CAHIERS DU BURO

B. ESCOFIER
J. PAGES

L'analyse factorielle multiple

Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 42 (1984), p. 3-68

<http://www.numdam.org/item?id=BURO_1984__42__3_0>

L'ANALYSE FACTORIELLE MULTIPLE

B. ESCOFIER[1] ; J. PAGES[2]

(1) I.R.I.S.A., avenue du Général Leclerc 35042 RENNES Cedex

(2) E.N.S.A.R., 65, rue de St Brieuc 35042 RENNES Cedex

SOMMAIRE

- INTRODUCTION -

Depuis plusieurs années, les statisticiens sont de plus en plus souvent consultés, aussi bien par des chercheurs de sciences de la nature que des sciences humaines à propos de problèmes impliquant l'étude simultanée de plusieurs tableaux. Il s'agit souvent de suites de tableaux indicés par le temps, ou de tableaux provenant d'un unique tableau de dimension trois, mais on rencontre aussi des ensembles de tableaux définis par différents groupes de variables mesurées sur les mêmes individus ou différents groupes d'individus caractérisés par la même variable, etc...

Ces problèmes s'inscrivent en droite ligne de l'évolution des quinze dernières années, permise par le développement de l'analyse des données, lui-même permis par celui de l'informatique : au raisonnement "toutes choses égales par ailleurs", au découpage cartésien d'un problème complexe en multiples problèmes plus simples, on substitue souvent une approche globale visant à prendre en compte simultanément tous les aspects d'un même phénomène.

Dans cet article, nous proposons une méthode, baptisée Analyse Factorielle Multiple, pour analyser simultanément ou comparer plusieurs tableaux croisant un même ensemble d'individus et différents groupes de variables (numériques ou qualitatives). Classiquement ces tableaux sont étudiés soit par une méthode usuelle de traitement de tableau unique (ACP ou AFC) moyennant une juxtaposition judicieuse de sous-tableaux, soit au moyen de méthodes spécifiques (Analyse Multicanonique, Statis, INDSCAL..). Chacune de ces analyses conduit à exhiber un aspect du problème et à le prendre en considération isolément. Or, du point de vue de l'interprétation, beaucoup de ces aspects sont liés et leur synthèse est d'autant plus difficile qu'ils sont abordés par des méthodes mises en oeuvre indépendamment.

Cette remarque est au coeur de l'Analyse Factorielle Multiple dont la caractéristique essentielle est d'exploiter le plus possible les relations intrinsèques entre des objectifs aussi apparemment divers que ceux de l'Analyse Canonique et du modèle INDSCAL par exemple.

Numériquement, l'analyse repose sur un calcul principal unique (une diagonalisation) dont les résultats sont repris par des calculs secondaires fournissant une réponse à chaque question particulière. Il résulte de cette démarche une grande cohérence dans l'interprétation simultanée des résultats associés aux différents aspects d'un tableau multiple.

Dans un premier chapitre, nous formalisons différents objectifs de l'étude d'un tableau multiple en référence aux méthodes classiques et en mettant en évidence les limites de ces méthodes. Chaque formalisation introduit une démarche qui aboutit à l'Analyse Factorielle Multiple.

Le second chapitre propose et discute une représentation des groupes de variables (qui induit une mesure de liaison entre ces groupes) ainsi qu'un système de pondération (qui précise la mesure de liaison et équilibre le rôle des groupes).

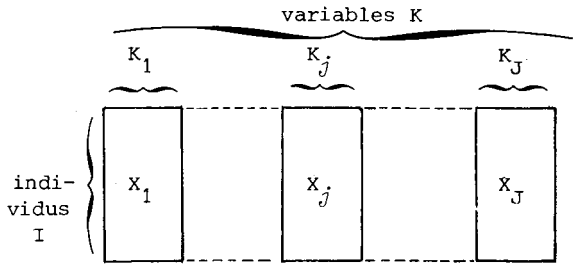
Les trois chapitres suivants décrivent l'Analyse Factorielle Multiple, chacun dans un espace vectoriel particulier. Ce découpage s'est avéré plus commode qu'un découpage par objectif avec lequel il ne se recoupe pas totalement.

L'évaluation de la méthode est réalisée dans le 7e chapitre qui la compare avec les principales méthodes présentant des objectifs analogues. Dans le cas de groupes réduits à une seule variable, on retrouve l'A.C.P. s'il s'agit de variables numériques et l'A.F.C. multiple s'il s'agit de variables qualitatives : d'où le nom d'Analyse Factorielle Multiple.

1. LES OBJECTIFS SOUS-JACENTS A L'ETUDE SIMULTANEE DE PLUSIEURS GROUPES DE VARIABLES

1.1 Notations

A chaque groupe de variables correspond un tableau. Tous les groupes de variables étant définis sur le même ensemble d'individus, tous les tableaux peuvent être juxtaposés et former ainsi un seul tableau individus × variables. L'ensemble initial de plusieurs tableaux apparaît alors comme un unique tableau structuré en sous-tableaux. Nous notons : X le tableau complet ; I l'ensemble des individus ; K l'ensemble des variables (tous groupes confondus) ; J l'ensemble de sous-tableaux ; K_j l'ensemble des variables du groupe j ($K = \bigcup_j K_j$) ; X_j le tableau associé au groupe j.



Le tableau X

Les symboles I,J,K ou K_j désignent à la fois l'ensemble et son cardinal. Une variable du groupe K_j est notée : $v_k, k \in K_j$. Nous supposons les individus et les variables munis de poids : p_i désigne le poids affecté à l'individu i ($\Sigma\, p_i = 1$), m_k le poids affecté à la variable v_k. Les matrices diagonales des poids des individus et des variables sont notées respectivement D, M_j (pour le groupe j) et M (pour K).

1.2 Variables et structures sur les individus

L'analyse des individus (vus au travers des variables) et l'analyse des variables (mesurées sur les individus) sont deux approches duales du même tableau. Il s'ensuit que pour étudier et comparer des groupes de variables, nous pouvons choisir entre deux points de vue :

raisonner sur les variables ou les groupes de variables eux-mêmes ou raisonner sur les structures induites sur le même ensemble d'individus. Les individus étant classiquement représentés par un nuage de points dans un espace euclidien, la structure de l'ensemble est souvent appelée "forme du nuage". Nous notons N_I^j le nuage des individus défini par le groupe K_j.

1.3 Pondération des groupes

Quel que soit le point de vue choisi, il est nécessaire d'équilibrer le rôle des groupes de variables : l'ensemble des résultats est perturbé si un groupe est prépondérant.

Théoriquement, un tableau soumis à l'Analyse des Données doit être, selon le mot de BENZECRI (cf. [1]) exhaustif. Ce n'est presque jamais le cas d'un tableau individus × variables. Le chercheur mesure plus souvent ce qu'il peut que ce qu'il veut et certains aspects des individus, pour des raisons pratiques, donnent lieu à un plus grand nombre de mesures que d'autres.

Si l'échantillonnage des individus est classiquement abordé, il n'en est pas de même des variables. Une analyse dans laquelle le rôle des groupes de variables est équilibré constitue un élément de solution à ce problème et réalise un peu l'analogue du redressement d'un échantillon d'individus.

Le déséquilibre entre les groupes peut provenir du nombre de variables, mais aussi de leurs relations. (La prépondérance d'un facteur peut ne traduire que la redondance entre mesures). Pour équilibrer le rôle des groupes, nous les pondérons en multipliant le poids des variables du groupe j par un même coefficient α_j. Il paraît souhaitable que ce coefficient α_j tienne compte, non seulement du nombre de variables, mais de la structure des groupes.

1.4 Comparaison globale des groupes de variables

La comparaison globale des groupes est l'un des objectifs de l'étude simultanée de plusieurs groupes de variables. Dans le contexte un peu plus général de la comparaison de tableaux, Y. ESCOUFIER et L'HERMIER DES PLANTES, intitulent ce problème "étude de l'interstructure" (cf. [2] et [12]).

Pour visualiser les proximités entre tableaux, Y. ESCOUFIER et H. L'HERMIER DES PLANTES recourrent à l'analyse factorielle d'un nuage dont les points représentent les tableaux. Mais les possibilités d'interprétation des axes ainsi obtenus sont limitées et il est difficile de préciser en quoi deux tableaux, proches selon une direction, se ressemblent. Pour cela, il faut pouvoir relier ces graphiques aux représentations des individus, des variables, ou des composantes principales des groupes de variables.

En A.F.M., pour comparer globalement les groupes de variables, on définit une mesure de liaison entre ces groupes. Puis, on cherche une représentation approchée des distances entre les groupes étudiés, complétée par des représentations des individus et des variables intervenant comme éléments explicatifs.

1.5 Comparaison des nuages d'individus

Dans ce paragraphe, on s'intéresse aux différents nuages d'individus définis par chacun des groupes de variables et on souhaite réaliser une comparaison analytique de ces divers nuages. Le terme "analytique" est utilisé en opposition au terme global du paragraphe précédent car le but poursuivi est maintenant d'analyser et non plus de mesurer les différences et les ressemblances entre les nuages.

Cette analyse elle-même peut avoir plusieurs points de vue. On peut étudier l'évolution de chaque individu à travers les groupes de variables, c'est à dire l'évolution de ses distances avec les autres individus. C'est dans cette optique qu'est introduite ce que Y. ESCOUFIER et L'HERMIER DES PLANTES appellent l'étude des "intra-structures". Pour cette comparaison des positions d'individus, la solution proposée est une représentation simultanée des nuages d'individus sur des plans ou dans un espace de petite dimension.

1.5-1 Le problème de la représentation simultanée

Deux propriétés nous paraissent essentielles pour qu'une telle représentation permette de comparer la position d'un même individu dans les différents nuages :

Chaque nuage N_I^j *est "bien représenté"*. Dans ce but, nous choisissons comme représentation du nuage N_I^j une projection orthogonale de ce nuage. La qualité d'une représentation peut alors être mesurée par son inertie : nous cherchons des projections des N_I^j d'inertie importante. (Notons que dans la méthode STATIS, les nuages ne sont pas représentés par des projections, leur qualité de représentation ne peut être mesurée à l'aide des critères habituels).

Les représentations des nuages N_I^j *se "ressemblent" entre elles*. En effet, il n'est pas possible de comparer les positions d'un même point dans les différents nuages si ces représentations sont, dans l'ensemble, très différentes. En particulier, des symétries, rotations ou homothéties peuvent masquer complètement de fortes ressemblances entre les nuages. Pour assurer cette propriété, nous utilisons l'un ou l'autre des deux critères suivants :

. Les points homologues (représentant le même individu) doivent être le plus proche possible les uns des autres.

. Les coordonnées des points homologues sur un même axe sont le plus corrélé possible : dans l'espace commun des représentations des N_I^j, les projections des N_I^j sur un même vecteur de base sont le plus homothétique possible.

Qualité de représentation des nuages et ressemblance entre ces représentations ne peuvent être optimisées simultanément. Ainsi, les meilleures représentations planes sont obtenues par les projections des nuages sur les plans engendrés par leurs deux premiers axes d'inertie ; mais le cas limite, où deux nuages ne diffèrent que par l'ordre de leurs axes d'inertie, montre bien que ces représentations de deux nuages très semblables peuvent ne pas être comparables. A l'inverse, une projection des nuages telle que tous

les points sont confondus à l'origine, optimise la ressemblance, mais ne présente aucun intérêt. Il faut donc trouver un compromis entre ces deux extrêmes. Nous l'obtenons en définissant un critère qui donne, a priori, une importance équivalente aux deux propriétés.

Pratiquement, pour obtenir une représentation simultanée, il faut projeter les nuages sur un espace de petite dimension et superposer ces nuages. Si la projection précède la superposition, le cas limite évoqué ci-dessus montre que la qualité de représentation est privilégiée aux dépens de la ressemblance. C'est l'inverse si la superposition (obtenue par exemple par des techniques d'analyse procustéenne (cf.[13]et[26]) précède la projection. Pour équilibrer les deux propriétés, il est nécessaire d'effectuer simultanément les deux opérations.

1.5-2 Le nuage moyen

On facilite beaucoup la comparaison des nuages en construisant un nuage moyen qui a la propriété de ressembler le plus possible à l'ensemble des nuages : on substitue alors aux comparaisons deux à deux (en nombre $J\times(J-1)/2$) J comparaisons à une moyenne. Dans la représentation simultanée, il est toujours possible de construire un nuage moyen au centre de gravité des différents nuages. Mais, il est préférable de définir a priori un nuage moyen et de le projeter avec les autres nuages.

En A.F.M., on cherche à représenter les J nuages d'individus, ainsi qu'un nuage moyen, dans un même espace de petite dimension. Les nuages sont représentés par des projections d'inertie importante et aussi semblables que possible entre elles. Elles sont complétées par des indices d'aides à l'interprétation (qualité de représentation de chaque nuage et ressemblance entre les nuages).

1.6 Comparaison des nuages de variables

Chaque groupe de variables est représenté par un nuage de points dans l'espace R^I. L'un des aspects de l'étude simultanée de plusieurs groupes des variables est une comparaison directe de ces

nuages ("directe" est employé en opposition avec le paragraphe précédent qui étudie les variables au travers des structures qu'elles induisent sur les individus).

Une analyse factorielle permet une représentation approchée de ces nuages et des corrélations inter et intra groupes. Mais dans une telle analyse, un groupe peut influencer de manière excessive les résultats et un équilibrage est nécessaire (cf §1.2).

On peut aussi étudier les corrélations entre les composantes principales des groupes par une A.C.P. de ces composantes préalablement pondérées.

1.7 Objectifs exprimés en termes de liaison : les analyses canoniques

C'est en ces termes que l'analyse simultanée de plusieurs groupes de variables a été d'abord formulée. Nous faisons allusion ici au cas de deux tableaux étudiés par HOTELLING en 1936 (cf.[16]) à l'aide ce qu'il appela l'analyse canonique. De nombreuses généralisations au cas de J (J>2) groupes de variables ont été proposées. L'objectif est alors de rechercher J combinaisons linéaires de variables (chaque combinaison est définie sur un groupe) telles que ces combinaisons soient liées le plus possible. (cf. HORST [15], KETTENRING [17], CARROLL [5]). Ces méthodes sont regroupées sous le terme d'Analyse Multicanonique ou Analyse Canonique Généralisée. Toutes ces techniques sont très peu utilisées du fait des difficultés d'interprétations. Nous donnons ici notre point de vue sur les analyses canoniques généralisées.

1.7-1 Primauté des variables générales

Dans la méthode proposée par CARROLL, on cherche d'abord une variable générale liée le plus possible à tous les groupes de variables, puis, cette variable générale obtenue, on cherche dans chaque groupe la variable qui lui est la plus liée.

Les variables générales constituent en fait un résultat essentiel et non une simple commodité de calcul. Dans le cas limite

(auquel on se réfère toujours plus ou moins explicitement) où les variables canoniques sont confondues, il existe une direction commune à tous les groupes. Or, même en dehors de ce cas limite, la recherche de directions communes est le fondement de ce type d'analyse. Un ensemble de J variables, même très proches entre elles représentent mal une direction ; une seule variable, proche de tous les groupes est d'emploi plus facile et correspond mieux à l'idée de direction commune. Naturellement, une variable générale doit être accompagnée d'un indice global mesurant sa liaison avec chacun des groupes.

L'idée de direction commune est déjà présente, quoique non exploitée, dans la première présentation de l'Analyse Canonique (cf. [16]). En effet, HOTELLING centre son analyse sur les coefficients de corrélation canonique : pour lui, un fort coefficient implique l'existence d'un facteur commun, cependant, il ne poursuit pas cette démarche et ne cherche pas à décrire ce facteur.

La primauté des variables générales amène à considérer les variables canoniques comme les représentantes, dans chaque groupe de la direction "commune" résumée par une variable générale. Le courant de pensée de l'Analyse des Données a incité à utiliser ces variables pour construire des représentations graphiques (difficilement interprétables) des variables et des individus.

Pour représenter les variables, les variables générales dont nous avons souligné l'intérêt au paragraphe précédent forment un système de référence bien plus adapté que les variables canoniques de l'un ou l'autre groupe.

Pour les individus, une amélioration importante des résultats serait que la représentation de l'ensemble des individus par 2 variables canoniques du même groupe j soit une projection orthogonale du nuage N_I^j. La superposition des graphiques de tous les groupes serait alors une représentation simultanée de ces nuages dans le sens précisé au paragraphe 1.5.1 .

1.7 - 3 Variance expliquée

En plus des difficultés d'interprétation, un reproche fait à l'analyse canonique (généralisée ou non) est que la variance des groupes de variables "expliquée" par les variables canoniques obtenues peut être très faible.

Pour obtenir des variables canoniques représentant mieux les groupes (au sens de la variance expliquée i.e de l'inertie du nuage de variables dans la direction de la variable canonique), des techniques un peu différentes de l'Analyse Canonique ont été proposées (cf. [27]). Elles s'appuient sur un indice de liaison qui, contrairement au coefficient de corrélation multiple, tient compte de l'inertie du nuage de variables dans les différentes directions.

L'A.F.M. peut être vue comme une Analyse Multicanonique dans laquelle l'indice de liaison entre groupes de variables prend en compte la variance expliquée. Elle cherche des variables générales, (des directions communes aux nuages de variables), puis en déduit des variables canoniques qui traduisent cette direction pour chaque groupe et qui définissent des projections des N_I^j *susceptibles d'être superposées dans une représentation simultanée.*

1.8 Le modèle INDSCAL

Cette approche est différente des précédentes : un modèle est proposé, dont il faut calculer les paramètres. Le modèle INDSCAL (Analysis of Individual Differences in multidimensional Scaling) dû à CARROLL et CHANG (cf. [4]) a été développé à partir de besoins exprimés par la psychométrie pour décrire la situation où plusieurs personnes (appelées juges) décrivent leur perception des proximités d'un ensemble d'objets au moyen d'une matrice de similarité ou de distances). Il s'applique donc à des données plus générales que les nôtres : matrices de distances entre objets ou matrices de similarités. Les données auxquelles nous nous intéressons peuvent être vues au travers du modèle INDSCAL puisque chaque groupe de variables définit une matrice de

distance entre les individus ou objets.

Selon ce modèle, les distances entre individus peuvent se décomposer suivant un certain nombre de "facteurs" communs à tous les groupes, les poids affectés à chaque facteur différant suivant les groupes. Plus précisément en notant :

- $z_s(i)$ la valeur du $s^{\text{ième}}$ facteur pour l'individu (s=1,S) ;
- q_s^j le poids affecté à z_s par le $j^{\text{ième}}$ groupe ;
- $d_j(i,l)$ la distance entre i et l induite par le $j^{\text{ième}}$ groupe.

Le modèle s'écrit : $$d_j^2(i,l) = \sum_{s=1}^{S} q_s^j \{z_s(i) - z_s(l)\}^2$$

Il peut être vu comme une décomposition des nuages N_I^j suivant S projections axiales $q_s^j z_s$ homothétiques aux z_s dans tous les nuages.

Les données ne vérifient jamais exactement le modèle. Les paramètres calculés (facteurs et poids) par différents algorithmes minimisent des critères d'ajustement variés.

Les facteurs sont des images de dimension 1 des individus pour lesquelles il existe des projections des N_I^j "presque" homothétiques. Ils représentent des directions "communes" aux nuages N_I^j comme les variables générales représentent celles des nuages de variables. C'est une optique de comparaison des nuages N_I^j un peu différente de la représentation simultanée du §1.5, mais qui s'en approche si l'on prend comme critère de ressemblance entre les représentations des N_I^j l'homothétie de leur projection (cf 1.5.1). Les projections de chaque N_I^j peuvent, (comme les variables canoniques) jouer le rôle de représentant de la direction commune dans chacun des nuages. Dans l'A.F.M. nous exploitons ces deux analogies. Enfin, dans les méthodes proposées jusqu'ici pour estimer les paramètres, modèle (INDSCAL [4] et [5], ALSCAL [20] ; SUMSCAL [25]) deux difficultés apparaissent : a) les algorithmes sont itératifs et posent des problèmes de convergence b) les poids estimés peuvent être négatifs.

L'A.F.M. contient une interprétation géométrique du modèle INDSCAL et fournit une solution non itérative (pas de problème de convergence) qui aboutit systématiquement à des poids positifs.

1.9 Conclusion

Nous avons dégagé un certain nombre d'objectifs pour comparer les groupes de variables :

- comparaison globale des groupes ;
- comparaison des nuages d'individus par l'intermédiaire d'une représentation simultanée (i.e. projections de ces nuages se ressemblant entre elles) complétée par la projection d'un nuage moyen ;
- comparaison directe des nuages de variables et des composantes principales de chaque groupe ;
- recherche de combinaisons linéaires de variables de chaque groupe corrélées entre elles (cf. analyse multicanonique) ;
- modèle INDSCAL (facteurs communs aux groupes avec poids dépendant de chaque groupe) ;
- pondération des groupes.

Ces objectifs s'expriment en fonction d'objets de nature différente : les variables, les groupes de variables et les individus associés à chaque groupe de variables. Pour formaliser mathématiquement ces objectifs, nous plaçons ces objets dans des espaces euclidiens qui servent de cadre de référence. Nous cherchons pour chaque objectif une solution optimum. Ces solutions coïncident et forment une méthode unique qui s'interprète dans chaque cadre de référence.

2. PONDERATION ET REPRESENTATION DES GROUPES DE VARIABLES

La transcription géométrique du modèle INDSCAL, la comparaison globale des groupes de variables (§1.4), l'indice de liaison entre une variable et un groupe de variables de l'analyse multicanonique (§1.7) sont basés sur une représentation de ces groupes par des vecteurs (notés W_j) d'un espace euclidien. Dans ce chapitre nous étudions les propriétés de cette représentation ainsi que la pondération des groupes (§1.2).

2.1 La pondération des groupes de variables

La pondération des groupes remplit plusieurs rôles, soit principalement : équilibrer l'influence des groupes dans les analyses ; "normaliser" les nuages d'individus associés à chaque groupe de variables ; rendre compatibles et donc utilisables les indicateurs de liaison (et les poids dans INDSCAL).

Dans l'Analyse Factorielle Multiple, cette pondération s'effectue en multipliant le poids de toutes les variables du groupe j par un même coefficient $\alpha_j=1/\lambda_j^1$ où λ_j^1 est la première valeur propre de l'A.C.P. du tableau X_j.

Cette pondération peut être illustrée en considérant un groupe de deux variables normées v_1 et v_2 initialement de même poids. Si v_1 et v_2 sont non corrélées, le poids de chacune de ces variables, à l'issue de la pondération du groupe, vaut 1. Si, au contraire, v_1 et v_2 sont très corrélées (à la limite confondues), le poids final de chacune vaut 1/2 ; un tel groupe est ainsi rendu équivalent à un groupe d'une seule variable, ce qu'il est en réalité. L'inertie totale du premier nuage vaut 2 tandis que celle du second vaut 1. Ceci rend compte de la présence de 2 directions d'importance égale dans le premier et d'une seule direction dans le second. La pondération choisie égalise l'inertie des nuages le long de leur première direction d'inertie et non leur inertie totale.

Cette pondération est une des originalités de la méthode. Elle se justifie pour chacun des objectifs comme cela est explicité au cours des paragraphes suivants. Dans la suite le coefficient α_j est inclus dans le poids des variables.

2.2 Représentation des variables et des individus

Les variables sont représentées par des vecteurs de l'espace R^I des fonctions numériques définies sur l'ensemble des individus I. Cet espace est muni de la métrique diagonale des poids des individus. Le groupe de variables j forme un nuage noté N_K^j.

L'ensemble des individus est représenté par les nuages N_I^j dans les espaces R^{Kj} définis par chaque groupe de variables et par un nuage N_I dans l'espace R^K défini par toutes les variables. Ces espaces sont munis des métriques diagonales M_j et M des poids des variables.

La pondération des groupes de variables par le coefficient $\alpha_j = 1/\lambda_j^1$ (cf.§ 2.2.), joue, entre autres, un rôle de "normalisation" des N_I^j. Chaque nuage subit une homothétie de rapport $\sqrt{\alpha_j}$ qui rend égale à 1 son inertie maximum dans une direction donnée. De ce fait, deux nuages homothétiques deviennent égaux (ce qui justifie le terme "normalisation") et deux nuages dont les projections sur leurs premiers axes d'inertie se ressemblent sont normalisés de manière identique quelque soient leurs dimensions respectives.

2.3 Représentation des groupes

Le vecteur W_j appartient à un espace euclidien de dimension I^2 noté R^{I^2} et caractérise le groupe j suivant plusieurs points de vue (nuage d'individus, nuage de variables):

2.3-1 W_j matrice de produit scalaire entre individus

Le nuage d'individus N_I^j associé au groupe de variables j est situé dans l'espace R^{Kj}. Considérons les vecteurs joignant l'origine à chaque point i^j de N_I^j. Soit W_j la matrice symétrique, de dimension I×I, contenant les produits scalaires entre ces vecteurs pris deux à deux :

$$W_j = X_j M_j X_j' = \sum_{k \in K_j} m_k v_k v_k'$$

où X_j' est la transposée de X_j, v_k' le transposé du vecteur v_k et M_j la matrice diagonale des poids des variables du groupe j (i.e. la métrique de R^{Kj}). Pour qu'il y ait bijection entre matrice de produits scalaires et matrices de distance, il suffit de fixer l'origine des axes au centre de gravité du nuage, ce qui revient à centrer les variables. Dans la suite, sauf expression du contraire, les variables sont supposées centrées.

2.3-2 W_j tenseur d'inertie

La matrice W_j est aussi la matrice de la forme quadratique d'inertie, (ou tenseur d'inertie) du nuage N_K^j représentant les variables du groupe j dans R^I. En notation tensorielle et matricielle,

$$W_j = \sum_{k \in K_j} m_k \, v_k \otimes v_k = \sum_{k \in K_j} m_k \, v_k \, v_k'$$

où v_k' est le transposé du vecteur v_k.

Ce tenseur d'inertie peut être considéré soit comme un élément du produit tensoriel $R^I \otimes R^I$, soit comme une forme bilinéaire symétrique sur le dual de R^I, noté $(R^I)^*$, soit comme une application linéaire de $(R^I)^*$ dans R^I. Il permet de calculer, avec la métrique D, l'inertie de la projection du nuage N_K^j dans une direction quelconque z de l'espace R^I : on applique la forme bilinéaire symétrique W_j à l'image Dz du vecteur z (unitaire) par la métrique D (considérée comme une application linéaire de R^I dans son dual) :

$$W_j(Dz, Dz) = z' D W_j D z = \sum_{k \in K_j} m_k \langle v_k, z \rangle^2$$

2.3-3 $W_j D$ opérateur de R^I dans R^I

L'opérateur $W_j D$ où W_j est considéré comme une application linéaire de $(R^I)^*$ dans R^I a pour vecteurs propres et valeurs propres les axes principaux d'inertie de N_I^j et les moments d'inertie associés. Le groupe j est donc caractérisé par $W_j D$ dans l'espace des opérateurs sur R^I.

2.3-4 Comparaison entre W_j et le sous-espace engendré par un groupe de variables

En Analyse Canonique ou Multicanonique, on associe à chaque groupe de variables l'opérateur de projection orthogonale sur le sous-espace de R^I qu'elles engendrent. Ce projecteur, ainsi que $W_j D$, caractérisent tous deux ce sous-espace qui est leur image, mais $W_j D$ contient plus d'informations, ses valeurs propres non nulles n'étant pas toutes égales à 1. Le projecteur ne permet d'ailleurs pas de calculer les distances dans N_I^j. En outre, W_j est plus stable que l'opérateur de projection orthogonale, très sensible à de petites variations des variables lorsque ces dernières sont liées.

2.3-5 Cas des variables qualitatives

Une variable qualitative définit une partition de l'ensemble des individus. Si l'on associe à cette variable l'ensemble des variables indicatrices des classes de cette partition, on peut dire qu'une variable qualitative est un cas particulier de groupe de variables.

On peut caractériser la variable qualitative j (i.e. la partition qu'elle induit) par le sous-espace E_j de R^I engendré par ces indicatrices, constitué par les fonctions prenant la même valeur pour des individus appartenant à la même classe de la partition (ou par le sous espace $\bar{E}_j$ des fonctions centrées de E_j). Dans ce cas particulier, toutes les directions de E_j ou de $\bar{E}_j$ ont même importance et le projecteur suffit pour caractériser le groupe. Certaines pondérations (cf. ci-dessous) font coïncider l'opérateur $W_j D$ avec le projecteur et dans la suite nous appelons variable qualitative un groupe de variables indicatrices ainsi pondérées.

Les variables indicatrices non centrées étant orthogonales deux à deux, si leur inertie vaut 1 l'opérateur $W_j D$ coïncide avec le projecteur sur E_j. En analyse des correspondances multiples (qui traite bien ce type de variables), cette condition est réalisée (les variables sont transformées en "profil" et pondérées par leur effectif). Si on affecte aux variables centrées réduites le poids $(1-\bar{v}k)$ où $\bar{v}_k$ désigne la moyenne de la variable v_k, $W_j D$ coïncide avec le projecteur sur $\bar{E}_j$. En effet, si l'inertie d'un groupe de variables non centrées vaut 1 dans toutes les directions de E_j, l'inertie des projections de ces variables sur l'orthogonal $\bar{E}_j$ du vecteur constant (i.e. des variables centrées) vaut 1 dans toutes les directions de $\bar{E}_j$ et $\bar{v}_k$:

$$\| v_k - \bar{v}_k \|^2 = \bar{v}_k(1-\bar{v}_k)$$

La plus grande valeur de P_j étant égale à 1, une variable qualitative ainsi définie n'a pas besoin d'être surpondérée si on l'introduit comme un groupe de variables particulier dans l'A.F.M.

2.4 La structure euclidienne de R^{I^2}

2.4-1 Définition du produit scalaire dans $R^I \otimes R^I$

La métrique D de R^I induit sur $R^I \otimes R^I$ une métrique notée D ⊗ D. Par définition, pour les tenseurs de rang un qui s'écrivent z ⊗ v, (où z et v sont des vecteurs quelconques de R^I), le produit scalaire vaut :

$$\langle z \otimes v,\ x \otimes y \rangle_{D \otimes D} = \langle z,x \rangle_D \langle v,y \rangle_D = z'D\ x\ v'\ Dy$$

Sa valeur pour un tenseur de rang quelconque qui est une combinaison linéaire de tenseurs de rang un s'en déduit par bilinéarité. Ainsi, pour deux tenseurs W_j et $W_{j'}$, associés aux groupes de variables K_j et $K_{j'}$, affectés des poids $\{m_k / k=1,K_j\}$ et $\{n_l / l=1,K_{j'}\}$:

$$\langle W_j,\ W_{j'} \rangle = \Big\langle \sum_{k \in K_j} m_k\ v_k \otimes v_k\ ,\ \sum_{l \in K_{j'}} n_l\ v_l \otimes v_l \Big\rangle$$

$$= \sum_{k,l} m_k\ n_l\ (v'_k\ D\ v_l)\ (v'_k\ D\ v_l) = \text{trace}\ (W_j D\ W_{j'} D)$$

Ce produit scalaire est toujours positif.

2.4-2 Tenseurs W_j et opérateurs $W_j D$

On reconnaît dans le dernier terme le produit scalaire classique entre les opérateurs D symétriques $W_j D$ et $W_{j'} D$. Il est équivalent de travailler avec W_j dans l'espace des tenseurs ou avec $W_j D$ dans l'espace des opérateurs muni de ce produit scalaire. Suivant les cas, on choisit la notion la plus facile à manier. Cette équivalence justifie la pratique maintenant classique des opérateurs $W_j D$ avec ce produit scalaire. (cf. |12| et |21|).

2.4-3 Distance entre matrice de produits scalaires

La distance entre deux matrices W_j et $W_{j'}$, induite par D ⊗ D s'écrit en notant $a_{ii'}$, le terme général de la première et $b_{ii'}$, celui de la seconde :

$$D^2(W_j, W_{j'}) = \cdot \sum_{i,i'} (a_{ii'} - b_{ii'})^2\ p_i\ p_{i'}$$

2.5 Interprétation du produit scalaire. Liaison entre deux groupes de variables

Le produit scalaire défini sur les tenseurs ou opérateurs associés à un groupe de variables sert de mesure de liaison entre ces groupes dans l'Analyse Factorielle Multiple.

2.5-1 Les deux groupes possèdent chacun une seule variable

La pondération par α_j donne le poids 1 à une variable centrée réduite qui constitue à elle seule un groupe.

Soit z et v deux variables centrées réduites de poids 1. Les tenseurs associés à ces groupes ont pour norme 1 et leur produit scalaire est le carré du coefficient de corrélation entre z et v.

$$<W_1, W_2>_{D \otimes D} = <z \otimes z, v \otimes v>_{D \otimes D} = \{<z,v>_D\}^2$$

2.5.2. Groupe d'une variable et groupe quelconque

Notons z la variable (réduite et de poids 1) du groupe K_1 réduit à un seul élément et v_k (réduites de poids m_k) les variables du groupe K_2. Alors :

$$<W_1, W_2>_{D \otimes D} = <z \otimes z, \sum_{k \in K_2} m_k v_k \otimes v_k> = \sum_{k \in K_2} (\text{inertie de la projection de } v_k \text{ sur } z)^2$$

La variable la plus liée à un groupe de variables donné est la première composante principale de ce groupe. L'inertie des projections des v_k est alors le plus grand moment d'inertie du nuage, i.e. la plus grande valeur propre de $W_2 D$, qui avec la pondération par α_j vaut 1. Ainsi, quel que soit le groupe, l'intervalle de variation de cette mesure de liaison avec une variable quelconque est $]0,1]$. Une telle pondération est nécessaire pour que ces mesures soient comparables d'un groupe à l'autre.

Si l'inertie du nuage des v_k vaut 1 dans toutes les directions du sous-espace qu'il engendre, la mesure proposée se confond avec le carré du coefficient de corrélation multiple. C'est le cas pour un groupe constitué de variables réduites, de poids un et orthogonales deux à deux. C'est le cas aussi pour une variable qualitative.

Mais, si les variables du groupe sont corrélées entre elles, cette mesure diffère du coefficient de corrélation multiple : elle est alors d'autant plus grande que z est proche d'une direction d'inertie importante des v_k.

2.5 -3 Cas des variables qualitatives

La liaison entre 2 variables qualitatives est décrite entièrement par le tableau de contingence croisant les deux partitions. Elle est classiquement mesurée par le Φ^2. Un calcul simple (cf. [2]) montre que le produit scalaire entre les tenseurs vaut Φ^2+1 si l'on considère les variables indicatrices non centrées et Φ^2 si l'on prend les variables centrées.

3. L'ANALYSE FACTORIELLE MULTIPLE DANS R^K

Dans l'espace R^K associé à toutes les variables, nous cherchons une représentation simultanée des J nuages N_I^j dans l'optique du § 1.4.

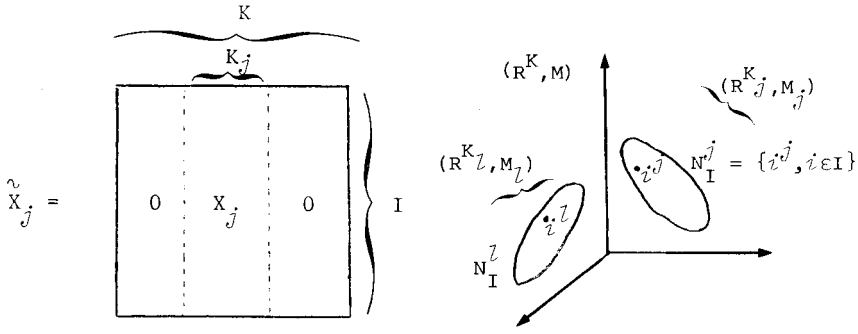
3.1 Les J nuages N_I^j dans R^K

Aux J groupes de variables correspondent J nuages différents N_I^j, chacun fournissant une image de I dans l'espace R^{K_j}.

Pour représenter simultanément les J nuages N_I^j dans l'espace R^K, il suffit de remarquer que R^K peut se décomposer en somme directe de sous-espaces orthogonaux deux à deux et isomorphes aux espaces R^{K_j}.

$$R^K = \bigoplus_j R^{K_j}$$

Sur chacun de ces sous-espaces, la métrique induite par M est la métrique M_j ; il s'agit donc d'un isomorphisme d'espaces euclidiens. Les coordonnées des points du nuage N_I^j sont contenues dans le tableau X_j. Les coordonnées de ces points dans l'espace R^K sont contenues dans un tableau noté $\tilde{X}_j$, de dimension I×K où X_j est complété par des zéros. Notons i^j le point représentant i dans le nuage N_I^j plongé dans R^{K_j}.



Les nuages N_I^j étant situés dans des sous-espaces orthogonaux, cette représentation simultanée est artificielle et inutilisable directement, mais sert de base à une véritable représentation simultanée obtenue par projection sur des sous-espaces de R^K.

3.2 Le nuage moyen N_I^* et le nuage N_I associé à toutes les variables

L'espace R^K contient le nuage associé à l'ensemble K de toutes les variables dont les coordonnées sont contenues dans le tableau X. Dans ce nuage noté N_I, le carré de la distance entre deux points i et l est la somme des carrés de leur distance dans les N_I^j :

$$d^2(i,l) = \sum_{k\in K} m_k (v_k(i) - v_k(l))^2 = \sum_{j\in J} \sum_{k\in K_j} m_k (v_k(i) - v_k(l))^2$$

$$= \sum_{j\in J} d^2(i^j, l^j)$$

L'influence des différents groupes n'est a priori équilibrée que si les distances dans les nuages sont du même ordre de grandeur. Multiplier les poids des variables du groupe j par un coefficient α_j, est un moyen d'équilibrer l'influence des groupes, puisque la distance s'écrit alors :

$$d^2(i,l) = \sum_{j\in J} \alpha_j \, d^2(i^j, l^j)$$

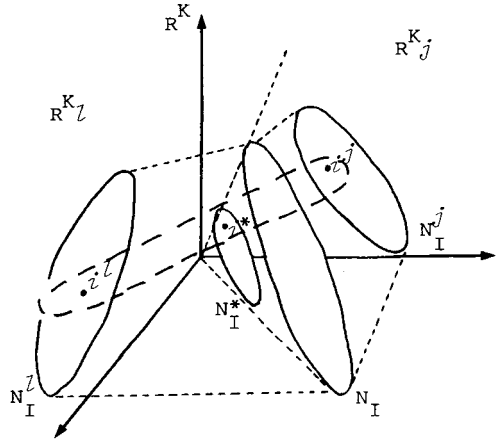
Dans chaque nuage N_I^j, $d^2(i^j, l^j)$ se décompose suivant des directions orthogonales dont le nombre, variable, est égal à la dimension du nuage. Avec la pondération $\alpha_j = 1/\sqrt{\lambda_j^1}$, aucune direction d'un seul sous

.../...

nuage ne peut être prépondérante dans la construction du nuage moyen. Le nombre de directions de N_I sur lesquelles N_I^j influe croît avec la dimension de N_I^j.

Soit N_J^* le nuage des centres de gravité notés i^* des J points i^j représentant le même individu i dans les N_I^j. Ce nuage se déduit de N_I par une homothétie de rapport 1/J. N_I^* est un nuage moyen pour les N_I^j, si l'on prend comme référence les carrés des distances entre deux points (critère raisonnable parmi d'autres critères possibles).

3.3 Bilan : les nuages en présence dans $R^K = \oplus R^{K_j}$



$N_I = \{i, i \in I\}$ images des individus par X

$N_I^j = \{i^j, i \in I\}$ images dans R^{K_j} des individus par X_j = projection de N_I sur R^{K_j}

$N_i^J = \{i^j, j \in J\}$ images du même individu i

$N_I^* = \{i^*, i \in I\}$ nuage des centres de gravité des N_i^J, homothétique de N_I dans le rapport 1/J

$N_I^J = \underset{j}{\cup} N_I^j = \underset{i}{\cup} N_i^I$

3.4 Le principe de la représentation simultanée des N_I^j

Dans la définition des objectifs (cf. §1.4-1), nous avons dégagé deux qualités essentielles pour qu'une représentation simultanée des J nuages N_I^j serve à les comparer efficacement :

(P1) Chaque nuage N_I^j est "bien représenté" : par une projection orthogonale de ce nuage d'inertie importante ;

(P2) Ces projections des J nuages N_I^j se"ressemblent entre elles". Cette qualité peut se traduire elle-même sous la forme : les J points i^j, images du même individu i sont proches entre eux.

Les nuages N_I^j étant tous situés dans R^K, il est possible d'en obtenir une représentation simultanée par projection sur un même sous-espace. Le choix du sous-espace doit être réalisé en fonction des deux qualités précédentes. Chacune est traduite par un critère à optimiser s'exprimant en termes d'inertie projetée.

(P1) Pour maximiser l'inertie projetée de tous les N_I^j, on maximise l'inertie de l'union des N_I^j, soit le nuage N_I^J qui comprend IxJ points ;

(P2) Le nuage N_I^J a été partitionné jusqu'ici en J nuages (contenant chacun I points) notés N_I^j représentant chacun l'ensemble des individus vus au travers d'un groupe de variables. Nous introduisons ici une autre partition de N_I^J : I nuages (contenant chacun J points) notés N_i^J représentant chacun le même individu vu au travers de chaque groupe de variables. Le centre de gravité de N_i^J n'est autre que i^*. Selon le principe de Huyghens appliqué à cette nouvelle partition , l'inertie totale de N_I^J se décompose en inertie intra (inertie des N_i^J autour des i^*) et inertie inter (inertie de N_I^*). Pour que les points associés au même individu i soient proches entre eux, il faut minimiser l'inertie projetée de chaque N_i^J donc l'inertie intra de N_I^J.

Compromis entre (P1) et (P2)

Pour optimiser les critères associés à (P1) et (P2), le sous-espace cherché doit être tel qu'en projection le nuage N_I^J ait une inertie totale maximum et une inertie intra minimum. Ces deux propriétés sont généralement incompatibles. Le théorème de Huyghens (inertie inter= inertie totale - inertie intra) suggère un compromis entre une inertie totale maximum et une inertie intra maximum : une inertie inter maximum. C'est celui que nous choisissons.

3.5 Interprétation en termes d'analyse factorielle

Le sous-espace de R^K sur lequel la projection de N_I^J a une inertie "inter" maximum est engendré par les premiers axes d'inertie, noté u_s, du nuage N_I^* des centres de gravité. Or, ce nuage est homothétique du nuage N_I associé à l'ensemble de toutes les variables. Le sous-espace cherché s'obtient par une Analyse en Composantes Principales du tableau X tout entier. La projection du nuage N_I^j notée F_s^j s'obtient par la formule de projection habituelle. Les coordonnées des points de N_I^j sont contenues dans un tableau $\tilde{X}_j$ de dimension I×K dans lequel X_j est complété par des zéros. Il suffit donc d'introduire dans l'A.C.P. les tableaux $\tilde{X}_j$ en supplémentaires pour obtenir la représentation simultanée des N_I^j.

La coïncidence de cette représentation simultanée avec une A.C.P. est précieuse : ses règles d'interprétation dérivent directement de celles d'une A.C.P..

3.6 Remarques sur la projection des nuages N_I^j

Le nuage N_I^j, qui appartient au sous-espace $R^K{}_j$, est projeté sur un vecteur u_s de R^K qui n'appartient par à $R^K{}_j$. La projection de N_I^j sur u_s , revient à réaliser successivement une projection sur un vecteur u_s^j (projection de u_s sur $R^K{}_j$) puis une projection sur u_s qui contracte le nuage en multipliant les coordonnées par $\cos(\theta_s^j)$ en notant θ_s^j l'angle entre u_s et u_s^j.

Ceci peut conduire à se demander s'il ne vaut pas mieux conserver les projections sur les u_s^j pour la représentation simultanée. En fait, il n'en est rien. Dans R^K, les axes u_s sont orthogonaux, ce qui n'est pas le cas des u_s^j. On superposerait alors des nuages dans des espaces munis de métriques différentes, ce qui est illisible. A la rigueur, on pourrait le faire en se limitant à un seul axe u_s. Mais, même dans ce cas simple, la propriété qui veut que le nuage moyen soit au centre de gravité ne serait plus vérifiée. En outre, les points homologues ne seraient plus proches entre eux.

3.7 Aides à l'interprétation

3.7-1 Qualité de représentation de chaque nuage N_I^j

Elle se mesure de manière classique par le rapport entre l'inertie projetée et l'inertie totale du nuage.

Cette qualité de représentation est toujours très faible. En effet, le vecteur u_s de R^K sur lequel N_I^j est projeté n'appartient pas au sous-espace dans lequel ce nuage est situé. Il fait avec sa projection u_s^j sur ce sous-espace un angle déjà noté Θ_s^j. D'où :

Qualité de représentation de N_I^j sur u_s = $(cos\ \Theta_s^j)^2$ (qualité sur u_s^j)

Les termes $cos^2\ \Theta_s^j$ sont en général, petits : ils sont en nombre J et leur somme vaut 1. Cette mesure de la qualité de représentation de N_I^j est donc systématiquement beaucoup plus faible que celle que l'on obtient dans l'A.C.P. du seul nuage N_I^j, même si un u_s^j est une composante principale de N_I^j. En d'autres termes, l'indicateur (inertie projetée)/(inertie totale) rend compte de façon pessimiste de la qualité de représentation en ce ce sens que la forme du nuage peut être bien respectée, même si le nuage est mal représenté au sens de l'indicateur.

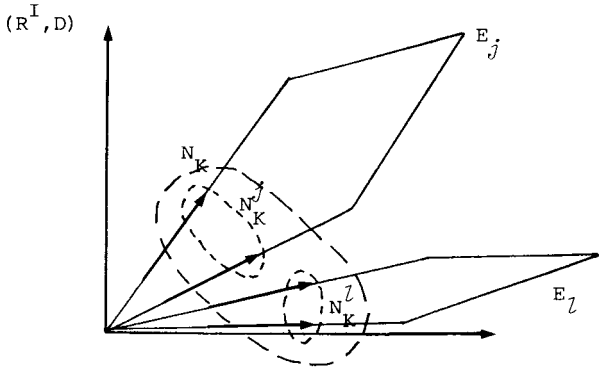
3.7-2 Ressemblance entre les représentations des différents nuages N_I^J

L'analyse cherche à rendre petite l'inertie intra du nuage N_I^J pour que les points i^j représentant le même individu i soient proches entre eux. Il est naturel de prendre comme critère de ressemblance sur un axe cette inertie intra. Mais cette valeur n'a de signification que comparée à l'inertie totale. On calcule donc, pour chaque facteur, et éventuellement chaque plan, le rapport : inertie inter/inertie totale.

4. L'ANALYSE FACTORIELLE MULTIPLE DANS R^I

Dans cet espace, l'Analyse Factorielle Multiple est décrite en premier lieu en tant qu'analyse des liaisons entre les groupes de variables dans une optique d'Analyse Canonique Généralisée (§4.2). Le lien avec la représentation simultanée du chapitre précédent et l'estimation des facteurs du modèle INDSCAL est assuré par une généralisation à plusieurs groupes de variables du concept de dualité (§4.3). Enfin, on cherche à comparer les nuages de variables (§4.4).

4.1 Notations



R^I est muni de la métrique D

v_k vecteur représentant une variable

N_K^j ensemble des vecteurs v_k appartenant au groupe j

$N_K = \bigcup_j N_K^j$

E_j sous-espace engendré par les variables du groupe j

4.2 L'analyse des liaisons entre les groupes de variables dans une optique d'Analyse Canonique Généralisée

Dans l'Analyse Canonique Généralisée, CARROLL recherche une première variable générale telle que la somme des carrés de ses coefficients de corrélation multiple avec tous les groupes soit maximum ; la variable canonique du groupe j est définie comme la combinaison linéaire des variables de ce groupe la plus corrélée à z.

Ce procédé est réitéré avec une contrainte d'orthogonalité sur les variables générales. La solution est donnée (cf. [17] et [24]) par la diagonalisation de la somme des opérateurs P_j de projection orthogonale sur les sous-espaces E_j. Les variables générales sont les

vecteurs propres de cette somme, ordonnées par leurs valeurs propres (qui sont égales aux quantités maximisées $\cos^2\theta j$). En effet, le carré du coefficient de corrélation multiple entre une variable normée z et le groupe j s'écrit : $\cos^2\theta j = \langle z, P_j\, z\rangle_D$ et $\Sigma\cos^2\theta j = \langle z, \sum_j P_j\, z\rangle_D$. En outre, les opérateurs P_j étant D-symétriques, leur somme l'est aussi et ses vecteurs propres sont orthogonaux deux à deux. La variable canonique z^j associée à la variable générale z est sa projection orthogonale : $z^j = P_j z$.

4.2-1 Les variables générales dans l'Analyse Factorielle Multiple

Dans l'analyse de CARROLL, toutes les directions des sous-espaces E_j jouissent de la même importance, ce qui pose problème lorsque les variables d'un même groupe sont corrélées (directions de E_j non significatives et instabilité de E_j)(cf. §2.3-4). Il est souhaitable que les variables générales expriment des directions communes "significatives", c'est-à-dire soient proches de directions d'inertie importante des nuages de variables N_{K_j}. Pour tenir compte de l'inertie de ces nuages dans les différentes directions de E_j, nous caractérisons le groupe j par W_j ou W_jD plutôt que l'opérateur de projection P_j qui ne définit que le sous-espace E_j (cf. §2.3-4). Nous prenons comme mesure de liaison celle introduite au §2.4 qui, appliquée à une variable normée z et au groupe K_j, s'écrit (cf. §2.5-2) :

$$\mathcal{L}(z,K_j) = \sum_{k\in K_j} m_k \left(\langle z, v_k\rangle\right)^2 = \langle z, W_jDz\rangle$$

Nous cherchons une première variable générale telle que la somme des liaisons entre z_1 et les J groupes K_j soit maximum ; ainsi les directions d'inertie importante sont favorisées.

$$\sum_j \mathcal{L}(z_1,K_j) = \sum_{k\in K} m_k \left(\langle z_1, v_k\rangle\right)^2 = \langle z_1, WDz_1\rangle$$

$$= \sum_{k\in k} \text{inertie des projections des } v_k \text{ sur } z_1$$

Le vecteur normé z_1 qui rend maximum cette expression est un vecteur propre de WD associé à la plus grande valeur propre (égale à la somme des liaisons entre z et les K_j), soit la première composante principale du tableau complet X. Les variables générales suivantes sont les vecteurs propres de WD rangés dans l'ordre décroissant des autres valeurs propres, soit les autres composantes principales de X.

Remarques :

. Les opérateurs W_jD jouent ici exactement le rôle des opérateurs P_j dans l'Analyse Canonique Généralisée au sens de CARROLL.

. Les variables générales z_s sont obtenues en cherchant à rendre maximum la somme de leurs liaisons avec tous les groupes. Pour que ces groupes jouent un rôle analogue, les liaisons (z_s, groupe j) doivent a priori avoir le même intervalle de variation pour tous les groupes. Avec la pondération proposée, la liaison entre z_s et K_j est comprise entre 0 et 1. Le rôle des groupes est ainsi équilibré, en ce sens que la contribution de chacun dans le critère global à maximiser $\sum_j \mathcal{L}(z_s, K_j)$ est bornée par 1.

. Pour chaque groupe K_j et chaque variable générale z_s, la quantité $\mathcal{L}(z_s, K_j)$ sert d'aide à l'interprétation.

4.2-2 Les variables canoniques dans l'Analyse Factorielle Multiple

Les variables canoniques expriment dans chaque groupe la direction "commune" qu'est la variable générale. En Analyse Canonique classique, un groupe de variables j est représenté par le sous-espace E_j qu'il engendre et par l'opérateur de projection associé P_j ; la variable canonique associée à une variable z est son image par P_j. En Analyse Factorielle Multiple, un groupe est caractérisé par W_jD et la variable canonique associée à une variable z est son image par W_jD. Nous montrons ci-

dessous que $W_j Dz$ extrait du groupe j une part de variance plus importante que la projection $P_j z$.

L'image de l'opérateur $W_j D$ est le sous-espace engendré par les variables du groupe j. Notons $l_1, \ldots, l_z$ une base orthonormée (pour D) de ce sous-espace composée de vecteurs propres de $W_j D$ ordonnés dans l'ordre décroissant des valeurs propres ($\lambda_1 \geq \lambda_2 \geq, \ldots \geq, \lambda_r \geq 0$). En notant x_s la $s^{\text{ième}}$ coordonnée de $P_j z$ dans cette base, $P_j z$ et $W_j Dz$ s'écrivent :

$$P_j z = x_1 l_1 + \ldots + x_r l_r$$

$$W_j Dz = \lambda_1 x_1 l_1 + \ldots + \lambda_r x_r l_r$$

Or, l'inertie des variables du groupe j suivant une direction quelconque z, vaut $z' D W_j Dz$ si z est normé. (cf.§ 2.3.2.) Cette inertie est grande si les coordonnées de z sur les premiers vecteurs propres de $W_j D$ sont grandes. Ainsi, $W_j Dz$ correspond à une direction de plus grande inertie que $P_j z$ (sauf dans le cas extrême où $P_j z$ est colinéaire à un vecteur propre, auquel cas $P_j z$ et $W_j Dz$ ont des directions identiques).

Remarque :

Les variables générales forment une base orthonormée de R^I. Les variables canoniques du groupe j sont leurs images par $W_j D$ et définissent exactement cet opérateur qui caractérise ce groupe.

4.2-3 La représentation des individus

Les variables générales permettent la représentation d'une structure moyenne des individus. Cette représentation coïncide avec celle du nuage moyen proposé dans R^K.

Nous montrons ci-dessous que les variables canoniques du paragraphe précédent coïncident, à la norme près, avec les projections des N_I^j dans la représentation simultanée : la solution proposée pour l'analyse des liaisons satisfait aux objectifs exprimés au §1.7.

Soit u_s l'axe d'inertie d'ordre s du nuage d'individus N_I associé au tableau X dans R^K. Il se déduit de la composante princi-

pale F_s par la relation : $u_s = (1/\lambda_s)X'DF_s$ dans laquelle λs est la valeur propre de WD associée à F_s. La projection de N_I^j sur u_s s'écrit (cf. §3.5) : $F_s^j = \tilde{X}_j M u_s = (1/\lambda_s)\tilde{X}_j M(X'D F_s) = (1/\lambda_s)W_j D F_s$.

4.2-4 Le cas des variables qualitatives : lien avec l'Analyse des Correspondances Multiples

Si chaque groupe de variables représente une variable qualitative, l'opérateur $W_j D$ est confondu avec l'opérateur de projection P_j (cf. §2.5-3). L'Analyse Factorielle Multiple se confond alors avec l'Analyse Multicanonique et l'Analyse des Correspondances Multiples.

4.3 La dualité généralisée à plusieurs groupes de variables (lien entre l'Analyse Canonique Généralisée, la recherche d'une représentation simultanée et le modèle INDSCAL)

L'identité des résultats de l'analyse des liaisons et de la représentation simultanée des nuages N_I^j est une forme - limitée - de la dualité, généralisée à plusieurs groupes de variables.

4.3-1 Rappel de la dualité pour un groupe

Les résultats rappelés ci-dessous pour le nuage de variables N_K situé dans R^I et le nuage d'individus N_I situé dans R^K valent pour tout couple de nuages définis par le même tableau, en particulier N_K^j et N_I^j. Le poids des variables du nuage N_K définit la métrique diagonale M de R^K et celui des individus de N_I définit la métrique D de R^I. Soit u un vecteur normé de R^K ; la projection du nuage N_I sur u définit une fonction numérique $\mathcal{F}$ sur I, (les coordonnées de N_I sur u) : $\mathcal{F}$=XMu. On est ainsi conduit à considérer l'application XM de R^K dans R^I. De même, la projection δ=X'Dz du nuage N_K sur un vecteur de R^I définit l'application X'D de R^I dans R^K. Or, les métriques D et M définissent des isomorphismes de R^I et de R^K dans leur dual noté $(R^I)^*$ et $(R^K)^*$. Il est donc naturel de considérer X comme une application de $(R^K)^*$ dans R^I et X' comme une application de $(R^I)^*$ dans R^K. Le schéma ci-après (cf.[2]) résume l'ensemble de ces applications auquel on ajoute l'application W=X'DX de R^{I*} dans R^I.

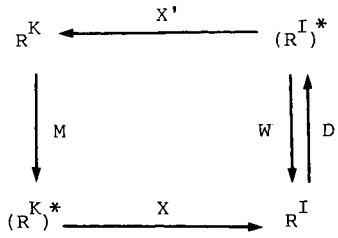
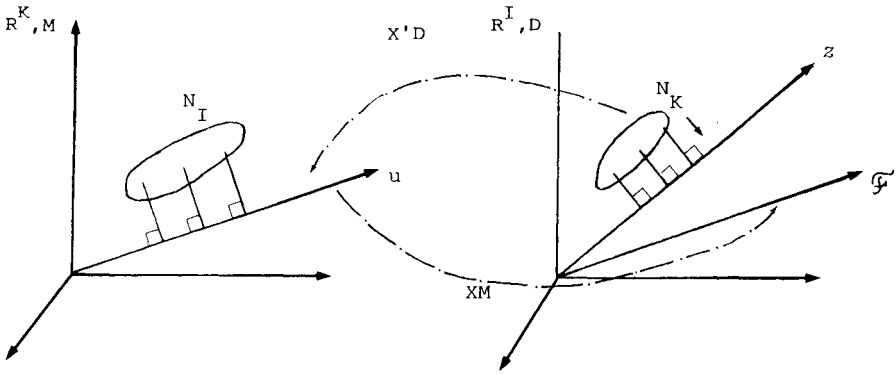


Schéma de dualité

Si l'on cherche une projection de N_I, $\mathcal{F}$ = XMu, d'inertie maximum et une projection δ=X'Dz de N_K d'inertie maximum, on obtient pour u et z les premiers axes d'inertie de N_I et de N_K qui se déduisent l'un de l'autre par les applications XM et X'D. Les projections de N_I et de N_K ont alors la même inertie. (Ces résultats remarquables ne sont pas vérifiés dans les autres directions : l'inertie de la projection $\mathcal{F}$ de N_I sur u est généralement différente de l'inertie de la projection de N_K sur l'axe directeur de $\mathcal{F}$; par ailleurs, si l'on applique XM (projection de N_I sur u) au vecteur X'Dz (projection de N_K sur z), le vecteur XMX'Dz n'est colinéaire à z que si z est vecteur propre de XMX'D, i.e. axe d'inertie de N_I).



δ = X'Dz=projection de N_K sur z

u = $\delta / ||\delta||$

$\mathcal{F}$ =XMu=projection de N_I sur u

=λz si z axe d'inertie de N_K

Les problèmes duaux de recherche de projections d'inertie maximum du nuage des variables et du nuage des individus aboutissent à une solution basée sur la diagonalisation d'une même matrice, où les axes sont liés par les applications XM et X'D. Les inerties projetées sont égales : il y a dualité des résultats.

4.3-2 Cas de plusieurs groupes de variables

L'analyse des liaisons d'une part, la représentation simultanée et les facteurs du modèle INDSCAL d'autre part, sont des problèmes duaux, i.e analogues mais exprimés l'un sur les variables et les autres à travers les nuages d'individus.

4.3-2 1 Projection des N_I^j et variables canoniques

La représentation simultanée vise à mettre en évidence des projections des J nuages N_I^j qui se "ressemblent" entre elles. La projection de N_I^j sur un axe u de R^{K_j} définit une variable $X_j M_j u$ de R^I, combinaison linéaire des variables du groupe j (comme une variable canonique). Si l'on adopte comme critère de ressemblance entre des projections axiales des N_I^j la corrélation entre ces variables, la représentation simultanée se traduit dans R^I par la recherche de combinaisons linéaires de variables de chaque groupe corrélées entre elles. C'est exactement la formulation classique de l'Analyse Canonique Généralisée.

En Analyse Factorielle Multiple, les directions d'inertie importante sont privilégiées : dans la représentation simultanée, on cherche les projections F^j des nuages d'individus N_I^j qui extraient de ces nuages une part d'inertie importante ; en analyse de liaisons, on cherche des variables canoniques F^j qui expliquent une grande part de la variance des variables du groupe j, i.e telles que l'inertie projetée sur F^j du nuage des variables soit grande.

Ces deux problèmes sont des généralisations à l'étude conjointe de plusieurs groupes, des deux problèmes duaux de l'A.C.P. ; à la recherche de directions d'inertie maximum des nuages de variables et d'individus, on ajoute la contrainte de ressemblance entre les facteurs de tous les groupes.

Dans les résultats, la dualité n'est pas totale : les composantes F_j^s du nuage des variables N_K^j se déduisent des axes U_s^j des individus N_I^j par l'application $X_j M_j$ (cf.§3.6.), mais $X'_j D$ ne joue plus le rôle inverse ; l'inertie des deux nuages sur les axes associés n'est pas égale.

Remarque :

Le critère de ressemblance entre les projections des J nuages N_I^j indiqué ici (la corrélation) permet d'exprimer le problème de la représentation simultanée dans l'espace R^I. Sa résolution aboutit au même résultat que dans R^K, ce qui prouve qu'il n'est pas lié à ce cadre un peu artificiel (cf. [10]). Bien que les critères de ressemblance soient différents, ils conduisent aux mêmes résultats car ils ne sont pas employés seuls, mais au sein d'un compromis incluant la qualité de représentation de N_I^j.

4.3.-2 2 Variables générales - Facteurs d'INDSCAL- Projection du nuage moyen

Les variables générales représentent des directions "communes" aux nuages de variables. Dans le modèle INDSCAL, les facteurs représentent des projections "communes" des nuages N_I^j (cf.§1.8.) : dans ce modèle, les distances se décomposent selon des facteurs z_s communs à tous les nuages (suivant $d^2(i^j,l^j) = \sum_s q_s^j (z_s(i)-z_s(l))^2$) qui peuvent être considérés comme des projections sur des axes orthogonaux 2 à 2. Dans les deux cas, l'Analyse Factorielle Multiple s'attache à obtenir des directions exprimant un pourcentage d'inertie important.

Les recherches de directions "communes" aux J nuages de variables d'une part, et "communes" aux J groupes d'individus, d'autre part, sont deux problèmes duaux. C'est une généralisation du double problème de l'A.C.P. qui fait jouer aux groupes de variables le rôle des variables de l'A.C.P.

Par ailleurs, les objectifs assignés au modèle INDSCAL et à la représentation simultanée sont assez proches. Dans le premier, l'accent est mis sur les directions communes (correspondant aux variables générales de l'Analyse Canonique Généralisée) ; dans le second, sur

les projections propres à chaque nuage (correspondant aux variables canoniques. Mais, dans ce dernier, les projections du nuage moyen peuvent jouer le rôle de représentants des directions communes et dans l'A.F.M., nous les proposons comme solution d'INDSCAL. L'intérêt d'une solution unique est grand : l'interprétation des projections du nuage moyen comme facteur commun d'INDSCAL leur donne un éclairage supplémentaire et permet d'associer à la représentation simultanée des N_I^j une représentation des groupes par leur "poids" ; l'interprétation des projections simultanées des N_I^j comme des traductions des facteurs communs dans chaque groupe permet de mesurer l'adéquation du modèle INDSCAL pour chaque facteur, chaque nuage et chaque individu. Pour le facteur F_s et le nuage N_I^j, cette adéquation peut être mesurée globalement par la corrélation entre ce facteur et la projection F_s^j de N_I^j qui représente F_s dans N_I^j, et examinée pour chaque individu i en comparant $F_s(i)$ et $F_s^j(i)$.

Les projections du nuage moyen construit en A.F.M. sont des approximations des facteurs communs d'INDSCAL tout à fait satisfaisantes ; En effet, si les nuages N_I vérifiant le modèle, alors le nuage moyen le vérifie aussi (cf.[14]et[4]):

pour un nuage $\quad d_j^2(i,l) = \sum_s q_s^j \, (z_s(i)-z_s(l))^2$

pour le nuage moyen (cf.3.2.) $\quad d^2(i,l) = \sum_s d_j^2(i,l) = \sum_s (\sum_j q_s^j)(z_s(i)-z_s(l))^2$

Les composantes principales du nuage moyen, prises dans l'ordre décroissant de leur valeur propre, constituent alors une solution qui classe les facteurs par ordre d'importance décroissante. Si le modèle s'ajuste assez bien aux données, les formules ci-dessus sont presque vérifiées. Cette solution d'INDSCAL est justifiée et complétée dans R^{I^2}.

4.4 La comparaison des nuages de variables

4.4-1 La comparaison des variables

L'A.C.P. de l'ensemble des variables fournit une image simplifiée des corrélations inter et intra-groupe. Les composantes prin-

cipales rendent maximum l'inertie des projections de toutes les variables. L'inertie projetée de chaque nuage N_K^j peut être interprétée comme la contribution d'un groupe. La pondération des groupes (par $\alpha_j=1/\lambda_j^1$) équilibre leur influence en ce sens que la contribution d'un groupe à la construction d'un axe est bornée par 1. On retrouve ici l'idée selon laquelle un groupe doit influer sur d'autant plus d'axes qu'il est de grande dimension. (Par contre, une pondération égalisant l'inertie totale de chaque groupe lamine le rôle des groupes de grande dimension dans la construction des premiers axes).

Remarque :

La contribution ainsi définie du groupe K_j à la construction de l'axe s se traduit dans R^K comme le carré du cosinus de l'angle entre l'axe s et le sous-espace $R^K j$ (noté θ_s^j au §3.7-1).

En effet, en notant $G_s(k)$ la coordonnée de la variable v_k sur l'axe s (dans R^I) et $u_s(k)$ la $k^{\text{ième}}$ composante de l'axe s dans R^K on a $(G_s(k)=\sqrt{\lambda_s}u_s(k))$ et :

$$\text{contribution de } K_j = \frac{1}{\lambda_s} \sum_{k\in K_j} m_k\{G_s(k)\}^2 = \frac{1}{\lambda_s} \sum_{k\in K_j} m_k\{\sqrt{\lambda_s}\, u_s(k)\}^2$$

$$= \sum_{k\in K_j} m_k\{u_s(k)\}^2 = \| u_s^j \|^2 = \{cos\ \theta_s^j\}^2$$

Enfin, cette contribution se confond aussi avec la mesure de liaison entre une composante principale et un groupe.

4.4-2 La comparaison des composantes principales de chaque groupe

Elle peut être réalisée à l'aide d'une A.C.P. de ces composantes. Pour qu'une telle analyse accorde plus d'importance aux composantes de grande inertie, on peut affecter à chaque composante normée un poids égal à son inertie. Ici encore, la pondération des groupes joue un rôle de "normalisation" en rendant égal à 1 le poids de la première composante principale de chaque groupe.

Cette A.C.P. des composantes principales des groupes est équivalente à l'A.C.P. de l'ensemble des variables. En effet, en notant H_j la matrice des composantes normées du groupe j et Δ_j la matrice diagonale des valeurs propres associées, on a : $W_j D = H_j \Delta_j H_j' D$. Les deux A.C.P. citées, qui conduisent à diagonaliser respectivement $\sum_j W_j D$ et $\sum_j H_j \Delta_j H_j' D$, sont équivalentes.

Ainsi, pour comparer les composantes principales des groupes, il suffit de les introduire en éléments supplémentaires dans l'analyse du tableau complet. On peut calculer en outre, situation paradoxale pour un élément supplémentaire, la contribution d'une composante d'un groupe à la construction des axes, par l'indicateur usuel.

On peut aussi adopter la démarche inverse : A.C.P. des composantes principales avec les variables en supplémentaire. En ne conservant que les premières composantes de chaque groupe, on obtient des résultats approchés satisfaisants (cf. [8]) nécessitant un temps calcul et une taille mémoire bien moindre, ce qui permet le traitement de très grands tableaux. La précision de l'approximation dépend de l'inertie maximum d'une composante abandonnée. On peut, par exemple, fixer le nombre maximum de composantes à conserver et éliminer celles dont l'inertie est la plus faible sans se préoccuper de leur appartenance aux groupes.

5. L'ANALYSE FACTORIELLE MULTIPLE DANS R^{I^2}

Dans l'espace R^{I^2}, chaque groupe de variables K_j est représenté par un vecteur noté W_j. Ce vecteur peut être considéré comme le tenseur associé au nuage N_K de R^I ou comme la matrice de produits scalaires associés au nuage N_I de R^{K_j} (cf. §2.3).

5.1 La comparaison globale des groupes et le modèle INDSCAL

Le produit scalaire entre W_j et W'_j est une mesure de liaison entre les groupes de variables j et j'. Pour comparer globalement les groupes, nous cherchons à décrire les proximités entre les W_j en les projetant sur un espace de faible dimension de R^{I^2}. (Dans la méthode STATIS, ce problème est appelé étude de l'interstructure). Les angles entre tenseurs doivent être bien représentés et il ne convient pas de centrer le nuage des W_j.

En exigeant uniquement une bonne qualité de représentation des W_j, on est conduit à une projection du nuage N_J sur ses axes d'inertie, c'est-à-dire une A.C.P., les vecteurs W_j étant considérés comme des variables (cf. STATIS). L'inconvénient de ce type d'analyse est de fournir un repère constitué d'axes difficilement interprétables. C'est pourquoi, nous imposons aux axes du repère, d'être des tenseurs symétriques de rang 1. Ces tenseurs de la forme $z_s \otimes z_s$ ou $(z_s z'_s)$ sont associés à des groupes d'une seule variable z_s et s'interprètent à partir de leurs liaisons avec les variables initiales.

La comparaison globale des groupes de variables, ainsi formulée comme une projection du nuage des W_j sur des tenseurs symétriques de rang 1, rejoint le point de vue du modèle INDSCAL. En effet, selon ce modèle (cf. §1.8), les distances entre individus définis par chacun des groupes de variables se décomposent suivant un certain nombre de "facteurs" communs à tous les groupes (les facteurs communs normés sont notés : $z_s, s=1,S$). Chaque groupe j affecte un "poids" q_s^j au facteur z_s et le modèle s'écrit :

$$(1) \qquad d^2(i^j, l^j) = \sum_s q_s^j (z_s(i) - z_s(l))^2$$

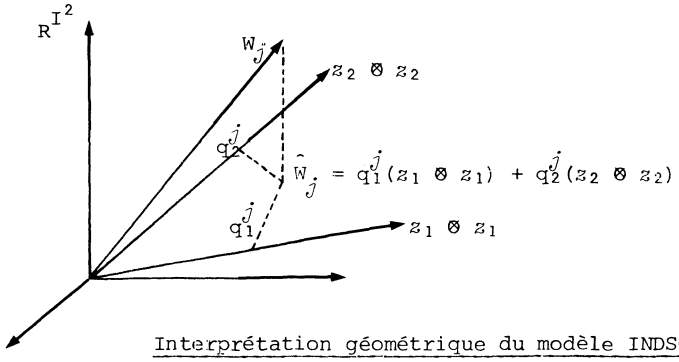
De cette écriture, on déduit une formulation en termes de produits scalaires :

$$(2) \qquad W_j(i,l) = \langle i,l \rangle_j = \sum_s q_s^j \, z_s(i) \, z_s(l)$$

soit matriciellement :

$$(3) \qquad W_j = \sum_s q_s^j \, z_s \, z_s'$$

La formule (3) traduite dans R^{I^2} exprime que les W_j sont décomposés sur un même repère composé de tenseurs symétriques de rang 1. Le poids q_s^j est la coordonnée de W_j sur l'élément $z_s \otimes z_s$ de ce repère. Chercher des paramètres z_s et q_s^j qui ajustent le modèle revient à chercher dans R^{I^2} un repère de ce type qui ajuste le nuage des W_j.



Interprétation géométrique du modèle INDSCAL

Partant de préoccupations très différentes, les deux points de vue aboutissent à la même formalisation dans R^{I^2}.

5.2 L'ajustement du nuage N_J

Nous cherchons un repère orthonormé dans R^{I^2}, de la forme $z_s \otimes z_s$ qui "ajuste" au mieux le nuage des W_j. Nous construisons ce repère progressivement en cherchant un premier vecteur, puis un second orthogonal au premier et ainsi de suite... Usuellement, on utilise le critère d'ajustement des moindres carrés, selon lequel on rend maximum la somme des carrés des projections des vecteurs du nuage. Du fait de la contrainte imposée aux vecteurs de base du repère,

c'est la somme des projections i.e. la somme des poids dans le modèle INDSCAL, et non de leur carré que nous maximisons.

Ce critère est plus facile à mettre en oeuvre que celui des moindres carrés (choisi souvent pour les facilités de calcul qu'il implique) et possède une signification puisque les coordonnées des W_j sur des tenseurs de type $z_s \otimes z_s$ sont toujours positives. En effet, la somme des projections de W_j sur $z_s \otimes z_s$ qui s'écrit $\sum_j <W_j, z_s \otimes z_s>$ est égale à l'inertie des variables (de tous les groupes) projetées sur z_s (cf.§ 2.5-2). La suite de tenseurs orthonormés qui maximisent cette somme est celle qui est associée aux composantes principales du tableau X, l'orthonormalité des z_s dans R^I étant équivalente à celle des $z_s \otimes z_s$ dans R^{I^2}. Les calculs nécessités par l'analyse dans R^{I^2} se déduisent directement des résultats de l'A.C.P. de X : les facteurs sont les composantes principales normées et le poids q_s^j la contribution du groupe j à l'inertie de la composante z_s.

Remarques :

Cette analyse dans R^{I^2} aurait pu être posée a priori en décidant de projeter le nuage N_J sur les tenseurs associés aux composantes principles de N_K. Mais, dans une telle présentation, la représentation des groupes apparaît seulement comme une "aide à l'interprétation". Or, elle est optimale en elle-même.

Cette estimation des facteurs du modèle INDSCAL correspond à celle proposée dans R^I. Une fois ces facteurs obtenus, les poids $\{q_j; s=1,S\}$ sont estimés par projection des W_j, c'est-à-dire minimisation du critère $\| W_j - \hat{W}_j \|$. Lorsque l'on considère un sous-espace engendré par S facteurs, ce n'est pas la somme des longueurs des projections des W_j qui est maximisé mais la somme des poids : $\sum_{s=1}^{S} \sum_{j=1}^{J} q_s^j$
(Cette somme des poids s'interprète dans R^I comme l'inertie du nuage N_K projeté sur les S facteurs).

5.3 La représentation graphique des W_j

La coordonnée de W_j sur l'axe factoriel $z_s \otimes z_s$ s'interprète aussi comme a) le poids q_s^j du modèle INDSCAL, b) la contribution absolue du groupe j à l'inertie de la composante principale z_s du tableau X, c) la mesure de liaison entre le groupe j et la variable générale z_s de l'analyse multicanonique. Les graphiques doivent être étudiés selon ces diverses optiques.

.../...

Du fait de la pondération des groupes, les coordonnées des W_j sont comprises entre 0 et 1 (sur un plan, W_j est toujours situé dans un carré de côté 1). Une coordonnée de W_j sur $z_s \otimes z_s$ voisine de 1 implique que a) le facteur z_s du modèle INDSCAL est présent (et important) dans le groupe j, b) le groupe contribue beaucoup à la détermination de la composante z_s, c) la direction commune z_s apparaît dans le nuage N_K^j (analyse multicanonique). Un groupe, selon la dimension des nuages associés, peut avoir plusieurs coordonnées proches de 1.

Les coordonnées des W_j étant comprises entre 0 et 1, les poids q_s^j du modèle INDSCAL sont comparables d'un groupe à l'autre, d'un axe à l'autre et même d'une analyse à l'autre. La somme de ces poids, pour un axe z_s , mesure l'importance de cet axe. Cette somme est la valeur propre associée à z_s dans l'A.C.P. du tableau X : l'Analyse Factorielle Multiple exhibe les facteurs du modèle INDSCAL par ordre d'importance décroissante.

Du fait de la contrainte sur les axes du repère (tenseurs symétriques de rang 1), la qualité de représentation des W_j par ces axes (mesurée au travers du critère usuel : inertie projetée/inertie totale) n'atteint en général pas 1, même si l'on augmente le nombre d'axes (qui atteint au plus I pour un espace de dimension I^2). En effet, projeter les W_j sur de tels axes revient à les approcher dans le cadre du modèle INDSCAL qui peut être plus ou moins adéquat aux données. Ces approximations correspondent à des nuages homothétiques axe par axe du nuage moyen.

6. ELEMENTS SUPPLEMENTAIRES

6.1 Individus et variables supplémentaires

Comme dans toute A.C.P., des individus peuvent intervenir en tant qu'éléments supplémentaires, c'est-à-dire avec un poids nul. Ces individus n'influent pas sur les représentations des individus actifs : on calcule simplement la projection de leur représentant dans le nuage N_I^* et dans les différents nuages N_I^j.

Dès que le nombre d'individus est assez grand, la lecture des graphiques des représentations simultanées est très complexe. En effet, le nombre des seuls points concernant les individus est égal à cardI × (cardJ+1). (i.e. nombre d'individus multiplié par le nombre de groupes de variables, plus 1 pour le nuage moyen). La lecture des aides à l'interprétation (stabilité de chaque individu, ressemblance globale entre la représentation d'un N_I^j et la représentation moyenne, etc...) facilite beaucoup le dépouillement. Mais il reste souvent nécessaire de remplacer l'étude de chaque individu par l'étude de classes d'individus ayant un caractère commun. Pour cela, on introduit en éléments supplémentaires les centres de gravité de ces classes.

Les variables supplémentaires sont traitées comme en A.C.P. : elles sont projetées sur les axes de l'analyse et leur qualité de représentation est calculée.

6.2 Groupes de variables supplémentaires

Tout un groupe de variables peut être mis en élément supplémentaire. Si ce groupe est homogène, il peut être intéressant de le comparer aux autres groupes avec tous les moyens mis en oeuvre pour ces derniers sans qu'il ait influé sur le nuage moyen et les résultats de l'analyse. La plupart des calculs (mais pas tous) effectués sur les groupes principaux s'appliquent à un groupe supplémentaire.

. La normalisation du nuage N_I^j : Pour comparer aux autres nuages le nuage associé au groupe supplémentaire, il faut le normaliser de la même façon en surpondérant les variables du groupe.

. La projection des composantes principales du groupe : Elle permet de comparer la forme générale du nuage de variables avec celle du nuage moyen et des autres nuages.

. Les variables canoniques du groupe : Dans l'analyse des liaisons, les variables générales sont calculées sans tenir compte du groupe supplémentaire : on peut cependant chercher les variables de ce groupe les plus "liées" à ces variables générales.

. Le nuage N_I^j dans la représentation simultanée : On peut chercher à projeter le nuage N_I^j sur les axes u_s de R^K. Ici, quelques difficultés apparaissent : l'ensemble K_j des variables supplémentaires n'est pas contenu dans l'ensemble K des variables actives, et R^{K_j} n'est donc pas contenu dans l'espace R^K où le nuage moyen a été construit.

Dans l'espace R^{KUK_j}, le nuage N_I^j est représenté, comme les nuages actifs, dans le sous-espace R^{K_j}. Mais, le nuage moyen N_I^*, construit en considérant seulement les groupes de variables actifs, (ou en donnant un poids nul aux variables du groupe supplémentaire) est contenu dans l'orthogonal du sous-espace R^{K_j}. Ses axes d'inertie (u_s) le sont aussi et les projections de N_I^j sur ces axes valent 0. C'est ce que l'on obtient, procédant au même calcul que pour les autres tableaux, en considérant comme individus supplémentaires les lignes du tableau X_j complété par des zéros.

On peut songer à utiliser comme facteur, la variable canonique $F_S^j = (\sqrt{\lambda_S}) W_j . Dz_s$, qui est, à un coefficient près, la projection de N_I^j sur un axe $\bar{u}_s$ de R^{KUK_j}. Mais cet axe est distinct de l'axe u_s sur lequel les autres nuages N_I^j et N_I^* sont projetés. En outre, les axes $\bar{u}_s$ ne sont pas orthogonaux entre eux.

Il semble que la présence, dans la représentation simultanée, d'un nuage associé à un groupe de variables supplémentaire n'admette pas de solution satisfaisante. Cette difficulté n'est sans doute pas sans rapport avec les raisons qui conduisent à introduire un groupe en supplémentaire. Si on craint qu'il ne perturbe les résultats, car présentant a priori de grandes différences avec les autres groupes, les indices globaux, projections des composantes principales, etc... permettent de mesurer et préciser ces différences, mais superposer le nuage N_I^j à des nuages qui ne lui ressemblent pas assez, n'a pas d'intérêt. S'il intervient uniquement en tant qu'élément explicatif lors de l'interprétation, on s'intéresse alors aux liaisons entre les variables de ce groupe et les autres et non à chaque individu.

. Projection des W_j : La présence d'éléments supplémentaires dans l'analyse du nuage N_J dans R^{I^2} ne pose pas de problème particulier. La coordonnée d'un élément supplémentaire W_j sur l'axe $z_s \otimes z_s$ coïncide avec la mesure de liaison entre le $s^{ième}$ facteur et le groupe j, le poids affecté à ce facteur par le groupe j dans le modèle INDSCAL et l'inertie des variables du groupe j le long de la direction s (qui ne s'interprète plus comme une contribution).

7. COMPARAISON AVEC D'AUTRES METHODES

L'Analyse Factorielle Multiple est comparée à chacune des méthodes évoquées dans le premier paragraphe en choisissant pour chacune l'aspect de l'Analyse Factorielle Multiple qui se rapporte au même objectif.

7.1 Analyse en Composantes Principales du tableau X

L'Analyse Factorielle Multiple s'appuie sur une A.C.P. du tableau complet, mais s'en distingue par : une pondération des variables qui équilibre le rôle des groupes ; une représentation simultanée des nuages d'individus associés à chaque groupe de variables ; une représentation des groupes de variables ; une interprétation en termes d'Analyse Canonique Généralisée ; une interprétation en termes de modèle INDSCAL ; des indices d'aide à l'interprétation spécifiques complétant les points précédents.

Dans le cas limite où chaque groupe est réduit à une seule variable numérique, l'Analyse Factorielle Multiple, compte tenu de la pondération décrite en 2.3-5, coincide avec l'Analyse en Composantes Principales normée.

7.2 Les analyses canoniques généralisées

L'analyse des liaisons dans l'A.F.M. a été présentée au §4.2 comme une variante de l'Analyse Multicanonique au sens de CARROLL dans laquelle l'opérateur de projection P_j est remplacé par l'opérateur $W_j D$. L'utilisation de $W_j D$ augmente la variance expliquée ; rend les résultats moins sensibles à de petites variations des données ; facilite considérablement l'interprétation des résultats qui s'appuient sur une A.C.P. classique et se traduisent par des représentations graphiques de projections des nuages de variables et de nuages d'individus N_I^j.

Remarque :

L'Analyse Canonique Généralisée au sens de CARROLL peut

aussi être considérée comme une A.C.P. du tableau complet X à condition de munir les espaces $R^K{}_j$ de la métrique de MAHALANOBIS $M_j = (X_j' \; D \; X_j)^{-1}$ et R^K de la métrique bloc-diagonale induite par les M_j. Outre le fait qu'en procédant ainsi la dualité entre l'analyse du nuage des individus et celle du nuage des variables est perdue (il faudrait pour la conserver introduire le concept de poids de couples de variables, puisque M n'est pas diagonale), ce point de vue rend ardue l'interprétation des résultats : le nuage d'individus est "déformé" par la métrique M, et la pondération des couples de variables discutable. Les variables canoniques peuvent, comme dans l'A.F.M., s'interpréter dans R^K comme des projections des nuages N_I^j sur les axes d'inertie d'un nuage moyen, mais ces nuages, rendus sphériques (i.e. tous leurs moments d'inertie sont égaux) par la métrique de MAHALANOBIS, présentent peu d'intérêt.

7.2-1 Cas où tous les groupes sont identiques

Dans ce cas, toute base orthogonale du sous-espace engendré par les variables est solution de l'analyse de CARROLL, à la fois en tant que variables générales et canoniques. Par contre, pour l'A.F.M., les variables générales et canoniques se confondent avec les composantes principales du tableau complet, prises dans l'ordre décroissant de leur inertie, et de ce fait représentent mieux les variables.

7.2-2 Cas de deux groupes

Dans le cas de deux groupes, l'Analyse Canonique Généralisée de CARROLL se confond avec l'Analyse Canonique classique : les variables canoniques d'un même groupe sont orthogonales 2 à 2 et s'obtiennent directement en diagonalisant le produit (à gauche ou à droite) des opérateurs $P_1^{E^2}$ et $P_2^{E^1}$ où $P_1^{E^2}$, par exemple, désigne la restriction de P_1 au sous-espace E_2.

On peut chercher à exploiter le cas particulier de deux groupes (c'est-à-dire obtenir directement les variables canoniques sans passer par l'intermédiaire de variables générales) avec les idées de l'Analyse Factorielle Multiple (variables canoniques correspondant à des directions

d'inertie importante. On est alors conduit à diagonaliser le produit $W_1 D W_2 D$; ici encore, l'opérateur $W_1 D$ remplace l'opérateur de projection P_1 pour représenter le 1er groupe de variables. (Ce travail est réalisé dans le mémoire de DEA de Michel LORENTER préparé à l'E.N.S.A. de Rennes et soutenu à l'U.E.R. d'Orsay (septembre 1983)).

Cependant, les variables canoniques ainsi obtenues ne sont pas en général identiques à celles de l'Analyse Factorielle Multiple. Ce résultat semble a priori négatif : en fait, l'identité de l'Analyse Canonique de CARROLL et l'Analyse Canonique classique dans le cas de deux groupes tient à l'idempotence des projecteurs, c'est-à-dire au caractère fruste des opérateurs représentant les groupes de variables. Quoiqu'il en soit, une telle identité n'est pas essentielle puisque les variables canoniques sont des représentants des directions communes permettant une représentation simultanée. Or, les axes de R^I issus de la diagonalisation du produit $W_1 D W_2 D$ ne sont pas orthogonaux, bien que les variables canoniques le soient et ils s'interprètent mal en terme de représentation simultanée. En outre, les variables générales possèdent un intérêt propre et doivent être calculées.

Pour deux groupes, l'A.F.M. se rapproche de la technique proposée par WOLLENBERG [cf. 27] comme préférable à l'analyse canonique. Remarquant que dans l'analyse canonique, la variance expliquée par les variables obtenues risque d'être faible, il cherche une suite de variables orthogonales deux à deux, combinaisons linéaires de variables du premier groupe, maximisant la somme des carrés des corrélations avec les variables du second groupe. On reconnait ici un critère analogue à celui de l'A.F.M. qui offre en plus la possibilité de pondérer variables et individus . Mais ce critère est appliqué directement sans introduction de variables générales. On est alors conduit à la diagonalisation d'un opérateur qui, dans nos notations s'écrit $P_1 W_2 D$. Le problème est posé de manière symétrique pour les deux groupes, mais il n'y a pas de symétrie des résultats : les valeurs propres de $P_1 W_2 D$ et $P_2 W_1 D$ sont différentes et il n'existe pas de relation simple entre leurs vecteurs propres.

L'A.F.M. peut être vue comme une généralisation de la méthode de WOLLENBERG parallèle à la généralisation de l'analyse canonique proposée par CARROLL.

7.3 Analyse en Composantes Principales des opérateurs

Le volet de l'A.F.M. ayant R^{I^2} comme espace de référence peut se comparer à ces analyses. Les représentations des groupes de variables sont isomorphes et dans les deux cas, sont projetées dans un espace de petite dimension. L'ajustement du sous-espace est meilleur dans l'analyse des opérateurs et les groupes de variables sont donc globalement mieux représentés. Mais, dans l'A.F.M., les composantes sont plus aisément interprétables du fait des représentations des individus et des variables associées, alors que l'analyse des opérateurs ne permet de conclure que sur la plus ou moins grande proximité des opérateurs entre eux, sans aucun élément d'interprétation. De plus, un indice unique est beaucoup trop pauvre pour exprimer la ressemblance entre deux structures multidimensionnelles. Une "bonne représentation" de ces indices présente beaucoup moins d'intérêt que leur décomposition suivant les mêmes structures unidimensionnelles (cf. interprétation d'INDSCAL).

7.4 La méthode STATIS et ses variantes

Cette méthode et les autres très proches proposées par Y. ESCOUFIER [cf.12] et M.C. PLACE [cf. 23] ont pour but en particulier :

. de construire un nuage "compromis" entre les nuages associés aux différents tableaux (ce sont dans notre cas les nuages d'individus N_I^j définis par chaque groupe de variables) ;

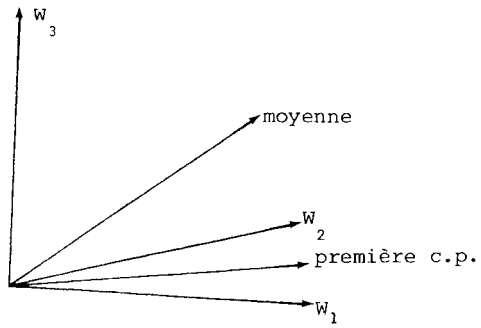
. de représenter simultanément ces nuages N_I^j sur des espaces de petite dimension en prenant pour base de départ le nuage compromis.

STATIS et l'A.F.M. diffèrent d'une part dans le choix du nuage compromis, d'autre part dans le choix des représentations des N_I^j associés à celle du compromis.

7.4-1 Choix du nuage compromis

Dans l'A.F.M., le nuage compromis est le nuage d'individus associé à toutes les variables. Dans STATIS, il se définit à partir d'un élément de R^{I^2}, une combinaison linéaire des W_j (ou W_jD) à coefficients positifs. Ces solutions sont comparables car : à l'union de 2 groupes de variables j et j' correspond la somme des tenseurs W_j et W'_j ; au groupe j surpondéré par β_j correspond le tenseur $\beta_j W_j$, et à l'ensemble des groupes surpondérés par β_j correspond donc $\sum_j \beta_j W_j$; réciproquement, à toute combinaison linéaire $\Sigma_j \beta_j W_j$ avec $B_j > 0$, on peut associer le nuage d'individus défini par les groupes surpondérés (par les β_j).

Au nuage moyen défini dans l'A.F.M., correspond la somme des W_j, tandis que le nuage compromis de STATIS est le nuage associé à la première composante principale de l'analyse des W_jD (cf. §8.3). Ces deux solutions sont raisonnables car les W_j étant situés dans un cône, leurs produits scalaires deux à deux sont positifs. De ce fait, a) l'ajustement des W_j par leur somme a un sens (il n'y a pas le risque de voir se neutraliser deux vecteurs opposés ; b) la première composante principale est une combinaison linéaire des W_j à coefficients positifs. La somme, (comme la moyenne), représente l'ensemble des groupes tandis que l'ajustement aux moindres carrés risque de privilégier un sous-ensemble des groupes de variables aux dépends des autres.



Par ailleurs, la somme présente l'avantage de la simplicité tant du point de vue des calculs que de celui de l'interprétation.

L'écart entre les solutions de l'A.F.M. et de STATIS provient beaucoup plus de la "normalisation" préalable que du mode de calcul du compromis. Dans STATIS, les vecteurs $W_j D$ sont normés dans R^{I^2}, (les N_I^j ont alors tous une inertie totale égale à 1). Un groupe de grande dimension influence peu les premières composantes principales du nuage compromis et, a contrario, un groupe très redondant risque de déterminer à lui seul ces composantes. Dans l'A.F.M., les variables du groupe j sont surpondérées par $1/\lambda_j$, ce qui équilibre l'influence des groupes dans chaque dimension du compromis. (Le nombre de dimensions du compromis sur lesquelles un groupe influe varie avec la dimension du nuage associé au groupe).

7.4-2 Représentation simultanée liée au compromis

Dans les 2 méthodes, les représentations des N_I^j, (dites des intrastructures dans STATIS) se déduisent des composantes principales du nuage compromis. En notant F_s ces composantes (différentes dans les deux cas), z_t la $t^{ième}$ composante du groupe j, la représentation F_s^j du groupe j s'écrit [cf. 21] pour STATIS :

$$F_s^j = \sum_t \cos(F_s, z_t^j) z_t^j$$

Elle se déduit de F_s par un opérateur dont le carré est $W_j D$, tandis que dans l'A.F.M., elle s'en déduit par $W_j D$.

Dans les deux méthodes, F_s^j appartenant au sous-espace E_j peut être vue comme la projection , à une homothétie près, du nuage N_I^j sur un axe. Lorsque l'on considère plusieurs composantes simultanément, par exemple les représentations planes des N_I^j, ces représentations sont des projections dans l'A.F.M. et ne le sont pas dans STATIS. Le nuage compromis est au centre de gravité des N_I^j dans l'A.F.M., propriété qui n'est pas vérifiée dans STATIS. (sauf si $W_j D$ est égal au projecteur sur E_j, auquel cas $W_j D$ et $\sqrt{W_j D}$ sont égaux). Enfin, la représentation simultanée de l'A.F.M. possède une propriété optimale : elle réalise un compromis global entre la ressemblance et la qualité des représentations des N_I^j.

Dans les variantes de la méthode STATIS, Y. ESCOUFIER [12] et M.C. PLACE [23] proposent de projeter les premières composantes principales du groupe K_j sur le sous-espace engendré par les premières composantes principales du compromis.

Or, le compromis peut être considéré comme un nuage associé à l'ensemble de toutes les variables (surpondérées par les β_j). Le sous-espace engendré par ses composantes principales est donc le sous-espace engendré par toutes les variables et contient les composantes principales de chaque groupe. En retenant toutes les composantes du compromis (ce qui est peu réaliste), cette méthode revient à superposer les composantes principales de tous les groupes.

7.5 Le modèle INDSCAL

Diverses solutions ont été proposées pour estimer les paramètres z_s et q_s^j du modèle INDSCAL, $W_j = \sum_s q_s^j z_s z'_s$. Elles dépendent du critère à optimiser : minimiser la somme des écarts entre les carrés des distances dans le modèle et dans les données réelles (stress) (cf 20) ; son carré (cf 25) ; la somme des écarts entre les produits scalaires du modèle et des données (cf 4). Nous comparons l'A.F.M. à l'estimation la plus couramment utilisée proposée par CARROLL et CHANG (cf 4). Dans celle-ci, le critère minimisé s'écrit, en notant $\bar{W}_j$ les produits scalaires obtenus par le modèle, et $\|\ .\|$ la norme de l'espace R^{I^2} (cf §5).

$$V = \sum_j \| \bar{W}_j - W_j \|^2 = \sum_j \| \sum_s q_s^j z_s z'_s - W_j \|^2$$

Les z_s orthonormés obtenus, les poids q_s^j qui minimisent V sont, dans R^{I^2} les coordonnées des projections des W_j sur le système orthogonal des $z_s z'_s$. C'est la solution de l'A.F.M.. Les z_s qui minimisent V seraient obtenus par un ajustement au moindre carré des W_j, en maximisant la somme des carrés des projections des W_j. Dans l'A.F.M., c'est la somme des projections des W_j qui est maximisée : la solution numérique est simple (diagonalisation d'une matrice symétrique) et les poids q_s^j obtenus sont toujours positifs et interprétables (ce qui n'est pas toujours le cas dans la méthode courante). Les facteurs obtenus dans l'A.F.M. sont orthogonaux

deux à deux et les s-1 premiers facteurs de la solution à s facteurs constituent la solution à (s-1) facteurs. (Ceci n'est pas vérifié dans la méthode courante). Enfin, rappelons que, dans la méthode de CARROLL et CHANG, l'algorithme est itératif et converge lentement vers un optimum local.

Dans l'estimation usuelle, la normalisation des données revient à rendre égale à 1 la norme des W_j. Dans d'autres estimations (cf), la normalisation est en option. Dans l'A.F.M., la normalisation rend égale à 1 la valeur maximum de la projection de W_j vers un tenseur $z_s\ z'_s$. Les poids estimés sont toujours inférieurs à 1 et q_s^j s'interprète comme une mesure de la liaison entre le facteur z_s et le groupe j. Ainsi, les poids affectés à un même facteur ou à deux facteurs différents sont toujours comparables entre eux. Cette propriété caractéristique de la solution apportée par l'A.F.M. s'est révélée essentielle dans le dépouillement des résultats. Dans l'estimation usuelle, les facteurs z_s sont normalisés et l'importance d'un facteur est mesurée par la somme des carrés des poids $\sum_j (q_s^j)^2$. Dans l'A.F.M., les poids sont calculés pour des facteurs de norme 1 dont l'importance est mesurée par la somme des poids et le critère optimisé par les facteurs est celui qui permet de les classer. (Mais, dans la représentation graphique de la configuration moyenne, les facteurs ont une inertie (ou une norme carrée) égale à cette somme).

L'interprétation géométrique des q_s^j dans l'espace R^{I^2} comme coordonnée des projections des W_j sur $z_s\ z'_s$ et dans l'espace R^I comme inertie du groupe de variables j projetée sur z_s permet d'introduire très simplement des éléments supplémentaires dans la solution de l'A.F.M.. Mais, un des avantages essentiels de cette solution est d'être inclue dans une analyse complète : les résultats associés aux autres points de vue de l'A.F.M. peuvent être utilisés comme des aides à l'interprétation en permettant en particulier des mesures facteur par facteur et nuage par nuage de l'approximation donnée par le modèle. Réciproquement, l'estimation des paramètres du modèle INDSCAL peut jouer le rôle d'aide à l'interprétation des autres résultats.

7.6 Les variables qualitatives : Analyse des Correspondances Multiples

Une variables qualitative est, pour nous, l'ensemble des variables indicatrices d'une partition affectées de poids tels que l'opérateur $W_j D$ associé à ce groupe de variables soit confondu avec l'opérateur de projection sur le sous-espace engendré par les indicatrices (cf. 2.3-5).

7.6-1 Chaque groupe est une variable qualitative

Dans ce cas, les facteurs sur I de l'A.F.M., sont, en tant que variables générales de l'Analyse Multicanonique, confondues avec celles de l'Analyse Multicanonique de CARROLL (cf. § 4.2) ; et donc avec les facteurs de l'Analyse des Correspondances Multiples (cf. [8] et [19]). La variable canonique du groupe K_j, combinaison linéaire des indicatrices qui constituent ce groupe est une fonction sur l'ensemble des K_j classes. On montre que la restriction à K_j du facteur sur K des correspondances multiples se confond avec cette variable canonique.

La représentation graphique des facteurs de l'Analyse des Correspondances Multiples s'interprète donc comme une projection simultanée des nuages N_I^j et d'un nuage moyen : la restriction à K_j du facteur g_s de l'A.F.C.M. est, au coefficient $\sqrt{\lambda_s}$ près, la projection de N_I^j sur le $s^{ième}$ axe. (Dans N_I^j, les individus appartenant à la même classe sont confondus). Les représentations des classes dans les deux méthodes coïncident axe par axe au ceofficient $\sqrt{\lambda_s}$ (pour l'axe s) près. Dans l'A.F.M., un individu apparaît au barycentre des classes auxquelles il appartient ; dans l'A.F.C.M., cette propriété n'est vraie qu'aux coefficients $\sqrt{\lambda_s}$ près.

La valeur propre λ_s est l'inertie du nuage moyen. L'inertie de l'union des N_I^j vaut 1 sur chacun des facteurs. En effet, dans R^K, l'inertie de N_I^j vaut 1 dans toutes les directions de R^{K_j} ; sur un axe u_s de R^K, elle est égale au cosinus carré de l'angle θ_j entre u_s et R^{K_j} et $\sum_j \cos^2 \theta_j = 1$. La valeur $\lambda_s = \lambda_s / 1$, confondue avec le quotient de l'inertie inter de l'union des N_I^j et son inertie totale, mesure la ressemblance globale entre la projection de tous les N_I^j.

La contribution du groupe j à l'inertie λ_s du facteur F_s (multipliée par J) est le poids du modèle INDSCAL et la coordonnée du projecteur P_j sur le système d'axes de R^{I^2} ajustant ces projecteurs (et associé aux facteurs F_s). On retrouve la représentation des variables qualitatives proposées dans [7] comme aide à l'interprétation d'une Analyse des Correspondances Multiples.

L'interprétation des résultats classiques des correspondances multiples est un peu enrichie.

7.6-2 Chaque groupe comprend plusieurs variables qualitatives

L'analyse de chaque groupe est encore identique à l'Analyse des Correspondances Multiples, mais l'analyse du tableau entier ne l'est pas :

a) Les groupes sont surpondérés par l'inverse de la première valeur propre du groupe pour équilibrer leur influence et "normaliser" les nuages associés.

b) Une représentation simultanée des N_I^j est proposée. Dans chaque N_I^j, un individu i est représenté par le croisement des classes (des partitions) auxquelles il appartient. A l'intérieur de chaque groupe, ce sont les croisements des modalités appartenant à des variables différentes qui sont étudiés.

c) Les contributions de chaque groupe, poids du modèle INDSCAL,permettent de mesurer l'importance du facteur pour ce groupe.

7.6-3 Groupes de variables numériques et groupes de variables qualitatives : analyse de tableaux hétérogènes

L'Analyse Factorielle Multiple est un moyen d'appréhender des tableaux mixtes quantitatif-qualitatif en équilibrant le rôle des différents groupes. Elle permet de plus :

a) Une représentation simultanée des individus vus à travers des groupes de variables, soit quantitatives, soit qualitatives.

b) Une mesure de l'importance de chaque facteur pour chaque groupe. Si chaque groupe de variables numériques est réduit à une seule variable, on obtient les mêmes résultats que par la technique proposée dans [9], mais l'A.F.M. est à la fois plus souple et plus complète.

7.7 La pondération

La pondération n'est pas une méthode particulière, mais elle intervient quel que soit le point de vue adopté, et celle de l'A.F.M. diffère de celle de toutes les autres méthodes. Dans les méthodes évoquées précédemment, soit il n'y a aucune pondération préalable, soit il y a une pondération rendant égale à 1 l'inertie totale de chaque nuage. La pondération de l'A.F.M. qui rend égale à 1 l'inertie maximum du nuage dans une direction est originale. Elle diffère de la pondération par l'inertie totale si les nuages associés aux différents groupes ont des dimensions très variables. Elle tient compte de la structure multidimensionnelle des objets maniés et des techniques appliquées qui toutes, (A.C.P., INDSCAL, comparaison globale, représentation simultanée et analyse multicanonique), travaillent itérativement (dimension par dimension). Les critères optimisés s'expriment sur une seule dimension et la pondération choisie équilibre l'influence des groupes dans ce sens.

Cette pondération présente l'avantage de fixer à [0,1] l'intervalle de variation des poids du modèle INDSCAL (et des mesures de liaison entre facteurs et groupes de variables), ce qui les rend comparables d'un groupe à l'autre, d'un facteur à l'autre et d'une analyse à l'autre. Elle assure l'équivalence avec une A.C.P. normée (resp. l'Analyse des Correspondances Multiples) dans le cas de groupes réduits à une seule variable numérique (resp. qualitative).

8. UN PETIT EXEMPLE EN GUISE DE CONCLUSION

Ce chapitre décrit un exemple artificiel de très petite dimension destiné à fournir une vision synthétique de la plupart des résultats. Un exemple réel plus complexe, qui illustre l'intérêt de la méthode, est commenté dans [10] et [11].

8.1 Les données

Six individus sont étudiés au travers de trois groupes de variables. Les trois groupes sont engendrés par 2 variables orthogonales U et V.

7 variables en 3 groupes

		K_1			K_2			K_3
	Noms	U	V	-U	U	V	V	U
	A	1	1	-1	1	1	1	1
	B	1	-1	-1	1	-1	-1	1
6 individus	C	0	1	0	0	1	1	0
	D	0	-1	0	0	-1	-1	0
	E	-1	1	1	-1	1	1	-1
	F	-1	-1	1	-1	-1	-1	-1

Tableau des données

8.2 Les analyses partielles

L'Analyse Factorielle Multiple réalise, en premier lieu, les A.C.P. partielles de chaque groupe. Dans ces analyses, les variables sont centrées réduites et de poids 1.

Résultats des A.C.P. partielles

A.C.P. du groupe	1re valeur propre	1er facteur	2e valeur propre	2e facteur
$K_1\{U,V,-U\}$	$\lambda^1_1 = 2$	U	$\lambda^2_1 = 1$	V
$K_2\{U,V,V\}$	$\lambda^1_2 = 2$	V	$\lambda^2_2 = 1$	U
$K_3\{U\}$	$\lambda^1_3 = 1$	U	$\lambda^2_3 = 0$	-

Il en résulte que, dans l'analyse globale, les variables des groupes K_1 et K_2 seront pondérées par 1/2, et celle du groupe K_3 par 1.

8.3 L'analyse globale

	valeur propre	composante principale
1er axe	2,5	U
2e axe	1,5	V

Les valeurs propres sont toujours inférieures au nombre de groupes (J). Une valeur de 3 indiquerait que le facteur commun de l'A.F.M. coïncide avec les premières composantes principales de chaque groupe.

Les facteurs 1 et 2 se confondent respectivement avec les variables U et V.

8.4 Etude globale des groupes de variables

Le tableau suivant résume les principaux résultats.

	Facteur 1			Facteur 2			Facteurs 1 et 2
	coord.	qual.	corr.	coord.	qual.	corr.	qualité
K_1 : (U,V-U)	1	0.8	1	0.5	0.2	1	1
K_2 : (U,V,V)	0.5	0.2	1	1	0.8	1	1
K_3 : (U)	1	1	1	0	0	0	1
	2.5 (λ_1)			1.5 (λ_2)			

coord. (coordonnée) : projection sur un axe factoriel du point W_j, représentant le groupe K_j dans R^{I^2}. Pour un facteur, la somme des coordonnées est égale à la valeur propre.

qual. (qualité) : qualité de représentation, dans R^{I^2}, de W_j par un axe ou un plan factoriel.

corr. (corrélation) : corrélation entre F_s (variable générale de l'analyse multicanonique, projection de N_I^* et facteur commun d'INDSCAL) et W_j D F (variable canonique et projection de N_I^j).

La coordonnée des groupes K_1 et K_3 sur le premier facteur atteint la valeur maximum : 1. Cette coordonnée supporte plusieurs interprétations. En tant que contribution à l'inertie, elle indique que les groupes K_1 et K_3 ont contribué fortement et également à la construction du facteur commun. En tant que mesure de liaison entre le facteur commun et le groupe, elle indique ici que le premier facteur se confond avec la première composante principale des groupes K_1 et K_3.

Enfin, ces coordonnées peuvent être interprétées comme des poids du modèle INDSCAL : ce premier facteur joue, dans la représentation des individus (nuage N_I^j), un rôle beaucoup plus grand pour les groupes K_1 et K_3 que pour le groupe K_2.

La qualité de représentation (des W_j dans R^{I^2}) montre que le groupe K_3 ne possède aucune autre dimension que ce facteur. Il n'en est

pas de même du groupe K_1 : compte tenu de la pondération choisie (par $1/\lambda_j^1$), le fait, pour un groupe, de présenter une coordonnée de 1 sur un axe n'implique pas que ce groupe présente une coordonnée nulle sur tous les autres axes.

La corrélation entre (F_S et W_j D F_s) de 1 pour les groupes K_1 et K_2 découle de la coordonnée de ces groupes. Par ailleurs, la valeur de 1 pour le groupe K_2 montre que le facteur commun est présent dans ce groupe : toutefois, ce groupe possède au moins une autre dimension plus importante d'où la coordonnée de 0.5.

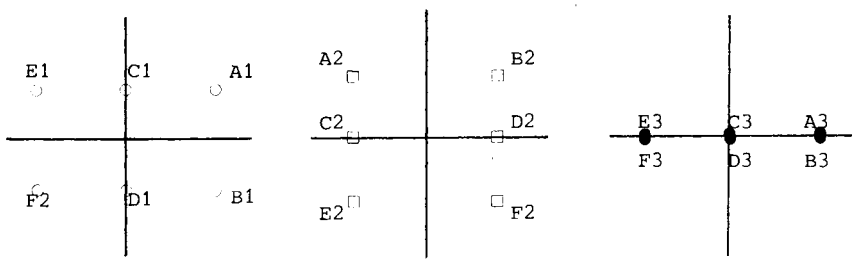
En résumé, ce premier facteur est également présent dans les 3 groupes (corrélations). Il correspond aux premières composantes principales des groupes K_1 et K_3, et à une dimension de moindre importance du groupe K_2 (coordonnées = 1 ou 0,5). Le groupe K_3 ne possède pas d'autre dimension. (qualité = 1).

Appliqué au deuxième axe, les mêmes raisonnements indiquent que ce facteur est présent dans les groupes K_1 et K_2 (corrélations = 1) et absent de K_3 (corrélations = 0). Il correspond à une première composante principale du groupe K_2 (coordonnée = 1) et à une direction de moindre importance de K_1 (coordonnée = 0.5).

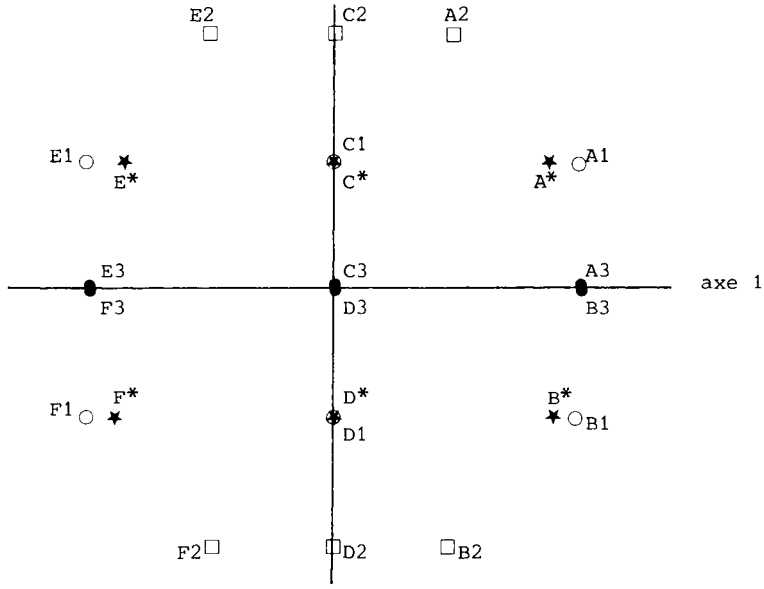
Enfin, la qualité de représentation des W_j sur le plan factoriel atteint 1 pour les 3 groupes. Dans R^{I^2}, les W_j peuvent être décomposés en une somme des mêmes tenseurs de rang 1 : le modèle INDSCAL est parfaitement vérifié. La représentation graphique des 3 nuages N_I^j se déduit de la représentation (par les deux facteurs U et V normés) du nuage moyen N_I^* par des homothéties dont les rapports sont les poids du modèle INDSCAL, c'est-à-dire les coordonnées des W_j.

8.5 La représentation simultanée

Les analyses partielles de chacun des trois groupes conduisent aux représentations suivantes.



La représentation simultanée de l'A.F.M. fournit un cadre de référence commun à ces trois configurations.



Cette représentation inclut un nuage moyen, représentation du nuage N_I^* fournie par les deux premiers facteurs non normés.

- Bibliographie -

Cette bibliographie ne prétend pas faire une revue complète des méthodes de traitement de plusieurs groupes de variables. Elle ne contient que les ouvrages ou articles apportant des compléments ou des précisions sur le texte.

[1] BENZECRI J.P. et coll. (1973)
L'analyse des données.
DUNOD, Paris.

[2] CAILLEZ F., PAGES J.P. (1976)
Introduction à l'analyse des données.
SMASH, Paris.

[3] CARROLL J.D.(1968)
A generalization of canonical correlation analysis to three or more sets of variables.
Proceedings of the 76th annual convention of the American Psychological association, p. 227-228.

[4] CARROLL J.D. et CHANG J.J. (1970)
Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart Young" decomposition.
Psychometrika, vol. 35 n°3, p. 283-319.

[5] CARROLL J.D. (1981)
INDSCAL. in SCHIFFMAN, REYNOLDS, YOUNG. Introduction to multidimensional scaling.
Academic Press INC.

[6] CAZES P., ESCOUFIER Y., PAGES J.P. (1976)
Opérateurs et analyse des tableaux à plus de deux dimensions.
Cahiers du Buro n° 25.

[7] ESCOFIER B. (1979)
Une représentation des variables dans l'analyse des correspondances multiples.
Revue de Statistiques Appliquées, 1979, n°4, p.

[8] ESCOFIER B. (1979)
Stabilité et approximation en Analyse Factorielle.
Thèse de Doctorat ès Sciences. Paris VI.

[9] ESCOFIER B. (1979)
Traitement simultané de variables qualitatives et quantitatives en analyse factorielle.
Cahier de l'Analyse des Données, vol IV n° 2, p. 137-146.

[10]ESCOFIER B. et PAGES J.P. (1982)
Comparaison de groupes de variables. 2ème partie : un exemple d'applications.
Rapports de Recherche I.N.R.I.A. n° 149 et n° 165

[11]ESCOFIER B. et PAGES J.P. (1983)
L'Analyse Factorielle Multiple : une méthode de comparaison de groupes de variables.
Actes des Troisièmes Journées Internationales Analyse des Données et Informatique.
INRIA. Versailles.

[12]ESCOUFIER Y. (1980)
L'analyse conjointe de plusieurs matrices.
Biométrie et temps.
Société Française de Biométrie.

[13]GOWER J.C. (1975)
Generalized procustres analysis.
Psychometrika, vol. 40 n°1, p. 33-51.

[14]HORAN C.B.
Multidimensional scaling : combining observations when individuals have different perceptuals structures.
Psychometrika 1969, 34,p139-165.

[15]HORST P. (1961)
Relations among m sets of measures.
Psychometrika, vol. 26 n°2, p. 129-149.

[16]HOTELLING H. (1936)
Relations between to sets of variables.
Biometrika, n°28, p.277-321.

[17]KETTENRING J.R. (1976)
Canonical analysis of several sets of variables.
Biometrika, vol.58 n°3,p.433-451.

[18]KOBILINSKY A. (1977)
Propriétés et utilisation de l'analyse multicanonique par la méthode de Carroll.
Analyse des données et Informatique, IRIA Rocquencourt.

[19]LEBART L., MORINEAU A., TABARD N. (1977)
Techniques de la description statistique : méthodes et logiciels pour l'analyse des grands tableaux.
Dunod, Paris.

[20]DE LEEUW J., PRUZANSKY S.
A new computational method to fit the weighted euclidean distance model.
Psychometrika 1978, vol 43 n°4, p. 479-490.

[21]L'HERMIER DES PLANTES H. (1976)
Structuration des tableaux à trois indices de la statistique.
Thèse de 3ème cycle, Université de Montpellier.

[22]PAGES J.P., CAILLEZ F., ESCOUFIER Y.
Analyse factorielle : un peu d'histoire et de géométrie.
Revue de Statistique Appliquée 1979, vol XXVII n°1, p.5,28.

[23]PLACE M.C. (1980)
Contribution algorithmique à la mise en oeuvre de la méthode STATIS.
Thèse de 3ème cycle, Université des Sciences et Techniques du Languedoc, Montpellier.

[24]SAPORTA G. (1975)
Liaisons entre plusieurs ensembles de variables et codage de données qualitatives.
Thèse de 3ème cycle, Paris VI.

[25]TAKANE Y., YOUNG F. , DE LEEUW J. (1977)

Nonmetric individual differences multidimensional scaling : an alternating least squares method with optimal scaling features.

Psychometrika, vol. 42 n°1, p.7-65.

[26]TEN BERGE (1977)

Orthogonal procustres rotation for two or more matrices.

Psychometrika, vol. 42 n°2, p.267-276.

[27]WOLLENBERG DEN M.L. (1977)

Redundancy analysis : an alternative for canonical correlation analysis.

Psychometrika, vol. 42 n°2, p.

Achevé d'imprimer
sur les presses de Copédith
7, rue des Ardennes, 75019 Paris
Mars 1984
Dépôt légal n° 6148