avec l'ACM.

La somme des pondérations des variables ne vaut pas 1. Cette caractéristique est induite par le fait que l'on impose à chaque groupe une inertie maximum de 1. Même pour les variables qualitatives, les calculs d'une AFM doivent reposer sur une ACP et non une AFC.

4.3. AFM et analyse multicanonique

4.3.1. Mesure \mathcal{L}_g et variables qualitatives

L'AFM peut être vue comme une analyse multicanonique au sens de (Carroll 1968) à condition de mesurer la liaison entre une variable z et un groupe de variables $K_j = \{v_k; k = 1, K_j\}$ par la mesure \mathcal{L}_g , inertie projetée des variables du groupe K_j le long de z . Soit :

$$\mathcal{L}_g(z, K_j) = \sum_{k \in K_j} \text{inertie de la projection de } v_k \text{ sur } z.$$

Dans le cas où le groupe de variables j est composé des indicatrices d'une seule variable V , cette mesure s'applique et vaut (z est centré) :

$$\mathcal{L}_g(z, K_j) = \eta^2(z, V)$$

en notant $\eta^2(z, V)$ le rapport de corrélation entre z et la variable V .

Si le groupe j comprend Q_j variables qualitatives V_q cette mesure devient :

$$\mathcal{L}_g(z, K_j) = \frac{1}{\lambda_1^j} \frac{1}{Q_j} \sum_{q \in Q_j} \eta^2(z, V_q)$$

Cette mesure est proportionnelle à la moyenne des rapports de corrélation entre la variable z et les variables qui composent le groupe j . Cette moyenne est rapportée à sa valeur maximum compte tenu des données du groupe j , à savoir la première valeur propre de l'ACM de ce groupe j . En effet, en ACM, la valeur propre de rang s est égale à la moyenne des rapports de corrélation entre le facteur F_s et les Q variables qualitatives. Finalement, la mesure \mathcal{L}_g vaut 0 si la variable z a un rapport de corrélation nul avec chaque variable du groupe j ; elle vaut 1 si z coïncide avec la direction principale d'inertie du groupe j .

Ces commentaires sont analogues à ceux faits à propos des variables quantitatives. Ils montrent que \mathcal{L}_g est adaptée pour servir de mesure de liaison dans une analyse multicanonique de variables qualitatives et que, par voie de conséquence, l'AFM peut être considérée comme une telle analyse.

Remarques. Dans la dernière relation, $\mathcal{L}_g(z, K_j)$ vérifie la propriété suivante : si l'on duplique chaque variable du groupe j , l'ACM de ce groupe, et en particulier λ_1^j , est inchangée; le coefficient $1/Q_j$ permet d'obtenir la même valeur de $\mathcal{L}_g(z, K_j)$ dans les deux cas.

Le fait que l'ACM puisse être vue comme une analyse multicanonique particulière a été vu depuis longtemps par (Saporta 1975).

4.3.2. Recherche des variables générales

La variable générale de rang s (notée z_s) de cette analyse multicanonique rend maximum :

$$\sum_j \mathcal{L}_g(z_s, K_j) = \sum_j \frac{1}{\lambda_1^j} \frac{1}{Q_j} \sum_{q \in Q_j} \eta^2(z_s, V_q)$$

avec les contraintes usuelles de norme et d'orthogonalité :

$$\|z_s\| = 1 \quad \langle z_s, z_t \rangle = 0 \text{ si } s \neq t$$

4.4. Représentation des individus

Un des concepts importants de l'AFM est celui d'individu partiel, c'est-à-dire considéré du point de vue d'un groupe seulement. On a déjà noté i^j , l'individu i «vu» par le groupe j . Le principe géométrique de représentation de ces individus partiels est identique dans le cas des variables qualitatives et celui des variables quantitatives (Escofier et Pagès 1998 p.152) : dans l'espace R^K , l'individu partiel i^j , déduit de i en remplaçant par 0 ses valeurs pour les variables (centrées) des groupes autres que j , est projeté sur les axes d'inertie de N_I .

La représentation des individus partiels jouit d'une propriété importante : une relation de transition «partielle», déduite de la relation usuelle exprimant la coordonnée d'un individu en fonction de celles des variables, en se limitant à un seul groupe de variables. Nous présentons ces deux relations dans le cas des variables qualitatives.

En AFM sur variables qualitatives, la représentation de N_I vérifie les propriétés de l'ACM pondérée (cf. § 3.5.). Ainsi, un individu est (à un coefficient près) au barycentre des modalités (affectées chacune de la pondération de la variable à laquelle elle appartient) qu'il possède. Soit :

$$F_s(i) = \frac{1}{\lambda_s} \sum_j \frac{1}{\lambda_1^j Q_j} \sum_{k \in K_j} y_{ik} F_s(g_k)$$

où λ_s est la valeur propre obtenue dans l'ACP avec les pondérations $c_j = 1/\lambda_1^j$ et non c_j/c .

La représentation de l'individu partiel i^j vérifie la relation de transition précédente en restreignant la sommation aux variables du groupe j (relation de transition «partielle»). Soit :

$$F_s(i^j) = \frac{1}{\lambda_s} \frac{1}{\lambda_1^j Q_j} \sum_{k \in K_j} y_{ik} F_s(g_k)$$

A un coefficient près, i^j est au barycentre des modalités (du groupe j) qu'il possède. Cette relation permet de comparer directement les positions des points d'un même nuage N_I^j (e.g. i^j et l^j). En revanche, la comparaison de points de nuages différents (e.g. i^j et l^h) est indirecte : elle consiste à comparer les positions relatives de ces points dans leurs nuages respectifs (N_I^j et N_I^h) ce qui ne semble véritablement utile que si les facteurs étudiés sont bien des facteurs communs.

Remarque. En pratique, on dilate le nuage N_I^j par une homothétie de rapport J afin de situer l'individu i au centre de gravité de ses points partiels $\{i^j; j = 1, J\}$. La relation précédente devient :

$$F_s(i^j) = \frac{1}{\lambda_s} \frac{J}{\lambda_1^j Q_j} \sum_{k \in K_j} y_{ik} F_s(g_k)$$

4.5. Représentation des modalités partielles

Une modalité étant représenté par un centre de gravité d'individus, on peut imaginer un concept de modalité partielle calqué sur celui d'individu partiel. Dans R^K , la modalité partielle g_k^j est déduite de g_k en remplaçant par 0 ses valeurs pour les variables (centrées) des groupes autres que j . Il résulte de cette définition qu'en AFM, une modalité partielle est au centre de gravité des individus partiels qui la possèdent. Soit, en projection sur l'axe de rang s :

$$F_s(g_k^j) = \frac{1}{I_k} \sum_i y_{ik} F_s(i^j)$$

En remplaçant, dans cette relation, $F_s(i^j)$ par son expression en fonction des modalités, on obtient :

$$F_s(g_k^j) = \frac{1}{\lambda_s} \frac{1}{I_k} \frac{J}{\lambda_1^j Q_j} \sum_{l \in K_j} I_{kl} F_s(g_l)$$

Cette expression est la restriction au groupe j de la relation de transition de l'ACM pondérée en tant qu'analyse du tableau de Burt pondéré (cf. § 3.4.1). A un coefficient près, la modalité k partielle au groupe j est au barycentre des modalités l (affectés des poids I_{kl} ; $\sum_{l \in K_j} I_{kl} = I_k Q_j$) du groupe j . On retrouve, limitée à un groupe, la règle selon laquelle une modalité se trouve globalement du côté des modalités auxquelles elle s'associe le plus.

Remarque. Le concept de modalité partielle est très important dans la pratique des variables qualitatives. Dans ce cas, les individus sont généralement nombreux : on ne les considère donc pas un par un mais au travers des sous-populations induites par les modalités. C'est le concept de « modalité partielle » qui justifie de représenter en AFM les modalités par les barycentres d'individus et non par les indicatrices (ou, ce qui revient au même, par les barycentres dilatés axe par axe par le coefficient $1/\sqrt{\lambda_s}$) comme cela est fait usuellement en ACM (tout au moins dans les logiciels français tels SPAD et LADDAD).

4.6. Analyse dans R^{I^2}

4.6.1. Nuage des groupes de variables

Dans R^{I^2} , le cas des variables qualitatives ne se distingue que très peu de celui des variables quantitatives. Chaque groupe de variables K_j est représenté par $W_j D' = X_j M_j X_j' D$. L'ensemble de ces points constitue le nuage des groupes de variables, noté N_J .

Le nuage N_J est projeté sur le sous-espace induit par les composantes principales z_s . La coordonnées du groupe j le long de z_s est égale à la mesure de liaison $\mathcal{L}_j(z_s, K_j)$ qui est aussi la contribution du groupe j à l'inertie de l'axe de rang s

(cf. § 4.3.2). A la différence des groupes de variables quantitatives, cette mesure est fondée sur le rapport de corrélation et non le coefficient de corrélation.

4.6.2. Interprétation du produit scalaire entre deux groupes

Rappelons que, dans cet espace, le produit scalaire entre deux groupes de variables quantitatives K_1 et K_2 s'interprète comme une mesure de liaison (Escoufier 1973) : COVV dans le cas général, RV si les groupes sont normés et, lorsque les inerties axiales maximum des nuages associés aux groupes sont égales à 1, $\mathcal{L}_g(K_1, K_2)$ (Escoufier et Pagès 1998 p.166). Cette propriété s'étend aux variables qualitatives (Cailliez et Pagès 1976), (Cazes *et al.* 1976).

- Si K_1 et K_2 sont composés chacun d'une seule variable qualitative (respectivement V_1 et V_2) :

$$\langle K_1, K_2 \rangle = \mathcal{L}_g(K_1, K_2) = \frac{1}{I} \chi_{V_1 V_2}^2$$

Dans ce cas particulier on retrouve la mesure de liaison usuelle.

- Si K_1 est composé d'une seule variable V_1 et K_2 de Q_2 variables notées V_q :

$$\langle K_1, K_2 \rangle = \mathcal{L}_g(K_1, K_2) = \sum_{q \in Q_2} \frac{1}{\lambda_1^2 Q_2} \mathcal{L}_g(V_1, V_q) = \frac{1}{I} \frac{1}{\lambda_1^2 Q_2} \sum_{q \in Q_2} \chi_{V_1, V_q}^2$$

Les deux groupes sont d'autant plus liés que la variable du groupe 1 est liée, au sens du χ^2 , à chacune des variables du groupe 2. La pondération par λ_1^2 (1^e valeur propre de l'ACM du groupe 2) appelle des remarques analogues à celles faites dans le cas de variables quantitatives. La liaison entre V_1 et K_2 est d'autant plus grande que les indicatrices de V_1 sont proches des principales directions d'inertie de K_2 . Remarquons que, si l'on duplique toutes les variables du groupe K_2 , λ_1^2 est inchangée et c'est le coefficient Q_2 qui assure l'invariance de la mesure de liaison.

- Si K_1 et K_2 comportent respectivement Q_1 et Q_2 variables (notées V_h pour le groupe K_1 et V_q pour le groupe K_2) :

$$\langle K_1 K_2 \rangle = \mathcal{L}_g(K_1, K_2) = \frac{1}{I} \frac{1}{\lambda_1^1 Q_1} \frac{1}{\lambda_1^2 Q_2} \sum_{h \in Q_1} \sum_{q \in Q_2} \chi_{V_h V_q}^2$$

Les deux groupes sont d'autant plus liés que chaque variable du groupe 1 est liée, au sens du χ^2 , à chacune des variables du groupe 2. La pondération par $1/(\lambda_1^1 \lambda_1^2 Q_1 Q_2)$ appelle des remarques analogues à celles faites dans le cas précédent.

5. Cas des données mixtes

5.1. Notations

Les J groupes de variables se décomposent en J_1 groupes de variables quantitatives et J_2 groupes de variables qualitatives. Les données constituent le tableau X dans lequel les variables qualitatives apparaissent sous la forme disjonctive complète (x_{ik} est une valeur numérique quelconque si la colonne k représente une variable quantitative et vaut 0 ou 1 si la colonne k représente une indicatrice).

5.2. Principe

En pratique, les données mixtes – mélange de variables quantitatives et qualitatives – sont fréquentes. Le traitement simultané par analyse factorielle de ces deux types de variables est classique lorsque les variables actives sont de même type : ainsi, on introduit des variables supplémentaires qualitatives en ACP (en représentant les centres de gravité des individus présentant une même modalité) et des variables supplémentaires quantitatives en ACM (à l'aide des coefficients de corrélation entre ces variables et les facteurs). Plusieurs méthodes ont été proposées pour introduire simultanément en actif les deux types de variables (Tenenhaus 1977), (Escofier 1979b), (Saporta 1990) mais elles ne prennent pas en compte la notion de groupes de variables. Or la possibilité de prendre en compte des variables des deux types structurées en groupes est souhaitée dans de nombreuses applications.

L'AFM offre une telle possibilité à la condition que les variables appartenant à un même groupe soient de même type.

Une telle analyse, qui traite un mélange des deux types de variables, ne pose pas de problème de principe : on peut souhaiter définir une distance entre individus qui prenne en compte à la fois les deux types de variables. Cette distance doit être celle de l'ACP si l'on se restreint à un groupe de variables quantitatives et celle de l'ACM si l'on se restreint à un groupe de variables qualitatives. Que l'on considère un tableau de mesures ou un TDC, ces deux restrictions ont la même forme analytique :

$$d^2(i, l) = \sum_k m_k (x_{ik} - x_{lk})^2$$

Seuls les coefficient m_k changent entre les deux types de variables.

Le problème de l'agrégation des deux types de variables au sein d'une seule formule de distance consiste essentiellement à choisir une pondération des variables qui équilibre leur influence a priori. Ainsi, dès lors que les variables d'un même groupe sont de même type, la problématique associée à cette pondération se confond avec celle des groupes de variables. En ce sens, l'AFM permet de traiter des données mixtes.

5.3. Pondération des variables

Le tableau de données juxtapose variables quantitatives et indicatrices. L'analyse doit prendre en compte simultanément ces deux types de données et fonctionner «localement» comme une ACP pour les variables quantitatives et comme une ACM pour les variables qualitatives. Pratiquement, ce résultat peut être obtenu directement en utilisant l'équivalence ACP-ACM (cf. § 2.3) : les indicatrices peuvent être traitées comme les variables quantitatives à condition de les pondérer de façon adéquate.

L'équilibre entre les groupes est assuré par la pondération de l'AFM. On divise le poids initial de la variable k du groupe j par :

- λ_1^j si le groupe j est quantitatif (λ_1^j : première valeur propre de l'ACP du groupe j);
- $(\lambda_1^j Q_j)$ si le groupe j est qualitatif (λ_1^j : première valeur propre de l'ACM du groupe j comportant Q_j variables; cf. § 4.2.2).

Le même résultat peut être obtenu en remplaçant les données brutes par les facteurs des analyses factorielles séparées des groupes (ACP ou ACM selon les cas). Ces facteurs seront soit non-normés, soit normés et affectés d'une pondération initiale égale à la valeur propre qui leur est associée. Cette propriété, couramment utilisée dans le cas quantitatif cf. (Escofier et Pagès 1998 p. 158), permet de calculer des contributions (au plein sens du terme) à l'inertie des axes aussi bien pour les variables initiales que pour les facteurs des analyses séparées.

5.4. Propriétés

5.4.1. Représentation des variables

Les variables quantitatives sont représentées par leurs coefficients de corrélation avec les facteurs. Les variables qualitatives n'apparaissent qu'au travers de leurs modalités, représentées par le centre de gravité des individus qui les possèdent.

Les modalités partielles sont définies comme dans le cas qualitatif pur et vérifient la même relation de transition restreinte présentée en 4.5.

5.4.2. Représentation des nuages d'individus partiels

L'égalisation des inerties maximum de ces nuages permet leur représentation simultanée. Cette représentation bénéficie des relations de transition partielles identiques à celles de l'AFM appliquée à des groupes de même type :

- un individu partiel au groupe qualitatif j se trouve, à un coefficient près, au barycentre des modalités du groupe j qu'il possède; soit :

$$F_s(i^j) = \frac{1}{\lambda_s} \frac{J}{\lambda_1^j Q_j} \sum_{k \in K_j} x_{ik} F_s(g_k) = \frac{1}{\sqrt{\lambda_s}} \frac{J}{\lambda_1^j Q_j} \sum_{k \in K_j} x_{ik} G_s(k)$$

- un individu partiel au groupe quantitatif j se trouve du côté des variables du groupe j pour lesquelles il présente de fortes valeurs et à l'opposé de celles du

groupe j pour lesquelles il présente de faibles valeurs; soit :

$$F_s(i^j) = \frac{1}{\sqrt{\lambda_s}} \frac{J}{\lambda_1^j} \sum_{k \in K_j} x_{ik} G_s(k)$$

Le coefficient J assure la propriété suivante (cf. aussi la remarque en fin du § 4.4) : l'individu i est à l'isobarycentre de ses points partiels i^j quel que soit leur type. Soit :

$$F_s(i) = \frac{1}{J} \sum_{j \in J} F_s(i^j)$$

5.4.3. Analyse multicanonique

La quantité rendue maximum par les facteurs de l'AFM peut s'exprimer en termes d'analyse canonique, ces facteurs étant les variables générales. On se limite à deux groupes, K_1 (groupe quantitatif comportant K_1 variables v_k) et K_2 (groupe qualitatif comportant Q_2 variables V_q).

La variable générale de rang $s(z_s)$ rend maximum :

$$\mathcal{L}_g(z_s, K_1) + \mathcal{L}_g(z_s, K_2) = \frac{1}{\lambda_1} \sum_{k \in K_1} r^2(z_s, v_k) + \frac{1}{\lambda_1^2 Q_2} \sum_{q \in Q_2} \eta^2(z_s, V_q)$$

avec les contraintes usuelles de norme et d'orthogonalité.

Cette quantité fait jouer un rôle analogue au carré du coefficient de corrélation et au rapport de corrélation. Elle est grande si z_s est fortement corrélée avec chacune des variables du groupe K_1 et présente un fort rapport de corrélation avec chacune des variables du groupe K_2 . En ce sens z_s est liée à la fois à K_1 et à K_2 . La pondération des variables limite à 1 la contribution maximum d'un groupe à cette quantité. Cette quantité vaut au maximum 2 (J dans le cas général), valeur atteinte lorsque z_s coïncide à la fois avec le premier facteur de l'ACP de K_1 et avec celui de l'ACM de K_2 .

5.4.4. Analyse dans R^{I^2}

Dans cet espace, le groupe j étant représenté essentiellement par la matrice W_j des produits scalaires (entre individus) qu'il induit, la présence de variables de natures différentes ne pose pas de problème particulier, ces matrices étant homogènes. La projection du groupe K_j , sur l'axe induit par la variable générale z_s de l'AFM, s'interprète comme la mesure de liaison $\mathcal{L}_g(z_s, K_j)$, déjà commentée en 5.4.3.

Dans cet espace, le produit scalaire entre K_1 et K_2 (à condition que les inerties axiales maximum des nuages associés aux groupes soient égales à 1) vaut $\mathcal{L}_g(K_1, K_2)$. Selon la nature des variables, ce produit scalaire dérive du coefficient de corrélation, du rapport de corrélation ou du χ^2 . Le tableau 3 explicite $\mathcal{L}_g(K_1, K_2)$ en fonction du nombre et de la nature des variables composant les groupes.

Dans le cas d'un groupe quantitatif et d'un groupe qualitatif, $\mathcal{L}_g(K_1, K_2)$ vaut 0 si chaque variable d'un groupe a un rapport de corrélation nul avec chaque variable de l'autre groupe. Elle n'est pas bornée supérieurement : elle est d'autant plus grande que les deux groupes présentent un grand nombre de directions communes d'inertie importante. Cette dernière propriété vaut pour tous types de groupes et est discutée dans le cas de deux groupes quantitatifs dans (Pagès 1996).

TABLEAU 3
Mesure de liaison $\mathcal{L}_g(K_1, K_2)$ entre deux groupes de variables selon leur nature (quantitatif / qualitatif) et le nombre de variables qu'ils comportent (une / plusieurs)

$r^2(z, v)$: coefficient de corrélation entre les variables quantitatives z et v ;

$\eta^2(z, V)$: rapport de corrélation entre la variable quantitative z et la variable qualitative V ;

χ_{zv}^2 : statistique du χ^2 entre les variables qualitatives Z et V

		Groupe K_2				
		quantitatif		qualitatif		
		v	$v_k, k=1, K_2$	V	$V_k, k=1, Q_2$	
Groupe K_1	quantitatif	z	$r^2(z, v)$	$\frac{1}{\lambda_1^2} \sum_{k \in K_2} r^2(z, v_k)$	$\eta^2(z, V)$	$\frac{1}{\lambda_1^2 \lambda_2^2} \sum_{k \in Q_2} \eta^2(z, V_k)$
	$z_l, l=1, K_1$		$\frac{1}{\lambda_1^2 \lambda_2^2} \sum_{l \in K_1} \sum_{k \in K_2} r^2(z_l, v_k)$	$\frac{1}{\lambda_1^2} \sum_{l \in K_1} \eta^2(z_l, V)$	$\frac{1}{\lambda_1^2 \lambda_2^2} \sum_{l \in K_1} \sum_{k \in Q_2} \eta^2(z_l, V_k)$	
qualitatif	Z			$= \frac{1}{I} \chi_{ZV}^2$	$\frac{1}{I} \frac{1}{\lambda_1^2 \lambda_2^2} \sum_{k \in Q_2} \chi_{ZV_k}^2$	
	$Z_l, l=1, Q_1$				$\frac{1}{I} \frac{1}{\lambda_1^2 \lambda_2^2} \sum_{l \in Q_1} \sum_{k \in Q_2} \chi_{Z_l V_k}^2$	

6. Application

On trouvera dans (Goldfarb et Pardoux 2001) un exemple d'AFM sur données mixtes provenant de données médicales et dans (Escofier et Pagès 1998 p. 143) un exemple sur données qualitatives issues d'enquêtes. Il est ainsi possible de juger l'intérêt de cette méthodologie dans l'analyse de problèmes concrets. Pour illustrer de façon minutieuse les propriétés de cette méthodologie, un exemple restreint de données choisies pour mettre en valeur ces propriétés est commode. Les résultats présentés peuvent être obtenu à l'aide des logiciels LADDAD (1997) et SPAD (2001).

6.1. Données, objectifs

La méthodologie proposée est appliquée à un petit ensemble de données construites de la façon suivante (cf. tab. 4) :

- Les individus sont décrits par 3 variables quantitatives dont deux sont fortement corrélées entre elles ($r(\text{long.}, \text{larg.}) = -0.71$) et peu corrélées à la troisième ($r(\text{long.}, \text{poids}) = -0.39$; $r(\text{larg.}, \text{poids}) = -0.13$).
- Ces trois variables sont codées en variables qualitatives en subdivisant leur plage de variation en 2 ou 3 classes.

TABLEAU 4

Données mixtes choisies

Six individus (A,...F) sont décrits par 3 variables quantitatives.

Ces données sont soumises à l'AFM au travers de deux transformations :

le centrage-réduction et un codage en (2 ou 3) classes

	Données brutes			Données centrées-réduites			Données codées en classes		
	Longueur	Poids	Largeur	Longueur	Poids	Largeur	Long. en 3 classes	Poids en 2 classes	Larg. en 3 classes
A	1	1	6	-1,464	0,447	1,464	1	2	3
B	2	0	5	-0,878	-0,894	0,878	1	1	3
C	3	2	3	-0,293	1,789	-0,293	2	2	2
D	5	0	4	0,878	-0,894	0,293	3	1	2
E	4	1	1	0,293	0,447	-1,464	2	2	1
F	6	0	2	1,464	-0,894	-0,878	3	1	1
moyenne	3,500	0,667	3,500	0	0	0			
ec.-type	1,708	0,745	1,708	1	1	1			

L'objet spécifique de cette application est de comparer deux images d'un même ensemble de données : celle fournie par les données centrées-réduites et celle fournie par un recodage en classes. Le point de vue «analyse canonique» de l'AFM, selon lequel la méthode met en évidence les facteurs communs aux tableaux et ceux qui sont spécifiques de l'un d'entre eux, correspond bien à cette comparaison. Le point de vue «analyse des facteurs des analyses séparées» permet une comparaison des méthodologies usuelles : l'ACP normée sur les données brutes; l'ACM sur données recodées en classes.

L'objet plus général de cette application est d'illustrer comment l'AFM prend en compte simultanément les deux types de variables.

Remarquons que le nombre d'individu est très faible, en particulier dans la perspective d'une ACM. Une telle analyse n'aurait a priori pas grand intérêt avec des données réelles; mais ces données ont été choisies, i.e. ici construite selon une structure simple et claire qui doit logiquement apparaître au travers des deux codages.

6.2. Analyses séparées

6.2.1. Inerties (tableau 5)

Les inerties projetées sont toujours inférieures à 1 en ACM; en ACP, la première valeur propre est toujours supérieure à 1. La pondération des variables est indispensable dans l'analyse simultanée de ces deux types de données.

La décroissance des inerties est plus lente pour le 1^{er} groupe (ACM). Cette observation est classique.

TABLEAU 5
Valeurs propres des analyses séparées

groupe	valeurs propres				pourcentages d'inertie			
	1	2	3	4	1	2	3	4
1 qualitatif (ACM)	0.667	0.605	0.333	0.061	40.0	36.3	20.0	3.7
2 quantitatif (ACP)	1.765	1.110	0.125		58.8	37.0	4.2	

Le premier groupe (qualitatif) est tridimensionnel, avec deux premières valeurs propres assez voisines. Le second groupe (quantitatif) est bidimensionnel, ce qui est en accord avec la façon dont il a été construit; les deux premières valeurs propres sont assez différentes, ce qui rend nécessaire une pondération.

6.2.2. Corrélations entre les facteurs des analyses séparées (tableau 6)

TABLEAU 6
Corrélations entre facteurs des analyses séparées
Exemple : .43 est le coefficient de corrélation entre le 2^e facteur de l'ACP
du groupe 2 et le 1^{er} facteur de l'ACM du groupe 1

		Groupe 1 (ACM)		
		F1	F2	F3
Groupe 2 (ACP)	F1	-0.82	-0.37	-0.41
	F2	0.43	-0.87	-0.04
	F3	0.33	0.27	-0.68

En se limitant aux 2 premiers, les facteurs de même rang sont assez fortement corrélés (.82; .87) : les plans $F_1 \times F_2$ issus d'une part de l'ACP et d'autre part de l'ACM sont donc apparentés mais loin d'être confondus.

6.3. Inerties de l'analyse globale

La séquence des valeurs propres (tableau 7) suggère d'interpréter 3 axes.

La première valeur propre (1.92) est proche de son maximum (le nombre de groupes). Le premier axe de l'AFM correspond à une direction d'inertie très importante dans chaque groupe. Les deux groupes contribuent également à ce premier axe, propriété attendue (et observée en pratique mais on peut imaginer des contre-exemples) suite à la pondération de l'AFM.

TABLEAU 7
Inertie des quatre premiers facteurs de l'AFM

	F 1	F 2	F 3	F 4	F 5
inertie totale	1.92	1.53	.61	.12	.02
groupe 1	.95	.91	.54	.10	.01
groupe 2	.97	.62	.07	.02	.02

Le deuxième axe correspond à une direction d'inertie importante pour le groupe 1 et sensiblement plus faible pour le groupe 2. Ceci également était attendu vu les deux premières valeurs propres des analyses séparées. Le troisième axe est spécifique du groupe 1.

Ceci est en accord avec le mode de construction des données.

On illustre ici la difficulté majeure inhérente à l'analyse factorielle simultanée de variables quantitatives et qualitatives : la pondération de l'AFM normalise les inerties axiales maximum des groupes; mais, du fait de la décroissance des inerties en général plus faible en ACM qu'en ACP, les groupes qualitatifs sont a priori plus influents, en terme de contribution à l'inertie, à partir du 2^e axe.

6.4. Coordonnées des facteurs des analyses séparées (figure 1)

Les deux premiers facteurs de l'AFM sont assez proches des facteurs de rang homologue des analyses séparées dont ils constituent une sorte de compromis. On remarque qu'ils sont un peu plus corrélés aux facteurs du groupe quantitatif. Il s'agit vraisemblablement d'une conséquence de la faible décroissance des inerties du groupe qualitatif qui avantage les facteurs du groupe quantitatif en ce sens qu'ils « arbitrent » entre des directions du groupe qualitatif ayant des inerties comparables.

Cette remarque atténue l'impact de la réserve émise en 6.3 quant au traitement simultané de variables des deux types. Quoi qu'il en soit, ceci souligne la nécessité de prendre en compte les inerties des analyses séparées. Ceci doit toujours être fait en AFM; cette nécessité est exacerbée dans le cas de groupes actifs de types différents.

Conclusion : la pondération de l'AFM a très bien fonctionné dans cet exemple en équilibrant les influences des deux types de variables.

6.5. Premier facteur

6.5.1. Individus et variables

Au niveau des variables quantitatives, des modalités et des individus, les règles d'interprétation sont celles de l'ACP et de l'ACM.

Le premier facteur :

- est corrélé négativement à *longueur* et positivement à *largeur* (quantitatives) (figure 1);
- classe les individus par ordre alphabétique (inverse) de façon à peu près régulière (figure 3);
- classe les modalités des variables qualitatives *longueur* et *largeur* dans leur ordre naturel et en accord avec les corrélations précédentes (figure 2).

Ce facteur oppose l'individu *A*, large et court, à *F*, long et étroit. Il s'agit bien d'un facteur commun, cette opposition étant claire dans les données avec les deux types de variables.

6.5.2 Individus partiels (figure 3)

De façon marginale, quelques différences entre les individus partiels des deux groupes apparaissent. Ainsi, les individus A et B sont plus différents du point de vue du groupe quantitatif (groupe 2) que qualitatif (groupe 1) : $|F_1(A^1) - F_1(B^1)| < |F_1(A^2) - F_1(B^2)|$. On retrouve ici une conséquence du codage en classes qui peut regrouper des individus (légèrement) différents : A et B diffèrent (peu) du point de vue de la longueur et de la largeur mais ceci est effacé dans le regroupement en classes (seule leur différence de poids, variable très légèrement liée au premier facteur, empêche A^1 et B^1 d'être confondus sur cet axe).

Les individus D et E attirent l'attention. Du point de vue de l'axe 1, E est plus caractéristique que D si l'on considère les données quantitatives : $|F_1(E^2)| > |F_1(D^2)|$; en effet, D est un peu plus long que large alors que E est beaucoup plus long que large; ceci est visible dans les données brutes, ces deux variables ayant le même écart-type.

En revanche, toujours du point de vue de F_1 , D est plus caractéristique que E si l'on considère les données recodées en classes : $|F_1(D^1)| > |F_1(E^1)|$. Remarquons d'abord que le codage en classes a gommé la différence précitée en ce sens que l'écart entre *longueur* et *largeur* n'est plus que d'une classe pour les deux individus. De ce point de vue, ces deux individus sont identiques, ce qui apparaît dans la proximité entre le barycentre (non représenté) de $\{\text{longueur} = 3; \text{largeur} = 2\}$ pour l'individu D et celui de $\{\text{longueur} = 2; \text{largeur} = 1\}$ pour l'individu E ; la différence entre les points partiels tient au *poids*, légèrement lié au premier facteur, et opposant franchement (surtout dans les données codées en classes) D et E .

Pour des raisons analogues, les individus C , D , E et F sont mieux séparés par le groupe 2 que par le groupe 1.

6.5.3. Relations de transition

Les relations de transition exprimant la coordonnée d'un individu en fonction de celles des variables quantitatives d'une part et des modalités d'autre part (§ 5.4.2) sous-tendent l'interprétation de la projection du nuage des individus et donc les commentaires du paragraphe précédent.

TABLEAU 8
Données mixtes choisies : contributions des valeurs du tableau cases aux coordonnées des individus sur le premier facteur

	Groupe 2, quantitatif				Groupe 1, qualitatif								
	Long	Poids	Larg.	Ind. Part G2	Long.=1	Long.=2	Long.=3	Poids=1	Poids=2	Larg.=1	Larg.=2	Larg.=3	Ind Part G1
A	1,165	0,080	1,016	2,261	0,927	0,000	0,000	0,000	0,225	0,000	0,000	0,927	2,079
B	0,699	-0,160	0,610	1,148	0,927	0,000	0,000	-0,225	0,000	0,000	0,000	0,927	1,630
C	0,233	0,320	-0,203	0,350	0,000	-0,228	0,000	0,000	0,225	0,000	-0,217	0,000	-0,221
D	-0,699	-0,160	0,203	-0,656	0,000	0,000	-0,699	-0,225	0,000	0,000	-0,217	0,000	-1,140
E	-0,233	0,080	-1,016	-1,169	0,000	-0,228	0,000	0,000	0,225	-0,710	0,000	0,000	-0,714
F	-1,165	-0,160	-0,610	-1,935	0,000	0,000	-0,699	-0,225	0,000	-0,710	0,000	0,000	-1,633

Le tableau 8 rassemble, pour le facteur 1, les termes, à droite dans les relations du § 5.4.2, dont la somme est la coordonnée d'un individu (à un coefficient près, ces termes résultent du produit d'une valeur du tableau par la coordonnée d'une

variable ou d'une modalité) : en ce sens il s'agit des contributions des variables aux coordonnées des individus. Entre les deux groupes, ces contributions aux coordonnées sont comparables, condition nécessaire pour une prise en compte simultanée des deux groupes de variables.

Ainsi :

- La variable qualitative *longueur* ne distingue pas *A* et *B* et, par rapport aux données quantitatives, à la fois fait apparaître (le long de F_1) *A* moins caractéristique ($.927 < 1.165$) et *B* plus caractéristique ($.927 > .699$);
- La variable *poids* est beaucoup moins liée à F_1 que la variable *longueur* ce qui se traduit par des contributions plus faibles et ce dans les deux codages; ces contributions sont un peu plus élevées dans le cas qualitatif, signe d'une plus forte liaison entre F_1 et *poids* dans le codage qualitatif ($\eta^2(F_1, \text{poids q.t.}) = .097$) que dans les données quantitatives ($r^2(F_1, \text{poids quant.}) = .048$).

6.5.4. Modalités partielles (figure 4)

Globalement, les modalités partielles homologues (i.e. relatives à une même modalité) sont plus proches entre elles que ne le sont les individus partiels homologues. Ce résultat est mécanique : parmi les individus présentant une même modalité, des écarts de sens contraires entre images partielles s'annulent dans le calcul des barycentres. Exemple : *largeur* = 3 (partagée par *A* et *B*) le long de F_1 . Contre-exemple : *largeur* = 1, le long de F_1 (partagée par *E* et *F*), tous deux plus extrêmes du point de vue de F_1 .

6.5.5. Conclusions

La corrélation négative entre *longueur* et *largeur*, principale structure des données en ce sens qu'elle concerne deux variables sur trois, apparaît clairement au travers des deux codages. La perte d'information due au codage en variable qualitative se traduit par une allure du nuage des individus légèrement différente, en particulier :

- un étalement régulier des individus pour les données brutes;
- un étalement suggérant deux classes ($\{A, B\}$ et $\{C, D, E, F\}$) dans l'autre.

La représentation superposée de l'AFM, classique dans le cas quantitatif, fonctionne bien dans les cas qualitatifs et mixtes. Son interprétation repose essentiellement sur les relations de transition partielles, presque identiques à celles de l'ACP et de l'ACM.

6.6. Deuxième facteur

Ce facteur :

- oppose essentiellement les individus *E* et *C* (contribution : 59,8%) aux autres;
- est corrélé positivement à la variable quantitative *poids* ($R^2 = .88$);
- est lié à la variable *poids* qualitative ($\eta^2 = .80$) et *longueur* qualitative ($\eta^2 = .902$), en notant η^2 le rapport de corrélation.

Selon ce facteur, les individus E et C sont donc caractérisés par un *poids* élevé et une *longueur* intermédiaire (ce que l'on retrouve dans les données brutes).

Ce deuxième facteur est commun aux deux groupes (schématiquement il oppose E et C aux autres) mais n'a pas exactement la même signification dans les deux cas. Il est lié aux deux variables *poids* mais sa signification est enrichie, pour le groupe 1, par l'opposition entre *longueurs extrêmes* et *longueur intermédiaire*.

On retrouve ici l'aptitude de l'ACM à mettre en évidence des liaisons non linéaires. Le point de vue quantitatif indique seulement que le 2^e facteur est non corrélé avec la longueur et la largeur. Le point de vue qualitatif montre une absence de liaison avec la largeur mais une liaison non linéaire avec longueur.

Du point de vue de la représentation superposée des nuages partiels, on remarque que la dispersion des individus partiels est plus grande pour le groupe 1 que pour le groupe 2. Ceci est une conséquence de la décroissance plus faible des inerties de l'analyse séparée du groupe 1; parallèlement, la liaison entre le facteur F_2 de l'AFM et le groupe 1 est plus grande qu'avec le groupe 2 : $\mathcal{L}_y(F_2, K_1) = .91$; $\mathcal{L}_y(F_2, K_2) = .62$.

Il en résulte que F_2 sépare plus fortement les modalités partielles du groupe 1.

6.7. Troisième facteur

Il est spécifique du groupe 1 et étroitement corrélé avec le 3^e facteur de l'ACM de ce groupe ($r = .95$). Étant tridimensionnel, ce groupe ne pouvait être parfaitement représenté sur le premier plan de l'AFM.

En ACM, chaque variable engendre, dans R^I , un sous-espace de dimension égal à : (nombre de modalités - 1). Le premier plan de l'AFM prend en compte, pour le groupe 1 : les deux dimensions de *longueur*, la dimension de *poids*, une dimension de *largeur*. «Mécaniquement», le 3^e facteur de l'AFM doit exprimer la dimension «restante» de *largeur*.

De façon directe (figure 5), ce 3^e facteur oppose les modalités 1 et 2 de *largeur* (ces modalités sont proches sur le premier plan, en particulier du point de vue du groupe 1 ce qui correspond bien à la façon identique dont elles s'associent avec les autres modalités) et les individus correspondants.

6.8. Représentation des groupes de variables

La représentation graphique des groupes de variables présente surtout de l'intérêt lorsque le nombre de groupes est élevé, ce qui n'est pas le cas de cet exemple. Toutefois, cette représentation peut être complétée par celles des variables en tant que groupes, ce que l'on obtient en pratique en introduisant dans le logiciel chaque variable comme un groupe supplémentaire (la coordonnée d'un groupe d'une variable s'interprète comme un coefficient de corrélation au carré ou un rapport de corrélation). Deux représentations de variables qualitatives quasiment analogues à celle-ci ont déjà été proposées dans le cadre de l'ACM. (Escofier 1979a) effectuée la même démarche mais norme au préalable les opérateurs $W_j D$ associés chacun à

une variable qualitative : ainsi la coordonnée de la variable q (ayant n_q modalités) le long de l'axe de rang s est égale au rapport de corrélation (entre q et F_s) au coefficient $\sqrt{n_q - 1}$ près. (Cazes 1982) prend comme coordonnées la racine de ces rapports de corrélation. L'utilisation du rapport de corrélation lui-même s'inscrit plus naturellement dans le cadre de l'AFM.

La figure 6 fournit une visualisation synthétique des grands traits de l'interprétation déjà faite :

- le premier facteur de l'AFM correspond à une direction d'inertie importante des deux groupes de variables; il est très liée à 2 variables (*longueur* et *largeur*) sur 3 de chacun d'eux;
- le deuxième facteur de l'AFM correspond à une direction d'inertie importante du premier groupe (il est lié à deux variables de ce groupe : *poids* et *longueur*) et d'importance moindre du second groupe (il est lié seulement au *poids*).

TABLEAU 9
Qualité de représentation des groupes sur les 3 premiers axes

	F 1	F 2	F 3
Groupe 1	0.43	0.40	0.14
Groupe 2	0.68	0.28	0.00
Ensemble	0.53	0.35	0.08
Longueur qualitative	0.40	0.41	0.00
Poids qualitative	0.01	0.64	0.00
Largeur qualitative	0.41	0.01	0.48
Longueur quantitative	0.90	0.00	0.00
Poids quantitative	0.00	0.78	0.00
Largeur quantitative	0.52	0.03	0.01

Les groupes de variables n'étant pas normés, leurs qualités de représentation (au sens inertie projetée/inertie totale) n'apparaissent pas sur le graphique et doivent être consultées à part (cf. tableau 9). Ainsi, sur le premier plan :

- l'éloignement entre les deux codages de *longueur* (bien représentés) correspond bien à l'unidimensionalité de l'un et la bidimensionalité de l'autre;
- la proximité entre les deux *largeurs* (mal représentées) ne correspond pas à une équivalence globale entre les deux codages, impossible du fait des dimensions différentes, la différence entre les deux codages s'exprime sur l'axe 3;
- la proximité entre les deux *poids* (assez bien représentés) suggère une équivalence entre les deux codages, possible du point de vue des dimensions et non contredite par les autres axes.

L'influence du codage telle qu'elle apparaît dans cette AFM est à rapprocher des indicateurs de liaisons \mathcal{L}_q et RV calculés entre deux codages d'une même variable (cf. tableau 10).

L'indicateur \mathcal{L}_g indique une plus faible structure commune entre les deux codages pour la variable poids; la variable qualitative permet de mieux reconstituer les données brutes dans le cas de la *longueur* et de la *largeur*.

L'indicateur RV , qui relativise l'importance de la structure commune par le nombre de modalités de la variable qualitative, indique que les images des individus données par les deux codages sont plus proches dans le cas de *poids*.

TABLEAU 10

Données mixtes choisies. Indicateur de liaison \mathcal{L}_g et RV entre les deux codages d'une même variable. Les groupes étant réduits à une variable quantitative d'une part et à une variable qualitative d'autre part, \mathcal{L}_g s'interprète comme le rapport de corrélation. La variable qualitative poids ayant deux modalités, \mathcal{L}_g s'interprète dans ce cas aussi comme le carré du coefficient de corrélation entre la variable quantitative et l'une des deux indicatrices.

Variable	Longueur	Poids	Largeur
\mathcal{L}_g	.914	.800	.914
RV	.646	.800	.646

6.9. Conclusion

Cet exemple montre que la pondération des variables incluse dans l'AFM rend possible l'analyse simultanée de groupes de variables de types différents. Les interprétations empruntent à la fois à l'ACP et à l'ACM. Elles ne posent pas de problèmes particuliers hormis la nécessité (classique en AFM mais exacerbée dans ce type d'application) de garder à l'esprit les inerties des analyses séparées.

L'intérêt de la méthodologie d'étude de variables quantitatives qui consiste à les coder en qualitatives et à en faire l'ACM est confirmé (si besoin était). Les liaisons linéaires décelées par l'ACP le sont aussi par l'ACM. Cette dernière met aussi en évidence d'autres liaisons. Dans cet exemple où les liaisons sont essentiellement linéaires, les premiers facteurs de l'ACP et de l'ACM sont apparentés mais ce n'est bien sûr pas toujours le cas.

L'AFM s'est avérée ici un outil riche et commode pour comparer deux méthodologies. On retrouve ici l'intérêt de cette méthode lorsque les groupes ont beaucoup en commun.

Une autre application est suggérée par cet exemple : l'AFM permet de ne pas choisir entre deux (ou plus) codages et de réaliser une analyse les prenant tous les deux en compte. Cette approche sera précieuse en particulier pour construire une classification des individus « robuste » vis-à-vis du codage.

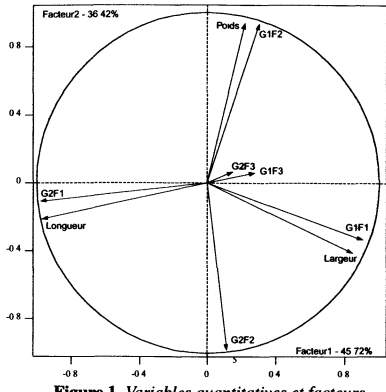


Figure 1. Variables quantitatives et facteurs des analyses séparées
G1F2 : 2^e facteur de l'analyse séparée du groupe1

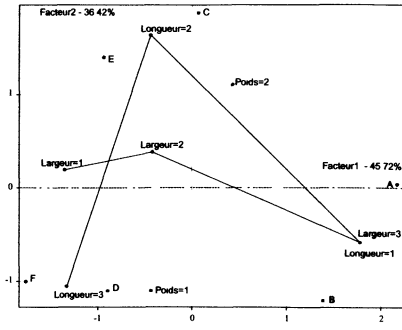


Figure 2. Individus et modalités

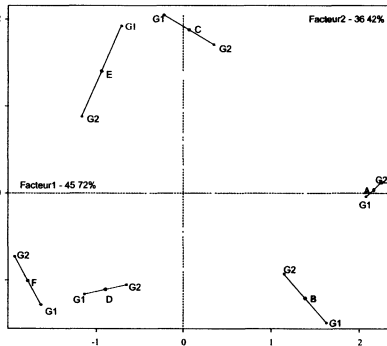


Figure 3. Individus, moyens et partiels

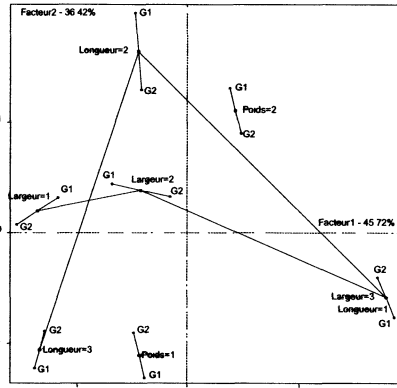


Figure 4. Modalités, moyennes et partielles

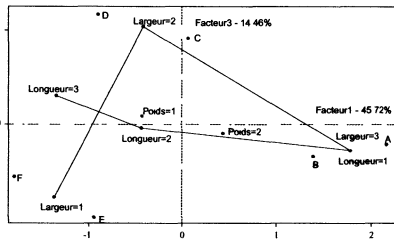


Figure 5. Individus et modalités sur le plan F1, F3.

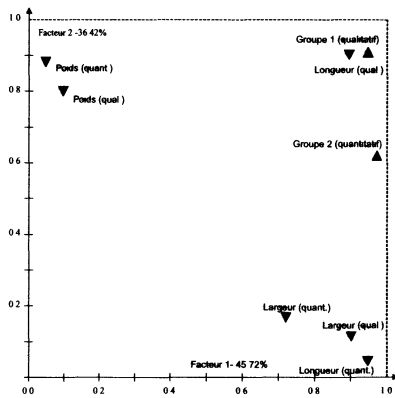


Figure 6. Groupes de variables

Données mixtes choisies : représentations issues de l'AFM.

Références

- BENZECRI J.-P. (1973), *L'analyse des données*. Dunod, Paris.
- CARROLL J.-D. (1968), A generalisation of canonical correlation analysis to three or more set of variables. In Proceedings of the 76th Annual convention of the American *Psychological Association*, 227-228.
- CAILLIEZ F. et PAGES J.-P. (1976), Introduction à l'analyse des données. *SMASH*, Paris.
- CAZES P., BAUMERDER A., BONNEFOUS S. et PAGÈS J.-P. (1977), Codage et analyse de tableaux logiques; introduction à la pratique des variables qualitatives. *Cahiers du B.U.R.O.*, (27) 3-47.
- CAZES P., BONNEFOUS S., BAUMERDER A. et PAGES J.-P. (1976), Description cohérente des variables qualitatives prises globalement et de leurs modalités. *Statistique et analyse des données*, 2-3, 48-62.
- CAZES P. (1980), L'analyse de certains tableaux rectangulaires décomposés en blocs. *Les cahiers de l'Analyse des données*, 5 (4) 387-403.
- CAZES P. (1982), Note sur les éléments supplémentaires en analyse des correspondances. *Les cahiers de l'analyse des données*, 7 (1) 9-23 et 7 (2) 133-154.
- CAZES P., CHESSEL D. et DOLEDEC S. (1988), L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. *Rev. Statistique Appliquée*, XXXVI (1) 39-54.
- ESCOFIER B. (1979a), Une représentation des variables dans l'analyse des correspondances multiples. *Rev. Statistique Appliquée* XXVII (4) 37-47.
- ESCOFIER B. (1979b), Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Les cahiers de l'analyse des données* 4 (2) 137-146.
- ESCOFIER B. et PAGES J. (1983), Méthode pour l'analyse de plusieurs groupes de variables. *Rev. Statistique Appliquée* XXXI (2) 43-59.
- ESCOFIER B. et PAGES J. (1998), *Analyses factorielles simples et multiples*. 3^e ed. Dunod.
- ESCOUFIER Y. (1973), Le traitement de variables vectorielles. *Biometrics*, 29, 751-760.
- GOLDFARB B. et PARDOUX C. (2001), Etude de données multidimensionnelles évolutives et comparaison de codages par l'analyse factorielle multiple. *Rev. Statistique Appliquée* XLIX, (1) 97-117.
- LEBART L., MORINEAU A. et TABARD N. (1977), *Techniques de la description statistique*, Dunod, Paris.
- LEBART L. (1989), Stratégies du traitement des données d'enquêtes. *La revue de Modulad*, (3) 21-29.
- LEBART L., MORINEAU A. et PIRON M. (1999), *Statistique exploratoire multidimensionnelle*, 2^e éd. Dunod.
- MORINEAU A. (1993), Le thémascope ou analyse structurelle des données d'enquêtes. *In traitement statistique des enquêtes*. Grangé D. et Lebart L.; Eds. Dunod.

- PAGES J. (1996), Eléments de comparaison entre l'analyse factorielle multiple et la méthode Statistique. *Rev. Statistique Appliquée* XLIV (4) 81-95.
- SAPORTA G. (1975), *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. Thèse de 3^e cycle, Paris VI.
- SAPORTA G. (1990), *Simultaneous analysis of qualitative and quantitative data*, Atti della XXXV riunione scientifica; società italiana di statistica, 63-72.
- TENENHAUS M. (1977), Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de statistique appliquée*, XXV (2) 39-56.
- TENENHAUS M. et YOUNG F. (1985), An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrica* 50 (1) 91-119.

Logiciels

- LADDAD (1997), Diffusé par : ADDAD, 151 Bd de l'Hôpital, F-75013 Paris.
- SPAD (2001), Diffusé par CISIA, 261, rue de Paris, F-93100 Montreuil.