Notions sur les méthodes factorielles

Gilbert Saporta

4

NOTIONS SUR LES METHODES FACTORIELLES

Gilbert Saporta

Conservatoire National des Arts et Métiers
Paris

Ce chapitre donne une présentation non mathématique des principales méthodes. Le lecteur est prié de se reporter à l'un des nombreux ouvrages consacrés à l'analyse des données et cités en référence pour des démonstrations formelles ou un exposé technique détaillé.

4.1 Introduction

Les méthodes factorielles ont l'ambition de représenter un grand nombre de variables dans un espace de faible dimension. On y représente également les unités statistiques, soit individuellement, soit selon des groupes ou catégories en utilisant un principe barycentrique.

La possibilité de réduire la dimension provient de l'existence de corrélations entre les variables. Cette réduction s'effectue non pas par sélection d'un sous ensemble de variables mais par la construction de variables synthétiques, combinaisons linéaires des variables initiales. L'interprétation de ces "composantes" nécessite la connaissance de certaines techniques et une bonne dose de pratique.

On distinguera les variables selon leur nature : *qualitatives* ou *quantitatives* et selon leur fonction dans l'analyse : *actives* ou *illustratives* (on dit aussi *supplémentaires*). Seules les variables actives participent à la détermination de l'espace de représentation appelé espace factoriel.

Les variables actives doivent être toutes de même nature ce qui conditionne la méthode d'analyse : composantes principales pour les variables quantitatives, correspondances pour les variables qualitatives.

Bien évidemment, avant d'effectuer une analyse multidimensionnelle sophistiquée, il est recommandé de prendre contact avec les données au

moyen des outils classiques de la statistique descriptive ou de ceux plus récents de la statistique exploratoire :

- *tris à plat* avec histogrammes, box-plot (appelés parfois boîtes à pattes ou boîtes à moustache), courbes de densité pour les variables quantitatives ; "camemberts" et autres diagrammes pour les variables qualitatives.
- *tris croisés* qui consistent à ventiler les observations selon deux variables afin d'étudier leur liaison.

4.2 Analyse d'un ensemble de variables numériques : ACP

4.2.1 Représentations graphiques, nuages associés ; notions d'espaces des individus et des variables

On considérera ici le tableau rectangulaire **X** à n lignes et p colonnes des valeurs relevées sur n individus de p variables quantitatives actives.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Chaque individu i sera considéré comme un point e_i d'un espace à p dimensions de coordonnées $x_{i1}\ x_{i2}\ \dots\ x_{ip}$: cet espace appelé espace des individus sera noté E. Lorsque les variables sont exprimées avec des unités incommensurables (poids en kg, âge en années etc.) on ne peut calculer directement les distances entre individus ; il faut pour cela donner une formule de distance appropriée.

La solution la plus courante, et la seule que nous retiendrons ici, consiste en fait à centrer et réduire toutes les variables : il n'y a plus alors de problèmes d'unités car les variables deviennent exprimées en nombre d'écarts-types.

Si s_j désigne l'écart-type de la variable j, le carré de distance entre i et i' s'écrit :

$$d^2(i, i') = \sum_{j=1}^{p} \frac{(x_{ij} - x_{i'j})^2}{s_j} ,$$

De plus cette transformation qui donne la même dispersion à toutes les variables ne privilégie aucune d'entre elles a priori. Seules les corrélations entre variables exprimées dans la matrice symétrique R détermineront les projections sur les axes factoriels. Cette transformation est faite implicitement par les logiciels.

$$R = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Les variables quant à elles sont des listes de n valeurs : on les considérera donc comme des vecteurs d'un espace à n dimensions appelé espace des variables et noté F. La longueur d'un vecteur dans cet espace est définie par la formule suivante :

$$\|X\| = \left[(1/n) \sum_{i=1}^{n} x_i^2\right]^{1/2},$$

On reconnaît ici l'expression de l'écart-type si la variable est de moyenne nulle.

Comme on travaille sur des variables centrées et réduites, les longueurs de toutes les variables sont égales à un : les extrémités des vecteurs représentant les variables sont donc toutes situées sur la sphère de rayon un de F.

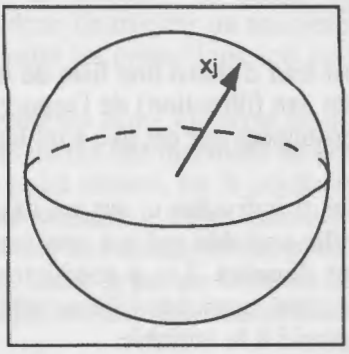

Figure 1

Les points représentant les variables centrées réduites sont tous sur une sphère

De plus l'angle formé par deux variables dans cet espace a pour cosinus le coefficient de corrélation linéaire entre ces variables.

Notons ici qu'il est facile d'introduire des pondérations p_i sur les individus (redressements d'échantillon par exemple) : il suffit de remplacer les moyennes arithmétiques simples par des moyennes pondérées.

$$\bar{x} = \sum_{i=1}^{n} p_i x_i \ ,$$

$$s^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2,$$

Les observations se présentent donc sous la forme de deux nuages de points : le nuage des n individus dans E et le nuage des p variables dans F.

L'étude de la forme du nuage des individus permettra de distinguer d'éventuels regroupements et de différencier des individus ou des groupes d'individus selon leurs réponses à l'ensemble des variables actives. La forme du nuage des variables décrit l'ensemble des corrélations entre celles-ci.

Au delà de trois dimensions, on ne peut appréhender directement ces espaces : il faut donc se placer dans des sous-espaces de dimensions acceptables pour visualiser les nuages.

Notons pour terminer que si à un individu on se contente d'associer un seul objet mathématique, un point de E, la représentation des variables est plus riche et plus complexe.

En effet une variable est tout d'abord une liste de n valeurs donc un vecteur de F ; mais c'est aussi un axe (direction) de l'espace des individus sur lequel on les projette : les coordonnées sur cet axe sont les valeurs de la variable.

Enfin si l'on projette les n individus e_i sur un axe quelconque Δ de E, on obtient donc une nouvelle variable qui est une combinaison linéaire des p variables du tableau des données. Les p coefficients de cette combinaison décrivent donc aussi une variable ; l'ensemble de ces coefficients s'appellera le facteur associé à la variable.

A une variable sont donc associés trois êtres mathématiques, ce qui mènera aux trois concepts de composante principale, d'axe principal et de facteur principal.

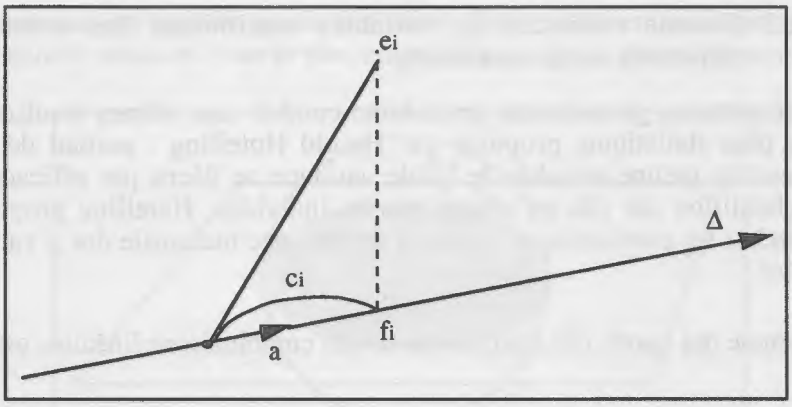

Figure 2

Projection de l'individu e_i sur l'axe Δ

4.2.2 Présentations de l'ACP

4.2.2.1 Comme projection d'individus

Cette présentation remonte à K. Pearson (1901).

On s'intéresse ici aux individus et on cherche à projeter le nuage de points sur un sous-espace de dimension fixée passant par l'origine (qui est le point moyen), de sorte que les distances entre individus projetés, ou plutôt leurs carrés, soient les plus proches possible des carrés des distances entre individus dans l'espace E. Comme les projections raccourcissent les distances, le critère sera donc de trouver un sous-espace tel que la moyenne des carrés des distances entre les projections soit maximale.

On montre que la moyenne des n(n - 1) carrés de distance entre points vaut deux fois la moyenne des carrés des distances au centre de gravité du nuage (le centre de gravité, ou point moyen, est le point dont les coordonnées sont les moyennes de chaque variable). Cette quantité est l'*inertie* du nuage. L'inertie totale est la somme des variances des p variables si on n'a procédé à aucune transformation. Dans le cas de données centrées réduites l'inertie vaut donc p puisque chaque variable est standardisée.

Les sous-espaces optimaux vérifient une propriété d'emboîtement : le sous-espace optimal de dimension k contient le sous-espace optimal de dimension (k - 1) etc. ce qui permet de se ramener à la recherche d'une suite d'axes orthogonaux appelés axes principaux du nuage de points.

4.2.2.2 Comme recherche de variables maximisant des critères de dispersion ou de corrélation

La présentation géométrique précédente conduit aux mêmes résultats que celle, plus statistique, proposée par Harold Hotelling : partant de l'idée élémentaire qu'une variable de faible variance ne décrit pas efficacement un échantillon car elle ne sépare pas les individus, Hotelling propose de rechercher les combinaisons linéaires de variance maximale des p variables initiales.

La somme des carrés des coefficients de ces combinaisons linéaires est fixée à un.

Les vecteurs renfermant les p coefficients de ces combinaisons, appelés facteurs principaux, sont les vecteurs propres de la matrice de corrélation **R** classés selon l'ordre décroissant des valeurs propres. Ces dernières sont les variances des combinaisons optimales que l'on appelle composantes principales. La somme des k premières valeurs propres est égale à l'inertie du nuage projeté sur le sous-espace de dimension k.

Une autre présentation de l'analyse en composantes principales, qui se généralise sous la forme d'un principe d'association maximale est la suivante : partant de p variables quantitatives on recherche de nouvelles variables notées c_k, non-corrélées entre elles, qui soient les plus liées aux p variables initiales au sens où la somme des carrés des coefficients de corrélation entre c_k et les p variables x_j est maximale.

$$\max \sum_i r^2 (c_k ; x_i)$$

Quelle que soit la présentation utilisée, l'analyse en composantes principales est une méthode factorielle linéaire : on construit de nouvelles variables, combinaisons linéaires des p variables de départ, non corrélées entre elles et de variance maximale.

4.2.3 Interprétation des résultats

4.2.3.1 Avec les variables actives

Les composantes de l'ACP sont de nouvelles variables. Pour les interpréter en fonction des anciennes variables on calcule les coefficients de corrélation

$$r(c_k ; x_j)$$

Ces coefficients sont les coordonnées des variables initiales dans l'espace défini par les composantes principales. La figure formée avec deux

composantes est appelée cercle des corrélations ; c'est en fait la projection de la sphère unité de F sur le plan engendré par c_1 et c_2 dans l'espace F.

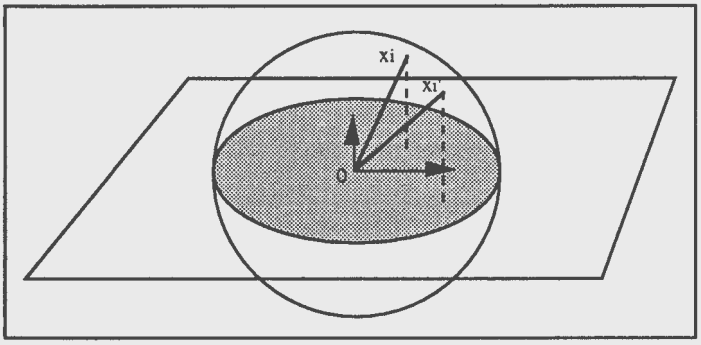

Figure 3

Le cercle des corrélations est une section (optimale) de la sphère de la figure 1

Les cercles de corrélation sont les éléments essentiels pour l'interprétation "interne" des résultats. On observe ce que l'on appelle un effet "taille" lorsque toutes les variables initiales sont corrélées positivement entre elles, ce qui conduit à un cercle des corrélations où toutes les variables sont d'un même coté de l'axe n°1.

4.2.3.2 Avec les variables illustratives

L'interprétation avec les variables actives peut être sujette à la critique classique de tautologie. L'utilisation de variables n'ayant pas servi à la détermination des axes apporte des éclairages différents, il s'agit alors d'une interprétation "externe".

Les variables quantitatives supplémentaires se représentent aisément dans le cercle des corrélations puisque l'on peut calculer simplement leurs corrélations avec les composantes principales. Les variables qualitatives quant à elles ne se représentent pas dans le même espace : en effet une variable qualitative se réduit à un ensemble de catégories ou modalités qui sont en fait des sous-groupes d'individus.

Chaque sous-groupe définit un sous-ensemble de points dans l'espace des individus ; il est certes possible de marquer chaque individu d'un sous-groupe par un symbole distinctif mais ceci ne permet de représenter qu'une seule variable à la fois. On utilise en général le principe barycentrique qui consiste à faire figurer le centre de gravité de chaque catégorie.

Il est clair que l'on perd ainsi la notion de dispersion et qu'il convient d'être prudent dans l'interprétation des proximités entre catégories. L'usage de domaines de confiance elliptiques autour des points moyens est alors une pratique recommandable.

La valeur-test associée à une modalité d'une variable qualitative (ou variable nominale) supplémentaire permet de vérifier rapidement si cette modalité diffère ou non de la moyenne générale.

Si a_j est la coordonnée sur un axe factoriel du centre de gravité de n_j individus de la catégorie j et λ_j leur variance sur ce même axe factoriel, on calcule :

$$v_j = \frac{a_j}{(\lambda_j)^{1/2}} \times \frac{(n-1)^{1/2}}{(n-n_j)^{1/2}} .$$

La quantité v_j représente la distance à l'origine de la modalité j exprimée en nombre d'écarts-types, dans l'hypothèse où les n_j individus seraient pris au hasard sans remise dans le plan factoriel.

4.2.3.3 Qualité des représentations

Les projections sur les plans principaux sont des représentations déformées de la réalité et il convient de prendre des précautions.

Un usage bien établi consiste à se servir des cosinus carrés entre les projections et les points initiaux. Des cosinus proches de 1 indiquent une bonne qualité de représentation. Cependant des cosinus proches de 0 ne sont pas toujours les témoins d'une mauvaise projection. Il vaut mieux dans ce cas se référer aux distances au plan de projection.

4.2.3.4 Choix du nombre d'axes

Le nombre d'axes à retenir est un problème délicat et qui n'a pas de solution rigoureuse. Il faut tout d'abord éliminer les critères liés à une valeur a priori du pourcentage d'inertie expliquée : on trouve trop souvent des affirmations du genre "il faut 80 % d'inertie expliquée" qui n'ont aucun sens.

S'il est souhaitable d'avoir beaucoup d'inertie expliquée sur un sous-espace cette valeur doit tenir compte du nombre de variables. Avoir 50 % d'inertie expliquée sur deux axes n'a pas le même sens avec 5 ou 50 variables au départ.

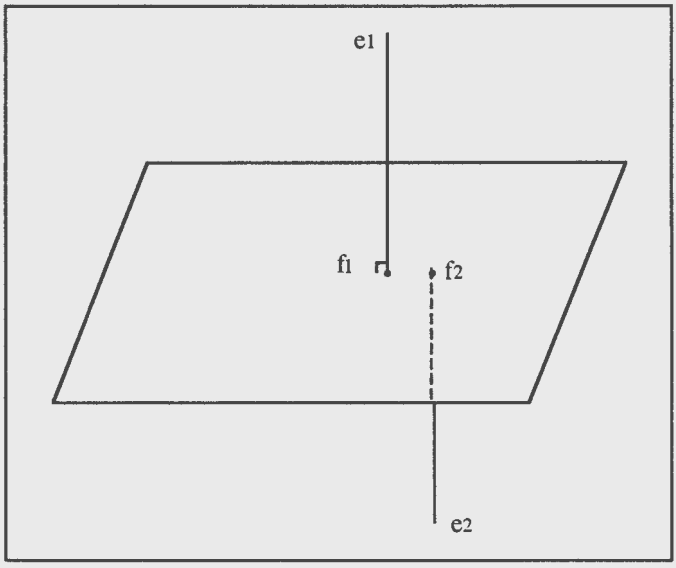

Figure 4

Les projections f_1 et f_2 sont proches, alors que les points e_1 et e_2 sont éloignés

Les critères théoriques liés aux propriétés de la loi normale multidimensionnelle n'ont qu'une valeur indicative, car ils ne résistent pas aux données réelles (le cas de distributions normales globales n'est d'ailleurs d'aucun intérêt en analyse des données).

Les seuls critères utilisables sont des critères empiriques : celui de Kaiser est un des plus connus qui consiste à ne retenir que les composantes associées à des valeurs propres supérieures à un. En effet en données centrées-réduites les variables de départ ont des variances égales à un et on cherche des combinaisons linéaires de variance maximale donc supérieures.

Le critère du "coude" ou de Cattell consiste à détecter un ralentissement dans la décroissance des valeurs propres : au delà, si les valeurs propres sont peu différentes entre elles, il n'y a plus que du bruit.

Enfin il faut obtenir des composantes interprétables : ce n'est pas seulement une question d'habileté ni d'aptitude à l'expression verbale ! On peut le formaliser en exigeant des corrélations suffisantes avec des variables supplémentaires (cf. l'application présentée au chapitre 9).

4.3 Analyse d'un tri croisé : AFC

Lorsque les données se mettent sous la forme d'un tableau de comptage N obtenu par croisement de deux caractères qualitatifs à p et q modalités respectivement, l'analyse factorielle des correspondances est la méthode privilégiée de description d'un tel tableau.

4.3.1 Profils lignes et profils colonnes, rappels sur le chi-deux

Le tableau N est en général ininterprétable directement et il faut tout d'abord étudier les pourcentages en lignes et en colonnes.
Le tableau des profils des lignes a pour terme général $n_{ij}/n_{i.}$ et vaut $D_1^{-1}\,N$ où D_1 est le tableau diagonal renfermant les effectifs totaux des catégories de la variable associée aux lignes.

Le tableau des profils des colonnes est de façon semblable égal à ND_2^{-1}.

Le problème essentiel consiste à étudier l'association entre lignes et colonnes, en d'autres termes l'écart à l'indépendance. On sait que l'indépendance se traduit par l'identité des p profils lignes (et l'identité des q profils colonnes) et que l'écart à l'indépendance peut se mesurer par le chi-deux :

$$\sum_{ij} \frac{(n_{ij} - n_{i.}\, n_{.j}/n)^2}{n_{i.}\, n_{.j}/n}$$

Chacun des éléments de cette double somme peut s'interpréter en termes d'écart positif ou négatif par rapport à une répartition "au hasard".

Lorsque les conditions usuelles sont vérifiées (échantillonnage aléatoire simple, n assez grand) on teste l'hypothèse d'indépendance en comparant la quantité précédente à une distribution du chi-deux à (p - 1) (q - 1) degrés de liberté.

4.3.2 AFC d'un tableau de contingence

4.3.2.1 Métrique du chi-deux, ACP des 2 tableaux de profils

Considérons pour commencer le tableau des profils des lignes. On peut l'envisager comme un tableau de données de type individus variables comme en ACP, où les individus seraient les lignes et les variables des pourcentages. On est donc en présence d'un nuage de p points dans un espace à q dimensions (en fait un sous-espace de dimension (q - 1) car les pourcentages ont pour somme cent). Etudier si les profils des lignes

diffèrent entre eux, c'est à dire étudier l'écart à l'indépendance, revient à regarder en quoi ce nuage diffère de son point moyen.

Effectuer l'ACP reviendra à représenter graphiquement les écarts à l'indépendance.

Ici deux différences notables interviennent : tout d'abord les points lignes ne sont pas équipondérés mais ont pour poids leurs fréquences marginales $n_{i.}/n$. D'autre part on utilise une formule spéciale de distance, la distance du chi-deux qui entre autres propriétés assurera la dualité entre ACP des profils des lignes et ACP des profils des colonnes.

$$d^2(i, i') = \sum_j \frac{n}{n_{.j}} \left(n_{ij} / n_{i.} - n_{i'j} / n_{i'.} \right)^2$$

Avec cette distance l'inertie du nuage des profils lignes est égale au chi-deux d'écart à l'indépendance divisé par n.

On montre alors que l'ACP du nuage des q profils colonnes conduit à des résultats similaires : mêmes valeurs propres, facteurs principaux égaux aux composantes principales de l'autre analyse.

Cette dualité est la source de la représentation simultanée des lignes et des colonnes de **N** sur un même graphique.

4.3.2.2 Contributions et interprétations

L'interprétation des résultats d'une AFC ne se fait pas exactement comme en ACP ; en effet le cercle des corrélations n'a guère de signification car on est en présence de catégories de variables qualitatives et non de variables quantitatives.

On utilise essentiellement les contributions à l'inertie. Chaque valeur propre λ étant égale à la moyenne pondérée des carrés des coordonnées des points profils, la contribution du profil i est ainsi égale à :

$$CTR\ (i) = \frac{n_{i.}}{n} (a_i)^2 / \lambda$$

où a_i est la coordonnée sur l'axe étudié.

Les contributions sont d'autant plus significatives qu'elles sont supérieures au poids $n_{i.}/n$.

Comme en ACP il est d'usage d'étudier la qualité des représentations graphiques à l'aide des cosinus carrés.

4.3.2.3 Commentaires sur la représentation simultanée

Un des attraits essentiels de l'AFC provient de la possibilité de projeter sur le même graphique lignes et colonnes du tableau **N**, contrairement à l'ACP où une telle opération est dénuée de sens.

Il faut cependant être prudent dans les interprétations des graphiques : si la proximité de deux profils lignes ou de deux profils colonnes s'interprète aisément, il n'en est pas de même pour la proximité entre un point ligne et un point colonne.

Deux considérations permettent de comprendre ce qui se cache derrière cette opération : tout d'abord la position du point colonne j s'interprète comme un pseudo barycentre des points lignes i pondérés par les poids $n_{ij} / n_{.j}$ mais surtout les points lignes et les points colonnes sont en fait les barycentres des groupes d'individus associés aux modalités des deux variables croisées. On est en fait en présence d'une figure analogue à celle représentant les catégories de deux variables qualitatives supplémentaires en ACP. Les lignes et les colonnes de **N** sont des éléments de même nature : des classes de deux partitions de l'ensemble des n individus.

4.3.2.4 Choix du nombre d'axes

Parmi les critères empiriques, seule la recherche d'un coude s'impose ; en effet la règle de Kaiser ne s'applique pas ici.

De nombreux travaux ont été effectués ces dernières années pour proposer des tests statistiques. Une des méthodes les plus efficaces est celle proposée par E. Malinvaud (1987) qui consiste à comparer par un test du chi-deux le tableau **N** à sa reconstitution approchée $\widetilde{\mathbf{N}}$ à l'aide de k axes :

$$\sum_{ij} (n_{ij} - \widetilde{n_{ij}})^2 / \widetilde{n_{ij}}$$

qui vaut approximativement n fois la somme des valeurs propres négligées. On compare ce résultat à un chi-deux à (p-k-1)(q-k-1) degrés de liberté.

4.3.3 Extensions à d'autres types de tableaux

Conçue au sens strict pour des tableaux de contingence, l'AFC peut s'étendre à d'autres tableaux de nombres positifs pourvu que la distance du chi-deux ait un sens.

4.3.3.1 Tableaux de comptage

De tels tableaux se rencontrent fréquemment dans des études de marché ; ainsi le tableau cité par J.P. Benzecri donnant pour des marques de cigarettes le nombre de fois où tel attribut a été associé à telle marque, un interviewé pouvant donner plusieurs évocations pour chaque marque.

4.3.3.2 Tableaux de contingence juxtaposés

Le tableau donnant pour chaque quartier de Paris les nombres de ménages par sexe, PCS, classe d'âge, taille du ménage est en fait la juxtaposition de quatre tableaux de contingence. Formellement il ressemble à un tableau de contingence et peut être soumis à l'AFC. L'existence de relations de somme constante entre les effectifs conduit à des valeurs propres nulles.

L'analyse d'un tel tableau a surtout pour but d'étudier les ressemblances et les différences entre quartiers vis à vis des quatre variables mais non les liaisons entre ces variables.

4.4. Analyse d'un ensemble de variables qualitatives : ACM

L'analyse de plus de deux variables qualitatives s'effectue à l'aide d'une extension de l'AFC que l'on appelle analyse des correspondances multiples (ACM).

4.4.1 Forme disjonctive et extension de l'AFC

Considérons n individus décrits par p variables qualitatives à m_1 m_2...m_p catégories respectivement. Le tableau donnant pour chaque individu les numéros ou codes des modalités ne se prête pas à des calculs algébriques. La façon naturelle de coder ces données est la forme disjonctive qui consiste à éclater chaque variable qualitative en autant de variables binaires qu'elle possède de modalités : un individu de sexe masculin sera ainsi codé 1 0. Pour p variables on aboutit donc à un tableau **X** comportant m_1+ m_2+...+m_p colonnes.

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Ainsi, ce tableau comprend trois blocs ayant respectivement $m_1 = 3$, $m_2 = 2$, $m_3 = 2$ colonnes. Ce tableau est très creux car il ne comporte que

p "1" dans chaque ligne (ici p = 3) et il est de plus redondant car la somme des indicatrices de chaque variable qualitative vaut 1.

Lorsque p vaut 2, on montre que l'analyse des correspondances de **X** conduit aux mêmes représentations que l'analyse des correspondances du tableau de contingence **N** croisant les deux variables, mais avec des valeurs propres différentes. L'extension à p variables s'en déduit immédiatement : l'analyse des correspondances multiples n'est en fait que l'analyse des correspondances du tableau disjonctif. Il convient d'utiliser cependant un programme d'ordinateur spécifique pour calculer les paramètres de façon économique et interpréter les résultats commodément.

4.4.2 Autres présentations

L'ACM peut s'obtenir en partant de nombreux autres points de vue, qui en font la richesse ; en voici deux :

4.4.2.1 Principe barycentrique généralisé

Si on représente les n individus et les $m_1 + m_2 + \ldots + m_p$ catégories simultanément on a la propriété suivante qui est caractéristique de l'ACM : à un coefficient près, chaque catégorie est située au barycentre des observations qu'elle regroupe, chaque individu est situé au barycentre des p catégories qui le décrivent.

4.4.2.2 ACP de variables qualitatives

Les composantes de l'ACM qui contiennent les coordonnées des n individus sur chaque axe sont des variables quantitatives vérifiant une propriété d'association maximale voisine de celle de l'ACP :

$$\sum_j \eta^2 (c ; x_j)$$

elles rendent maximale la somme des carrés des rapports de corrélation $\eta(c ; x_j)$ avec les p variables qualitatives. Rappelons que le carré du rapport de corrélation est la proportion de la variance d'une variable quantitative expliquée par une variable qualitative (variance des moyennes par catégories / variance totale).

En ce sens l'ACM apparaît comme une extension de l'ACP à des variables qualitatives. On n'aboutit pas cependant à des représentations des p variables mais seulement de leurs modalités.

4.4.3 Interprétation des résultats et problèmes spécifiques

Comme en AFC on utilise les contributions des catégories aux inerties de chaque axe, et comme en ACP les variables supplémentaires.

Compte tenu de la nature particulière du tableau disjonctif qui conduit à un nuage très éparpillé, les pourcentages d'inertie apportés par chaque axe sont en général très faibles ce qui surprend le néophyte. Il n'y a là rien d'inquiétant mais cela rend un peu plus délicat le choix du nombre pertinent d'axes à retenir.

4.5 Conclusion

Les trois méthodes d'analyse factorielle présentées ici permettent de décrire rapidement et efficacement les informations contenues dans de grands tableaux en utilisant simplement la structure des relations deux à deux entre variables.