Chapitre **4**

Analyse des Correspondances Multiples

4.1 DONNÉES ET NOTATIONS

4.1.1 Données

L'Analyse des Correspondances Multiples (ACM) permet d'étudier une population de I individus décrits par J variables qualitatives.

Une variable qualitative (ou nominale) est une application de l'ensemble I des individus dans un ensemble fini sur lequel on ne considère aucune structure : par exemple un ensemble de trois couleurs (bleu, blanc, rouge). Les éléments de cet ensemble sont appelés modalités de la variable et l'on dit par exemple qu'un individu bleu possède la modalité *bleu*.

L'application la plus courante de l'ACM est le traitement de l'ensemble des réponses à une enquête. Chaque question constitue une variable dont les modalités sont les réponses proposées (parmi lesquelles chaque enquêté doit faire un choix unique).

Nous commençons par passer en revue différentes façons de transcrire numériquement l'ensemble de ces données.

4.1.2 Codage condensé

Ces données peuvent être rassemblées dans un tableau de type $Individus \times Variables$ tout à fait analogue à celui étudié en ACP. Les lignes représentent les individus, les colonnes représentent les variables : à l'intersection de la ligne i et de la colonne j, se trouve la valeur x_{ij} (on dit aussi le codage condensé) de l'individu i pour la variable j (*cf.* **Figure 4.1**). Généralement, x_{ij} est le numéro de la modalité (de la variable j)

possédée par i mais beaucoup de logiciels acceptent pour x_{ij} une chaîne de caractères désignant la modalité (codage dit « alphabétique »).

Naturellement, même lorsque ce sont des nombres, les valeurs x_{ij} sont des codifications qui ne possèdent pas de propriétés numériques. Si la variable j est la couleur des individus, cette couleur peut être codifiée ainsi : bleu = 1, blanc = 2, rouge = 3. Il est clair que la moyenne entre *bleu* et *rouge* n'a pas grand sens et ne peut être considérée comme étant *blanc* ! Il n'est donc pas possible de traiter directement ce tableau par ACP (ou AFC) : les tableaux *Individus*×*Variables qualitatives* possèdent des spécificités et leur analyse factorielle nécessite une méthode spécifique.

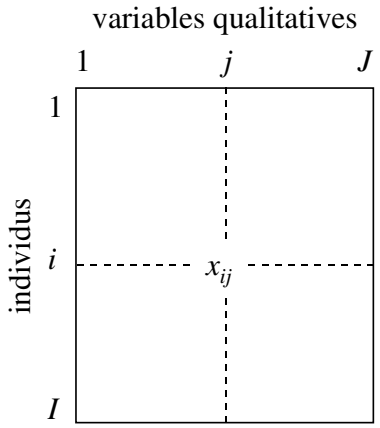

Figure 4.1 Tableau des données sous forme de codage condensé. I : nombre et ensemble des individus. J : nombre et ensemble des variables qualitatives. x_{ij} : codage condensé de la valeur de l'individu i pour la variable j (numéro ou chaîne de caractère).

4.1.3 Tableau Disjonctif Complet

Une autre façon de présenter ces mêmes données est de construire un Tableau Disjonctif Complet (TDC). Dans ce tableau, les lignes représentent les individus et les colonnes représentent les modalités des variables : à l'intersection de la ligne i et de la colonne k, on trouve x_{ik} qui vaut 1 ou 0 selon que l'individu i possède la modalité k ou non (*cf.* **Figure 4.2**). L'origine de la terminologie « Tableau Disjonctif Complet » est la suivante : l'ensemble des valeurs x_{ik} d'un même individu, pour les modalités d'une même variable, comporte la valeur 1 une fois (complet) et une fois seulement (disjonctif).

Les colonnes de ce tableau sont des fonctions numériques définies sur l'ensemble des individus appelées indicatrices de modalité.

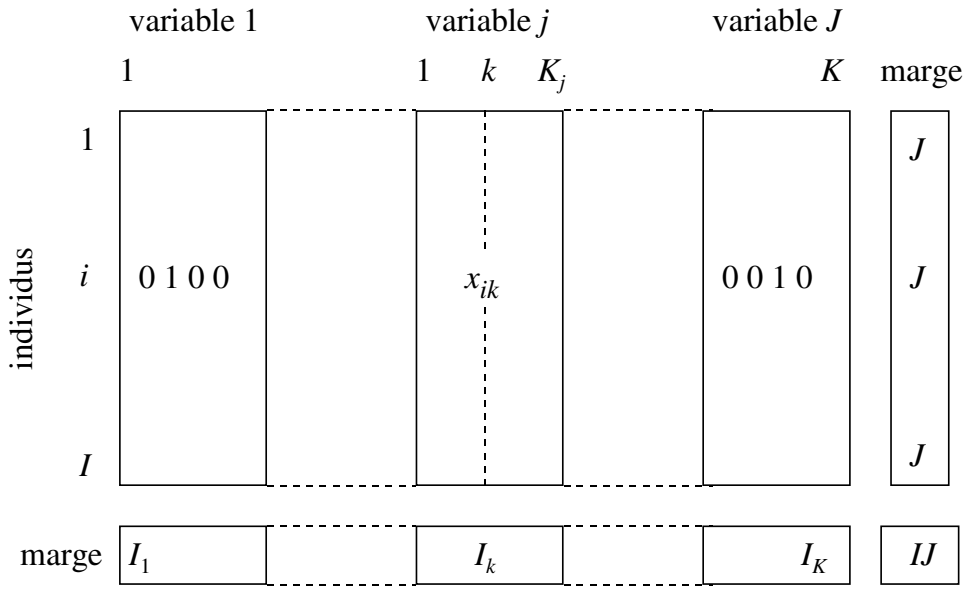

Figure 4.2 Tableau des données sous forme disjonctive complète. K_j = nombre et ensemble des modalités de la variable j. $K = \sum_{j=1}^{j=J} K_j$ = nombre et ensemble des modalités toutes variables confondues. $x_{ik} = 1$ si l'individu i possède la modalité k et 0 sinon $\sum_{k=1}^{k=K_j} x_{ik} = 1$ pour tout (i, j) $\sum_{k=1}^{k=K} x_{ik} = J$ pour tout i ; $\sum_{i=1}^{i=I} x_{ik} = I_k$ pour tout k ; $\sum_{k=1}^{k=K_j} I_k = I$ pour tout j

4.1.4 Hypertableau de contingence

Lorsque le nombre de variables J est réduit à 2, ces mêmes données peuvent être présentées sous la forme d'un tableau de contingence mettant en correspondance les deux ensembles de modalités.

Une généralisation directe du cas où $J = 2$ suggère de concevoir, sinon de construire explicitement, l'hypertableau de contingence dont chaque dimension est une variable. La **figure 4.3** représente cette construction quand $J = 3$. Cet hypertableau est bien équivalent aux données initiales. Néanmoins, son nombre de cases croît si rapidement avec J que, dans la plupart des situations concrètes, presque toutes les cases ont un effectif nul (si l'on mesure sur 1 000 plantes 10 variables à 5 modalités, l'hypertableau associé possède environ 10^7 cases dont au plus une sur 10 000 sera d'effectif non nul). Le développement de méthodes générales d'analyse de cet hypertableau est sans intérêt pratique immédiat. En revanche, le cas où $J = 3$ conduit à un hypertableau de dimension raisonnable et mérite une attention particulière : nous lui consacrons le chapitre 10.

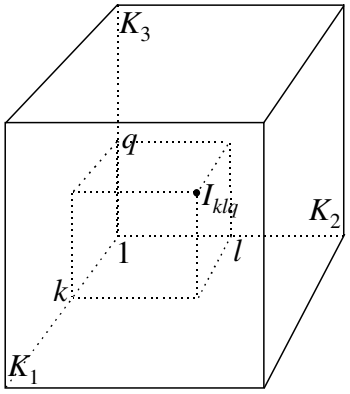

Figure 4.3 L'hypertableau de contingence associé à 3 variables qualitatives. K_1 : nombre de modalités de la première variable. I_{klq} : nombre d'individus possédant les modalités k (de la variable 1), l (de la variable 2) et q (de la variable 3).

4.1.5 Tableau de Burt

L'hypertableau étant la plupart du temps impossible à manier, pour généraliser l'analyse des correspondances à l'étude des croisements entre plus de deux variables, on peut construire un tableau contenant l'ensemble des tableaux de contingence entre les variables prises 2 à 2. Le « tableau de Burt » (*cf.* **Figure 4.4**) n'est pas exactement un tableau de contingence mais une juxtaposition de tels tableaux ; chaque individu y apparaît J^2 fois. Les tableaux contenant la diagonale croisent chaque variable avec elle-même : ils ne contiennent que des 0 sauf sur la diagonale qui contient les effectifs totaux I_k des modalités.

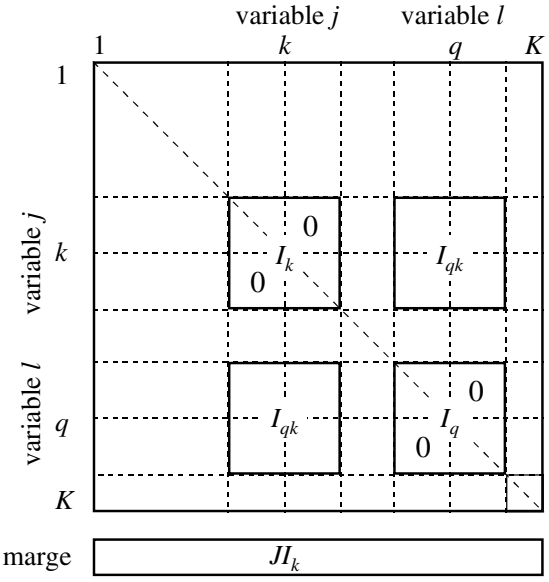

Figure 4.4 Tableau de Burt. Le tableau est symétrique. Les tableaux J situés sur la diagonale sont diagonaux. I_{qk} : nombre d'individus possédant à la fois la modalité q (de la variable l) et la modalité k (de la variable j). I_k : nombre d'individus possédant la modalité k (de la variable j).

Ce tableau est analogue à une matrice des corrélations en ce sens qu'il récapitule l'ensemble des liaisons entre les variables prises 2 à 2. Il contient beaucoup moins d'information que l'hypertableau et ne permet pas de reconstruire le TDC.

4.2 OBJECTIFS

La problématique de l'ACM est apparentée à celle de l'ACP (étude d'un tableau *Individus*×*Variables*) mais peut être considérée aussi comme une généralisation de celle de l'AFC (étude de la liaison entre plusieurs variables qualitatives). Ces deux aspects sont toujours plus ou moins explicitement présents dans les objectifs de l'ACM, présentés ici à partir des trois familles d'objets qui interviennent en ACM : les individus, les variables et les modalités des variables.

4.2.1 Étude des individus

De façon analogue à l'ACP, l'un des objectifs de l'ACM est de réaliser une typologie des individus. Cette typologie doit s'appuyer sur une notion de ressemblance telle que deux individus sont d'autant plus proches qu'ils possèdent un grand nombre de modalités en commun.

En outre, dans la plupart des applications de l'ACM, les individus sont très nombreux et ne sont connus que par leurs caractéristiques présentes dans le tableau de données. Par exemple, dans une enquête d'opinion, on ne dispose pour chaque individu d'aucune autre connaissance que ses réponses au questionnaire. En ce cas, les individus sont étudiés au travers des classes définies par les variables. Ainsi, dans les enquêtes d'opinion, on s'intéresse, par exemple, aux femmes, aux jeunes, aux retraités, etc. Une analyse des individus au travers de ces classes doit être telle que deux classes se ressemblent d'autant plus que leurs profils de répartition sur l'ensemble des modalités sont proches.

4.2.2 Étude des variables

Procédant encore de façon analogue à l'ACP, on peut adopter deux points de vue dans l'étude des variables.

Le premier est celui du bilan des liaisons entre les variables. L'étude de la liaison entre deux variables qualitatives nécessite de considérer le tableau de contingence croisant leurs modalités. Un bilan un tant soit peu détaillé de ces liaisons implique donc de se situer au niveau des modalités plus qu'à celui des variables.

Le second consiste à résumer l'ensemble des variables (qualitatives) par un petit nombre de variables numériques. Par exemple, on peut chercher à résumer un ensemble de variables socio-professionnelles par un indicateur de « statut social ». L'intérêt de ces variables synthétiques provient de ce qu'elles sont liées à l'ensemble des variables

étudiées. Ainsi, une variable ne pourra être considérée comme un indicateur de « statut social » que si elle est liée à la fois à la catégorie socio-professionnelle, au type de diplôme, etc.

Remarque. Par rapport à l'ACP, on cherche, selon ce second point de vue, une variable quantitative pour synthétiser un ensemble de variables qualitatives (et non quantitatives) ce qui implique, d'une façon ou d'une autre, d'affecter un coefficient à chaque modalité de chaque variable ; pour un individu, la valeur de la variable synthétique est alors la somme des coefficients des modalités qu'il possède.

4.2.3 Étude des modalités

Etudier l'ensemble des modalités revient à dresser un bilan de leurs ressemblances. Or une modalité peut être considérée selon deux points de vue :

1. en tant que variable indicatrice définie sur l'ensemble des individus, soit une colonne du TDC (*cf.* section 4.1.3) ;
2. en tant que classe d'individus dont on connaît la répartition sur l'ensemble des modalités, soit une ligne ou une colonne du tableau de Burt (*cf.* section 4.1.5).

La notion de ressemblance entre modalités diffère selon le point de vue adopté. Dans le premier cas, la ressemblance entre deux modalités doit reposer sur leur association mutuelle : deux modalités se ressemblent d'autant plus qu'elles sont présentes ou absentes simultanément chez un grand nombre d'individus. Les autres modalités n'interviennent pas.

Dans le second cas, la ressemblance entre deux modalités est analogue à celle que l'on utilise dans les tableaux de fréquence. Une ligne du tableau de Burt caractérise l'association de la modalité avec les modalités de toutes les variables : deux modalités se ressemblent d'autant plus qu'elles s'associent beaucoup ou peu aux mêmes modalités.

4.2.4 Conclusion sur les objectifs

L'étude d'un tableau *Individus*$\times$*Variables qualitatives* met en jeu trois familles d'objets : individus, variables et modalités. Il en résulte une problématique beaucoup plus riche et complexe que le triptyque classique : typologie des lignes, typologie des colonnes, mise en relation des deux typologies. Cette richesse ne doit cependant pas faire oublier l'unicité du tableau : il ne peut être question d'étudier séparément les différents aspects de la problématique par des méthodes sans rapport entre elles. Pratiquement, cette unicité est réalisée en articulant les interprétations autour de la typologie des modalités. En effet, cette typologie permet d'étudier l'association mutuelle entre les modalités, c'est-à-dire les liaisons entre les variables. Elle permet aussi d'aborder celle des individus en examinant le comportement moyen de classes d'individus.

Les objectifs indiqués dans l'étude des variables et des individus s'expriment ainsi en grande partie à l'aide des modalités.

4.3 AFC APPLIQUÉE À UN TABLEAU DISJONCTIF COMPLET

4.3.1 ACM et AFC

Lorsque les programmes d'AFC ont commencé à être diffusés, l'idée est venue d'appliquer ces programmes à des TDC. Rapidement, on s'est rendu compte que cette méthodologie fournissait des résultats intéressants, c'est-à-dire faisait apparaître des structures du tableau des données mettant en jeu un grand nombre de lignes et de colonnes.

En fait, conçue pour traiter des tableaux de fréquence, l'AFC en tant que méthode ne peut s'appliquer aux tableaux *Individus*$\times$*Variables qualitatives*. En revanche, les calculs de l'AFC, c'est-à-dire concrètement le programme, peuvent bien sûr être appliqués aux TDC. Mais, dans ce cas, ces calculs doivent être réinterprétés en fonction de la nature particulière du tableau. Ces calculs, munis de cette nouvelle interprétation, constituent une méthode à part entière ; d'où l'introduction du vocable Analyse des Correspondances Multiples. L'AFC d'un TDC n'est qu'une façon pratique de réaliser les calculs, d'ailleurs incomplète puisqu'elle ignore la notion de variable et donc ne fournit aucun résultat les concernant.

Cela étant, nous suivrons cette démarche historique et commode pour présenter l'Analyse des Correspondances Multiples.

Un TDC possède non seulement une nature différente de celle d'un tableau de contingence (ils codent les données différemment) mais aussi des propriétés numériques particulières. Les plus importantes sont celles-ci (*cf.* **Figure 4.2**) :

1. les valeurs dans le tableau ne sont que des 0 et des 1 ;
2. les colonnes peuvent être regroupées par paquets (qui correspondent chacun à une variable) dont la somme est une colonne composée de 1 ;
3. la somme des nombres d'une même ligne est constante et égale à J, nombre total de variables.

Les sections suivantes montrent que les distances, les poids et les facteurs de l'AFC d'un TDC correspondent aux objectifs préalablement fixés.

4.3.2 Nuage des individus

La marge sur I étant constante, la transformation en profils-lignes ne modifie guère les données. Un individu est représenté par les modalités qu'il possède. Deux individus se ressemblent s'ils présentent globalement les mêmes modalités. Plus précisément, la

distance entre deux individus i et l est définie par :

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

L'expression $(x_{ik} - x_{lk})^2$ vaut 0 ou 1 et ne diffère de 0 que pour les modalités k possédées par un seul des deux individus i et l. La distance d(i,l) croît avec le nombre de modalités qui diffèrent pour les individus i et l (ce qui est logique !). Une modalité k intervient dans cette distance avec le poids I/I_k, inverse de sa fréquence : la présence d'une modalité rare éloigne son ou ses possesseurs de tous les autres individus.

La distance induite par l'AFC appliquée à un TDC est donc satisfaisante. Le poids affecté à chaque individu l'est aussi puisqu'il est identique pour chacun (du fait de la marge constante).

Le centre de gravité de ce nuage, noté G_I, a pour coordonnée, pour la modalité k, I_k/IJ , proportion, au coefficient J près, des individus ayant choisi la modalité k . Il peut s'interpréter comme un individu théorique « moyen » (dans une enquête, cet individu aurait pu « partager » sa réponse à une question dans les différentes modalités, et ce proportionnellement aux réponses de l'ensemble des individus). On retrouve ici le fait qu'un individu est d'autant plus éloigné de G_I qu'il possède des modalités rares.

4.3.3 Nuage des modalités

La modalité k est représentée par le profil de la colonne k. Les nombres du TDC ne pouvant prendre que les valeurs 0 ou 1, le profil de la colonne k ne contient à son tour que deux valeurs possibles : 0 ou $1/I_k$. En outre, le centre de gravité du nuage des modalités, noté G_K , qui se confond avec le profil de la marge sur I, est caractérisé par un profil constant égal à $1/I$ (équivalent à une modalité que tous les individus auraient choisie). Il en résulte que le profil de la colonne k ressemble d'autant plus au profil moyen que l'effectif de la modalité k est grand. Réciproquement, une modalité rare sera toujours loin du centre de gravité du nuage des modalités.

La distance entre deux modalités k et h est définie par :

$$d^2(k,h) = \sum_i I\left(\frac{x_{ik}}{I_k} - \frac{x_{ih}}{I_h}\right)^2$$

En utilisant le fait que $(x_{ik})^2 = x_{ik}$ et en développant le terme carré, on obtient :

$$d^2(k,h) = \frac{I}{I_h I_k}[\text{nombre d'individus possédant une et une seule des modalités } h \text{ et } k]$$

Cette distance croît avec le nombre d'individus possédant une et une seule des deux modalités h et k, et décroît avec l'effectif de chacune de ces modalités. Deux

modalités d'une même variable sont obligatoirement assez éloignées l'une de l'autre dans l'espace. Deux modalités possédées par les mêmes individus sont confondues. Les modalités rares sont éloignées de toutes les autres. Cette distance traduit bien le premier des deux points de vue sur la ressemblance entre modalités indiqués dans les objectifs.

En appliquant ce calcul à la distance entre une modalité k et le centre de gravité G_K du nuage des modalités (correspondant à une modalité possédée par tous les individus), on trouve : $\mathrm{d}^2(k, G_K) = (I/I_k) - 1$; cela spécifie l'influence de l'effectif d'une modalité sur sa distance au point moyen.

Le poids de la modalité k vaut I_k/IJ ; il est proportionnel à l'effectif I_k.

➤ Remarques

Un élément (ligne ou colonne) influence la construction des axes par l'intermédiaire de son inertie par rapport au centre de gravité. Un calcul simple donne :

$$\text{Inertie de } k \text{ par rapport à } G_K = \frac{1}{J}\left(1 - \frac{I_k}{I}\right)$$

Ce résultat montre que, dans l'influence d'une modalité rare, le faible poids ne suffit pas à compenser leur éloignement. Par exemple, une modalité présente dans 1 % seulement de la population possède une inertie (c'est-à-dire une influence) presque deux fois plus grande qu'une modalité présente dans 50 % de la population. Concrètement, il est courant de voir les premiers facteurs d'une ACM déterminés presque exclusivement par quelques modalités très rares partagées par les mêmes individus. Comme il est souvent beaucoup plus intéressant de dégager des phénomènes généraux plutôt que ces phénomènes ponctuels, on cherche, en pratique, à éviter les modalités trop rares (en effectuant des regroupements).

En sommant les inerties des modalités, on montre facilement que l'inertie totale du nuage étudié vaut $(K/J) - 1$. En ACM, comme en ACP et à la différence de l'AFC, l'inertie totale des nuages n'intervient pas dans l'interprétation.

L'inertie des K_j modalités de la variable j vaut $(K_j - 1)/J$. Cette inertie, étant liée directement au nombre de modalités de la variable j, incite à exiger des nombres de modalités égaux pour toutes les variables actives. En fait, cette différence d'inertie entre variables ayant des nombres de modalités différents vaut pour l'espace entier R^I. Dès l'instant que l'on considère une seule direction de R^I, ce qui est le cas des axes factoriels, l'inertie du nuage des K_j modalités d'une même variable j est toujours inférieure à $1/J$, quantité ne dépendant pas de K_j. Il en résulte qu'il n'est pas gênant, de ce point de vue, de faire intervenir simultanément en actif des variables ayant des nombres de modalités différents. Ce problème sera à nouveau abordé en section 4.3.5.

4.3.4 Relations de transition et représentation simultanée

Avec les notations déjà utilisées en ACP et en AFC, les relations de transition de l'AFC, appliquées à un TDC, s'écrivent :

$$\begin{aligned} F_s(i) &= \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K} \frac{x_{ik}}{J} G_s(k) \\ G_s(k) &= \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{x_{ik}}{I_k} F_s(i) \end{aligned}$$

Du fait que x_{ik} ne prend que les valeurs 0 ou 1, ces relations de transition s'interprètent simplement. En projection sur l'axe s, l'individu i est placé, au coefficient $1/\sqrt{\lambda_s}$près, au barycentre des modalités qu'il possède. Parallélement, la modalité k est placée, au coefficient $1/\sqrt{\lambda_s}$ près, au barycentre des individus qui la possèdent. Il en résulte que, sur un axe, une modalité (colonne du TDC) représente à une dilatation près la moyenne des individus qui la possèdent (lignes du TDC). Aussi, dans l'étude de sa projection, on peut considérer une modalité aussi bien comme barycentre d'une classe d'individus (*i.e.* une ligne du tableau de Burt = somme des lignes du TDC correspondant aux individus possédant la modalité concernée) que comme indicatrice d'une variable (*i.e.* une colonne du TDC). Le coefficient de dilatation varie avec les axes, ce qui n'est pas gênant lorsque l'interprétation des résultats se fait facteur par facteur et milite pour examiner conjointement de préférence des axes d'inerties comparables (principe commun à toute les analyses factorielles).

Cette équivalence entre facteurs ne doit pas faire oublier que les modalités, d'une part en tant qu'indicatrices et d'autre part en tant que barycentres, sont situées dans des espaces différents. Il en résulte que les qualités de représentation d'une même modalité selon chacun des points de vue ne sont pas liées. En outre, les notions de proximité entre ces deux types d'objets diffèrent.

En effet, la proximité entre indicatrices mesure leur association mutuelle (*cf.* section 4.3.3). D'autre part, la proximité des moyennes de classes d'individus découle des distances définies entre les individus : deux classes d'individus k et h sont d'autant plus proches qu'elles possèdent des caractéristiques identiques quant à l'ensemble des variables, c'est-à-dire que les modalités k et h s'associent de la même manière aux modalités de toutes les variables. Cette notion de proximité correspond au second point de vue sur les ressemblances entre modalités dégagé dans les objectifs. Il est remarquable de constater, qu'à des dilatations axiales près, les deux notions de proximité fondées sur des principes différents conduisent aux mêmes graphiques dans l'analyse du TDC.

En pratique, les deux notions de proximité s'utilisent conjointement ; en particulier, on interprète souvent la proximité entre modalités de variables différentes en tant qu'association de modalités et la proximité entre modalités d'une même variable en

tant que ressemblance entre deux classes d'individus. Par exemple, en décrivant un plan factoriel sur lequel apparaissent différents repères sociaux, on interprète la proximité entre les modalités *retraités* et *plus de 65 ans* en terme d'association (ce sont presque les mêmes individus qui possèdent ces deux modalités) et la proximité entre *60 à 65 ans* et *plus de 65 ans* en terme de ressemblance (ces deux classes d'individus possèdent des caractéristiques identiques quant aux autres variables). Ainsi, les relations de transition, même si elles ne sont pas utilisées dans le cadre strict d'une représentation simultanée, confèrent à la représentation des modalités les propriétés souhaitables dégagées dans l'exposé des objectifs.

4.3.5 Les variables à travers leurs modalités

Les variables qualitatives ne sont pas introduites explicitement dans l'AFC d'un TDC. Elles n'apparaissent qu'à travers l'ensemble de leurs modalités. Les sous-nuages des modalités d'une même variable ont des propriétés qu'il est intéressant de connaître pour interpréter des résultats mais aussi pour coder des variables quelconques en vue de les traiter en variables qualitatives dans une ACM (*cf.* section 4.5).

a) Barycentre des modalités d'une variable

Comme le montre la relation ci-dessous, le barycentre des modalités d'une même variable se confond avec celui de l'ensemble du nuage.

$$\sum_{k \in K_j} \frac{I_k}{I} \frac{x_{ik}}{I_k} = \frac{1}{I}$$

La projection conserve cette propriété. L'ensemble des modalités d'une même variable est donc centré sur l'origine pour tous les graphiques ; les facteurs opposent entre elles à la fois l'ensemble de toutes les modalités et l'ensemble des modalités de chaque variable.

b) Sous-espace engendré par les modalités d'une variable

Du fait du caractère disjonctif du TDC, les vecteurs de R^I joignant l'origine (avant centrage) aux modalités d'une même variable sont orthogonaux entre eux. L'ensemble des r modalités d'une variable engendre un sous-espace de dimension égale à r. Du fait du caractère complet du TDC, tous ces sous-espaces possèdent une direction commune : celle qui relie l'origine au centre de gravité du nuage. Cette direction étant éliminée lors du centrage (*cf.* section 3.3), on peut considérer que, en ACM, une variable présentant r modalités engendre un sous-espace de dimension égale à $r - 1$. Il en résulte que, pour représenter parfaitement les r modalités d'une même variable, au moins $(r - 1)$ facteurs sont nécessaires.

Cette propriété a plusieurs conséquences pratiques :

1. quelle que soit la structure du tableau, le pourcentage d'inertie associé à chaque facteur, en particulier au premier, est nécessairement faible lorsque les variables présentent beaucoup de modalités ;
2. même si un facteur est très lié à une variable (en ce sens qu'il regroupe clairement les individus possédant la même modalité pour cette variable), il est impossible que toutes ses modalités soient bien représentées par ce seul facteur ;
3. dans l'élaboration d'un tableau de données, même si le nombre d'individus est très grand, il n'est pas utile de multiplier de façon importante les modalités d'une même variable : le gain de finesse obtenu risque de ne pas pouvoir être valorisé dans l'analyse.

L'inertie d'une variable à r modalités (égale à $(r-1)/J$; *cf.* section 4.3.3) est donc répartie dans un sous-espace à $r-1$ dimensions. On peut montrer en outre qu'elle est égale à $1/J$ dans toutes les directions de ce sous-espace. Il en résulte qu'une variable ayant un grand nombre de modalités, bien qu'engendrant une inertie importante dans R^I, n'a aucune raison d'infléchir le premier axe de façon privilégiée puisque cette inertie importante est en quelque sorte diluée dans un sous-espace de grande dimension.

4.3.6 Synthèse des variables qualitatives

Un aspect de l'étude d'un ensemble de variables est la mise en évidence d'un petit nombre de variables synthétiques, c'est-à-dire liées le plus possible à l'ensemble des variables initiales (*cf.* section 4.2.2). Pour montrer que les facteurs de l'ACM constituent ces variables synthétiques, nous utilisons le rapport de corrélation, qui mesure la liaison entre une variable numérique (ici le facteur) et une variable qualitative.

Rappelons la définition de ce rapport. Une variable qualitative définit une partition sur l'ensemble des individus en autant de classes qu'elle a de modalités. Utilisant le théorème de Huygens, l'inertie totale (ou variance) d'une variable numérique peut se décomposer en somme de l'inertie inter (*i.e.* inertie des centres de gravité des classes) et des inerties intra (*i.e.* inertie des individus par rapport au centre de gravité de la classe à laquelle ils appartiennent). Le carré du rapport de corrélation est le quotient de l'inertie inter par l'inertie totale. Il varie entre 0 et 1. Lorsqu'il est proche de 1, les individus d'une même classe sont très regroupés et les classes sont séparées les unes des autres : c'est une situation de liaison très forte entre la variable qualitative et la variable numérique. Lorsqu'il est proche de 0, les moyennes des classes sont très proches de la moyenne générale et les individus d'une même classe sont très dispersés : la variable qualitative et la variable numérique ne sont pas liées. La **figure 4.5** illustre ces deux cas extrêmes.

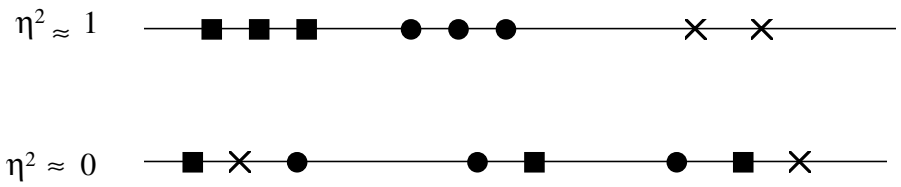

Figure 4.5 Illustration des deux valeurs extrêmes du rapport de corrélation. 8 individus, représentés par un symbole différent selon leur modalité pour une variable qualitative, figurent sur un axe représentant une variable numérique.

En notant G_k le barycentre des individus présentant la modalité k, le carré du rapport de corrélation entre une variable j et le facteur F_s vaut :

$$\eta^2(F_s, j) = \frac{\text{inertie inter}}{\text{inertie totale}} = \frac{\sum_{k \in K_j} (I_k/I)(F_s(G_k))^2}{\lambda_s}$$

En utilisant le fait que, en ACM, la modalité k a le poids I_k/IJ et se trouve, à un coefficient près, au barycentre des individus qui la possèdent, soit :

$$G_s(k) = F_s(G_k)/\sqrt{\lambda_s}$$

on trouve :

$$\eta^2(F_s, j) = J \sum_{k \in K_j} (\text{inertie de la modalité } k, \text{ projetée sur l'axe d'ordre } s)$$

Notons que le rapport de corrélation étant compris entre 0 et 1, l'inertie du sous-nuage des modalités d'une même variable sur un axe est comprise entre 0 et $1/J$: elle vaut $1/J$ si F_s appartient au sous-espace engendré par les modalités de la variable.

La quantité maximisée par les axes factoriels dans l'espace R^I est l'inertie projetée du nuage de l'ensemble des modalités. En regroupant les modalités d'une même variable, ce critère n'est autre que la moyenne des carrés des rapports de corrélation entre le facteur et chacune des variables. Il en résulte que les facteurs F_s de l'ACM sont les variables numériques les plus liées à l'ensemble des variables qualitatives étudiées et, en ce sens, constituent bien les variables synthétiques annoncées.

La première relation de transition (*cf.* section 4.3.4) fournit un éclairage sur la façon dont le facteur F_s est calculé pour chaque individu. À chaque modalité k, l'ACM affecte le poids $G_s(k)$; $F_s(i)$ est la moyenne de ces coefficients pour les modalités possédées par l'individu i (à $\sqrt{\lambda_s}$ près).

Les propriétés énoncées dans ces deux derniers paragraphes permettent de préciser l'influence relative d'une variable en ACM : **pour un axe donné, l'importance *a***

***priori* de chaque variable est la même mais le nombre d'axes sur lesquels une variable peut influer est directement lié au nombre de ses modalités**. Cela implique notamment que, si quelques variables très riches en modalités sont liées entre elles, les premiers facteurs peuvent n'exprimer que ces liaisons et il faudra alors chercher très loin dans la suite des facteurs pour percevoir d'autres liaisons.

4.3.7 Représentation des variables en ACM

Le concept de variable (et non plus de modalité) apparaît en ACM et conduit à des aides à l'interprétation. Ces indices complètent ceux déjà obtenus dans une simple AFC du TDC et qui concernent les individus et les modalités.

La contribution d'une variable à l'inertie d'un facteur est la somme des contributions de toutes ses modalités. Elle permet aussi de mesurer la liaison (rapport de corrélation) entre la variable et le facteur. Il est intéressant de commencer l'analyse des résultats d'une ACM par la consultation systématique de ces coefficients, qui met en évidence les variables les plus liées à chacun des facteurs.

Il peut être utile de construire le graphique suivant (*cf.* **Figure 4.6**) dit « carré des liaisons ». En abscisse et en ordonnée figurent deux facteurs, par exemple F_s et F_t. Dans ce repère, on peut représenter chaque variable j par un point dont la coordonnée sur F_s (respectivement F_t) est le carré du rapport de corrélation entre la variable j et F_s (respectivement F_t).

On montre (*cf.* section 8.6.2) que ce graphique s'interprète aussi comme la projection d'un nuage dans lequel chaque point représente une variable, la proximité entre deux points-variables traduisant la ressemblance entre les partitions engendrées par les deux variables.

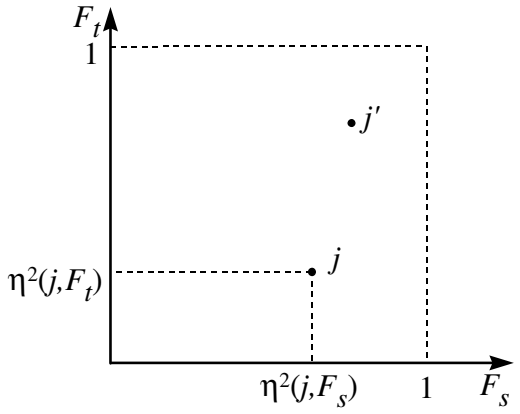

Figure 4.6 Représentation des variables en ACM (carré des liaisons). $\eta^2(j, F_s)$: rapport de corrélation entre la variable qualitative j et le facteur F_s. Par construction, pour tout j et tout s : $0 \leqslant \eta^2(j, F_s) \leqslant 1$. Ce graphique montre que les variables j et j' sont très liées au facteur F_s et que seule j' est liée à F_t .

4.4 ANALYSE DES CORRESPONDANCES D'UN TABLEAU DE BURT

4.4.1 Tableau de Burt et Tableau Disjonctif Complet

Nous avons vu, dans la section précédente, que la représentation des modalités dans l'analyse du TDC fournissait, à des dilatations axiales près, des représentations des barycentres de classes d'individus. Mais cette représentation est-elle optimum ? Autrement dit, si au lieu de calculer les axes d'inertie du nuage d'individus et de projeter les barycentres sur ces axes nous avions analysé directement le nuage des barycentres, aurions-nous obtenu le même résultat ? Très curieusement, et ce n'est pas la moindre surprise que réserve l'ACM, la réponse est oui.

Remarquons tout d'abord que la k^{e} ligne du tableau de Burt est la somme des lignes du TDC qui présentent la modalité k. Géométriquement, cela signifie que dans l'espace R^K, le profil de la modalité k (défini dans le tableau de Burt) se trouve au barycentre des profils des individus i (définis dans le TDC) qui la possèdent.

De plus, le TDC et le tableau de Burt ont la même marge sur l'ensemble K (*cf.* **Figures 4.2** et **4.4**). La métrique induite sur R^K dans l'AFC de chacun de ces deux tableaux est la même : les individus (définis dans le TDC) et leurs barycentres (définis dans le tableau de Burt) sont situés dans le même espace euclidien.

Enfin, dans l'AFC du TDC, tous les individus ont un poids identique tandis que dans l'AFC du tableau de Burt, le poids affecté au barycentre d'une classe est proportionnel à son effectif.

L'analyse du nuage des barycentres s'obtient donc par une AFC du tableau de Burt.

Or, on montre (*cf.* section 5.7) que **l'AFC du tableau de Burt et celle du TDC aboutissent au même résultat**. Plus précisément, les axes d'inertie du nuage des individus (lignes du TDC) et ceux de leurs barycentres (lignes du tableau de Burt) sont confondus. Il en découle que, pour obtenir simultanément les projections optimales des uns et des autres, il suffit d'appliquer une AFC au tableau juxtaposant en colonne le TDC et le tableau de Burt, en mettant indifféremment l'un ou l'autre des deux tableaux en « supplémentaire ». Cette équivalence présente un intérêt théorique important : l'optimalité simultanée des représentations des individus et des barycentres des classes. Elle présente aussi un intérêt pratique : la possibilité d'analyser le tableau de Burt à la place du TDC, le premier étant en général bien plus petit.

Attention : dans l'analyse du TDC, il faut bien distinguer la représentation des modalités en tant que colonnes (ou variables indicatrices) et la représentation des barycentres (ou moyennes de lignes). Selon les relations de transition, les deux représentations sont homothétiques dans le rapport $\sqrt{\lambda_s}$ pour l'axe d'ordre s. C'est la deuxième représentation qui est confondue avec celle des lignes du tableau de Burt (dans l'analyse de ce dernier, lignes et colonnes ont d'ailleurs la même représentation

du fait de la symétrie). Il en découle que les facteurs définis sur le même ensemble de colonnes K des deux tableaux ne sont pas égaux, mais homothétiques dans le rapport $\sqrt{\lambda_s}$. Les inerties (dans lesquelles les distances interviennent par leur carré) associées aux facteurs du tableau de Burt sont les carrés de leurs homologues dans le TDC.

4.4.2 Analyse des liaisons binaires et décomposition des χ^2

Le tableau de Burt est composé de J^2 tableaux de contingence croisant les variables deux à deux. Tous ces tableaux étant calculés à partir du même ensemble d'individus, les marges du tableau de Burt correspondant aux modalités des variables j et l sont égales, au coefficient J près, aux marges du tableau binaire croisant ces deux variables (*cf.* **Figure 4.4**). Le profil d'une modalité, ligne du tableau de Burt, n'est autre que la juxtaposition des J profils de cette même modalité dans les tableaux binaires où elle apparaît.

Dans l'AFC du tableau de Burt, il est intéressant d'interpréter l'inertie totale du nuage étudié. Rappelons que, dans l'AFC d'un tableau de contingence, cette inertie est proportionnelle au χ^2 d'indépendance. En utilisant le fait que les marges du tableau de Burt sont proportionnelles aux marges des sous-tableaux croisant les variables 2 à 2, on peut montrer que l'inertie totale est égale à la somme des χ^2 d'indépendance associés à chacun des J^2 sous-tableaux. La projection sur les facteurs décompose l'inertie des nuages. On peut interpréter un facteur comme une part de la somme de ces χ^2. En ce sens, cette AFC est une étude simultanée des liaisons binaires.

Dans cette somme de χ^2, les tableaux croisant deux variables différentes interviennent deux fois et les tableaux diagonaux croisant une variable avec elle-même interviennent une seule fois. Or les tableaux croisant une variable avec elle-même sont diagonaux, puisque les modalités d'une même variable s'excluent entre elles, et leur χ^2 n'est jamais nul (de ce fait l'inertie d'un tableau de Burt n'est pas nulle même lorsque tous les couples de variables sont indépendants). Le « biais » introduit par ces tableaux diagonaux dans l'étude simultanée des liaisons binaires est nul. En effet, on peut montrer que l'analyse d'un nouveau tableau, dérivé du tableau de Burt en remplaçant les tableaux diagonaux par le produit de leurs marges, aboutit, à un coefficient près, aux mêmes facteurs que celle du tableau de Burt.

> ***Remarque : cas de deux variables*** L'ACM peut théoriquement s'appliquer à l'étude de deux variables seulement. Dans ce cas, il est aussi possible d'analyser par l'AFC le tableau binaire croisant ces deux variables. On montre que ces deux analyses aboutissent encore aux mêmes résultats, en ce sens que si l'on juxtapose les facteurs de même rang obtenus sur les lignes et les colonnes du tableau binaire, on obtient, à une homothétie près, les facteurs du tableau de Burt.

4.5 CODAGE EN CLASSES DES VARIABLES QUANTITATIVES

Dans la pratique, les variables qualitatives étudiées en ACM résultent souvent d'une transformation de variables numériques (*e.g.* : l'âge est souvent pris en compte au travers de l'appartenance à une tranche d'âge). En outre, même lorsque la variable est par nature qualitative, il existe souvent, pour la prendre en compte, un choix entre plusieurs partitions plus ou moins fines (*e.g.* : les catégories socio-professionnelles). Les résultats dépendant du choix des partitions associées aux variables, ce problème est crucial.

En analyse des données, on appelle généralement **codage** la construction, à partir de données brutes, d'un tableau prêt à être analysé : en ce sens, le problème du choix des classes est un problème de codage. Il n'y a pas de méthode systématique pour réaliser un codage. La pratique et la théorie ont cependant dégagé un certain nombre de principes qu'il est prudent de respecter. En outre, les résultats d'une analyse permettent une validation ou une remise en question du codage utilisé. Seuls seront détaillés ici quelques problèmes relatifs au codage des variables numériques en variables qualitatives.

4.5.1 Pourquoi transformer des variables quantitatives en variables qualitatives ?

Deux objectifs principaux conduisent à coder par classes des variables continues en découpant leur intervalle de variation.

Tout d'abord, on peut vouloir **rendre homogènes** des données qui se composent initialement de variables numériques et de variables qualitatives. Ainsi, dans l'analyse d'un ensemble de repères sociaux (sexe, profession, âge, revenu, etc.), le fait de transformer les variables numériques *âge* et *revenu* en variables qualitatives permet de traiter l'ensemble de ces variables par l'ACM.

On peut aussi avoir intérêt à réaliser un codage qualitatif même lorsque l'on dispose d'un ensemble de variables numériques sur lequel une ACP peut tout à fait s'appliquer. En effet, une ACM sur ces mêmes variables codées en classes donne une autre approche des données. En représentant chaque variable par autant de points qu'elle possède de classes, l'ACM peut mettre en évidence, si elles existent, **des liaisons non linéaires** entre les variables. Ce type de liaison est assez fréquent car beaucoup de phénomènes présentent des effets de seuil : un état pathologique peut être caractérisé par une valeur « trop faible » ou « trop élevée » ; un fromage sera d'autant plus apprécié qu'il est salé mais jusqu'à un certain point (de ce point de vue, les deux extrémités de l'intervalle de variation du caractère « salé » sont plus proches entre elles qu'elles ne le sont des valeurs moyennes). Concrètement, sur les graphiques, la proximité de modalités extrêmes démontre l'aptitude de l'ACM à mettre en évidence des liaisons non linéaires.

De tels phénomènes sont naturellement invisibles dans les résultats d'une ACP qui ne tient compte que des liaisons linéaires. Paradoxalement, en réduisant l'information traitée (l'appartenance à une classe ou un intervalle est moins précise qu'une valeur numérique), on augmente la richesse du résultat ! Notons par exemple que la moyenne d'une classe d'individus comprenant des individus très grands et des individus très petits correspond à un individu moyen pour une variable numérique alors qu'elle correspond à une répartition dans les deux extrêmes pour cette même variable codée en qualitative.

L'ACM de variables numériques codées en variables qualitatives est une approximation d'une analyse non linéaire dans le sens suivant : on cherche des variables synthétiques qui soient des combinaisons linéaires de fonctions quelconques des variables étudiées et non, comme en ACP, des variables elles-mêmes. Ce problème n'a de sens que dans le cadre d'un modèle où la population est infinie. En pratique, en ACM sur une population finie, au lieu de considérer l'ensemble des fonctions d'une variable, on divise l'intervalle des valeurs de la variable en sous-intervalles et l'on considère l'ensemble des fonctions constantes sur chaque sous-intervalle. En effet, quand on traite par l'ACM une variable qualitative j, cette variable est représentée dans R^I par le sous-espace E_j engendré par les indicatrices de ses classes ; E_j n'est autre que l'ensemble des variables ayant une même valeur pour tous les éléments d'une même classe. Le premier facteur est la combinaison linéaire des éléments de ces J sous-espaces E_j (chaque élément est une fonction constante sur les classes d'une variable) la plus proche possible de tous ces sous-espaces.

Ce codage permet aussi d'étudier des variables dont les distributions sont très irrégulières et pour lesquelles le coefficient de corrélation est une mesure de liaison inadaptée. Par exemple, si un élément a une valeur très éloignée des valeurs des autres éléments, il influe de manière prépondérante sur les coefficients de corrélation et un codage qualitatif le neutralise.

4.5.2 Choix du nombre de classes

Pour coder par classes une variable continue, c'est-à-dire découper son intervalle de variation en sous-intervalles qui définissent autant de modalités, il faut déterminer d'une part le nombre de classes et d'autre part leurs limites. Cette séparation est un peu formelle dans la mesure où les deux choix sont souvent effectués simultanément.

Combien de classes faut-il utiliser ? Ni trop, ni trop peu.

En diminuant à l'excès le nombre de classes, on regroupe des individus de plus en plus différents et de ce fait on perd beaucoup d'informations. Les modalités recouvrent alors des situations très variées et leur étude ne peut mettre en évidence que des phénomènes très généraux.

En augmentant le nombre de classes, on risque d'obtenir des classes d'effectif faible avec tous les inconvénients que cela comporte. Si l'effectif de la population est très grand, ce risque est écarté et l'on peut être tenté de prendre un grand nombre de classes. Toutefois, un nombre de classes excessivement grand n'est pas sans poser de problèmes. Plus on éclate les classes, plus on risque de faire apparaître des liaisons ponctuelles entre quelques modalités. D'autre part, chaque variable intervient dans l'analyse par le sous-espace de dimension $r - 1$ engendré par ses r modalités. Lorsque l'on augmente r, le nombre de facteurs sur lesquels une variable peut influer augmente et l'aspect synthétique de l'analyse n'est pas amélioré, bien au contraire !

Indiquons, pour fixer les idées, que l'expérience montre qu'il n'est pas utile de dépasser le nombre de 8 modalités dans le codage de variables quantitatives et que 4 ou 5 sont souvent bien suffisantes.

4.5.3 Choix des classes

Pour choisir les classes, on examine tout d'abord s'il n'existe pas des seuils naturels ou classiques pour la variable mesurée. Ainsi, dans une étude sociale, l'âge du départ à la retraite est une limite « naturelle ».

Lorsque ce point de vue ne suffit pas, on étudie les irrégularités de la répartition des valeurs. Pour cela, on construit un histogramme avec de nombreuses classes. Les « creux » dans la répartition suggèrent des coupures de l'intervalle de variation.

Lorsque les deux principes précédents n'imposent aucun seuil, on réalise un découpage systématique de l'intervalle de variation. Le principe à respecter dans cette opération est d'obtenir des **classes de même effectif** plutôt que des intervalles de même amplitude. Cette procédure de découpage est toujours prévue dans les programmathèques complètes d'analyse des données.

Il existe des justifications théoriques à cette pratique. Un certain nombre d'arguments directs militent pour ce choix.

1. Les modalités représentant un ensemble d'individus, il est souhaitable, pour que leur comparaison ait un sens, que ces ensembles soient analogues du point de vue de leur effectif. Cela est particulièrement important en ACM où la distance d'une modalité au barycentre croît quand son effectif décroît.
2. Cette procédure évite les modalités d'effectif trop faible dont nous avons souligné l'effet perturbateur. Par ailleurs le profil de ces modalités est très sensible à de faibles variations des individus étudiés ; cela est particulièrement gênant lorsque ces individus proviennent d'un échantillonnage dans une population.

4.6 ANALYSE FACTORIELLE DE DONNÉES MIXTES (AFDM)

Il est fréquent de souhaiter réaliser une analyse factorielle sur un tableau croisant des individus et des variables des deux types, quantitatives ou qualitatives, ce que nous appelons des données mixtes. Dans cette perspective, il convient de bien distinguer deux cas, selon que les variables actives sont de même type ou mixtes.

Lorsque toutes les variables actives sont quantitatives, le problème revient à introduire des variables qualitatives illustratives dans une ACP (*cf.* section 1.10). Lorsque les variables actives sont qualitatives, le problème revient à introduire des variables quantitatives illustratives dans une ACM. Pour cela, on calcule les coefficients de corrélation entre les variables quantitatives et les facteurs de l'ACM ; cette démarche est la même qu'en ACP et conduit au même type de graphique : le cercle des corrélations.

La prise en compte simultanée des deux types de variables en tant qu'éléments actifs d'une même analyse a été l'objet du paragraphe précédent : le codage, en classes, de variables quantitatives est une méthodologie excellente mais qui trouve ses limites dans deux cas :

– Lorsque le nombre d'individus est faible, disons inférieur à 100 pour fixer les idées, l'ACM est souvent instable vis-à-vis de l'ajout ou de la suppression d'un petit nombre d'individus et de variables.
– Lorsque le nombre de variables qualitatives et très faible en regard du nombre de variables quantitatives ; concrètement, l'utilisateur qui pressent surtout des liaisons linéaires hésitera à coder en classes vingt variables quantitatives avec pour seul objet de prendre en compte (en actif) une seule variable qualitative.

Dans ces deux cas, on pourra recourir à l'Analyse factorielle de Données Mixtes (AFDM). Le principe tient en quatre points.

1. On considère l'espace R^I des fonctions définies sur I. Dans cet espace (muni de la métrique des poids des individus), on représente simultanément les variables quantitatives comme en ACP normée (une variable = un vecteur de longueur 1) et les variables qualitatives comme en ACM (une variable = l'ensemble des indicatrices de ses modalités = le sous-espace engendré par ces indicatrices).

2. On adopte le point de vue de l'analyse factorielle selon lequel les facteurs F_s sont liés le plus possible aux variables actives. Ainsi, en ACP, la quantité maximisée s'écrit (en notant r le coefficient de corrélation ; *cf.* section 1.6)

$$\sum_k r^2 (k, F_s)$$

En ACM, elle s'écrit (en notant η^2 le carré du rapport de corrélation ; *cf.* section 4.3.6) :

$$\sum_j \eta^2 (j, F_s)$$

Dans le cas de données mixtes, il est naturel de proposer le critère suivant :

$$\sum_k r^2(k, F_s) + \sum_j \eta^2(j, F_s)$$

Ce critère équilibre le rôle de chacune des variables quel que soit son type ; cet équilibre implique que les variables quantitatives soient centrées et réduites.

3. Pour réaliser pratiquement une AFDM (en l'absence d'un logiciel *ad hoc*), on juxtapose le tableau des variables quantitatives centrées réduites et le tableau disjonctif complet dans lequel les valeurs « 1 » pour la modalité k sont remplacées par $\sqrt{I_k}$. Ce tableau est ensuite soumis à une ACP non normée.

4. Les trois graphiques de base de l'AFMD représentent :

– les individus comme en ACP ou en ACM ;
– les variables quantitatives comme en ACP (cercle des corrélations) ;
– les modalités des variables qualitatives comme en ACP c'est-à-dire à l'exact barycentre des individus qui les possèdent (et non pas au coefficient $\sqrt{\lambda_s}$ près comme en ACM).

À ces graphiques, on ajoute celui des variables des deux types construit de la façon suivante : la coordonnée de la variable quantitative k sur l'axe de rang s est $r^2(k, F_s)$; celle de la variable qualitative j vaut $\eta^2(j, F_s)$. Ce graphique a déjà été introduit pour l'ACM (**Figure 4.6**) ; il montre simultanément les liaisons entre les variables des deux types et les facteurs (d'où sa dénomination « carré des liaisons ») mais s'interprète aussi, pour les variables actives, en terme de contributions au critère (une autre interprétation, géométrique, sera donnée en 8.4 à propos de l'AFM). Le carré des liaisons peut-être construit à partir de n'importe quelle analyse factorielle appliquée à un tableau dont les lignes sont des individus (ACP, ACM, AFDM, AFM).

4.7 CONCLUSION

L'ACM est une méthode d'étude de plusieurs variables qualitatives définies sur un ensemble d'individus. Sa problématique est très riche et va bien au-delà d'une simple mise en œuvre de l'AFC sur un tableau particulier.

C'est là un des aspects de l'équivalence entre l'AFC sur le TDC et sur le tableau de Burt. Il existe d'ailleurs d'autres équivalences que celles déjà citées ; des points de vue très différents sur l'étude de variables qualitatives ont induit la conception de méthodes qui conduisent, au moins partiellement, aux mêmes résultats que l'AFC sur le TDC (*cf.* section 8.6).

Outre qu'elles permettent de considérer l'ACM comme une méthode à part entière, ces convergences la renforcent. Les mécanismes de l'ACM, supportant plusieurs interprétations, sont d'une part adaptés à une vaste palette de problèmes concrets et d'autre part fournissent des résultats en accord avec plusieurs points de vue.