
*Cours et études de cas*

**Masters et Écoles d'ingénieurs**

# ANALYSES FACTORIELLES SIMPLES ET MULTIPLES

## Objectifs, méthodes et interprétation

4e édition

***Brigitte Escofier***
***Jérôme Pagès***



DUNOD

# ANALYSES FACTORIELLES SIMPLES ET MULTIPLES

## Objectifs, méthodes et interprétation

# Consultez nos parutions sur dunod.com

# ANALYSES FACTORIELLES SIMPLES ET MULTIPLES

## Objectifs, méthodes et interprétation

***Brigitte Escoffier***
Ancien professeur à l'Université de Rennes et à l'IUT de Vannes

***Jérôme Pagès***
Ingénieur agronome, professeur à l'Agrocampus de Rennes

**4e édition**

DUNOD

Illustration de couverture : © *Digitalvision*





ISBN 978-2-10-053809-6

# Table des matières

# Introduction

### *L'analyse des données : outil de connaissance dans les domaines les plus divers*

Depuis une trentaine d'années, les méthodes d'analyse des données ont largement démontré leur efficacité dans l'étude de grandes masses complexes d'informations. Ce sont des méthodes dites multidimensionnelles en opposition aux méthodes de la statistique descriptive qui ne traitent qu'une ou deux variables à la fois. Elles permettent donc la confrontation entre de nombreuses informations, ce qui est infiniment plus riche que leur examen séparé. Les représentations simplifiées de grands tableaux de données que ces méthodes permettent d'obtenir s'avèrent un outil de synthèse remarquable. De données trop nombreuses pour être appréhendées directement, elles extraient les tendances les plus marquantes, les hiérarchisent et éliminent les effets marginaux ou ponctuels qui perturbent la perception globale des faits.

Nées à l'université, elles ont d'abord été connues essentiellement des chercheurs et appliquées à des domaines scientifiques comme l'écologie, la linguistique, l'économie, etc. Elles ont permis d'aborder des études nouvelles plus riches et plus complexes. Mais leur domaine d'application déborde depuis longtemps ce cadre universitaire, surtout depuis que l'acquisition et le stockage des informations sont facilités par le développement de l'informatique. Dans tous les domaines (marketing, assurance, banque, etc.), d'importants fichiers de données sont accumulés. Le premier objectif est de conserver les informations et de pouvoir les consulter facilement. Mais on s'aperçoit vite que pour exploiter l'ensemble de l'information contenue dans ces fichiers, dont le recueil est souvent coûteux, il est nécessaire de disposer d'outils statistiques adaptés.

### *Puissance des représentations géométriques de l'analyse factorielle*

Parmi les méthodes de l'analyse des données, l'analyse factorielle tient une place primordiale. Elle est utilisée soit seule, soit conjointement avec des méthodes de classification (alors que ces dernières sont rarement appliquées seules). Cette place de choix tient en partie aux représentations géométriques des données, qui transforment en distances euclidiennes des proximités statistiques entre éléments.

Elles permettent d'utiliser les facultés de perception dont nous usons quotidiennement : sur les graphiques de l'analyse factorielle, on voit, au sens propre du terme

(avec les yeux et l'analyse assez mystérieuse que notre cerveau fait d'une image), des regroupements, des oppositions, des tendances, impossibles à discerner directement sur un grand tableau de nombres, même après un examen prolongé.

Ces représentations graphiques sont aussi un moyen de communication remarquable car point n'est besoin d'être statisticien pour comprendre que la proximité entre deux points traduit la ressemblance entre les objets qu'ils représentent.

### *L'analyse factorielle ou les analyses factorielles ?*

Les deux expressions se justifient.

1. Il existe plusieurs méthodes adaptées à différents types de données : ainsi, pour citer les plus connues, l'analyse en composantes principales (ACP) traite des tableaux croisant des individus et des variables quantitatives, l'analyse factorielle des correspondances (AFC) traite des tableaux de fréquence et l'analyse des correspondances multiples (ACM) s'applique à des tableaux croisant des individus et des variables qualitatives.
2. Le principe de ces méthodes est unique. Deux nuages de points, représentant respectivement les lignes et les colonnes du tableau étudié, sont construits et représentés sur des graphiques. Les représentations des lignes et des colonnes sont fortement liées entre elles.

### *Rigueur et souplesse des méthodes d'analyse factorielle*

Le fait que l'analyse factorielle ne s'applique qu'à des tableaux rectangulaires peut paraître au premier abord une limitation importante à la fois sur le type de données et sur la manière de les aborder. En réalité, la plupart des études de données peuvent être formalisées comme une analyse de tableaux rectangulaires. D'autre part, un même fichier de données peut conduire à un grand nombre de tableaux différents et donc à des analyses différentes qui permettent chacune d'étudier un des aspects du problème.

La construction de tableaux à partir d'un fichier initial est appelée codage. Ce terme de codage inclut la transformation de données brutes en variables quantitatives ou qualitatives, le choix des lignes et des colonnes du tableau, celui des éléments à traiter en actif, etc. Dans cette étape de codage, la marge de manœuvre est presque infinie. Le résultat d'une analyse factorielle est unique, ce qui en assure la rigueur, mais les analyses possibles sont nombreuses, ce qui en assure la souplesse et la faculté d'adaptation.

### *Les tableaux multiples*

Les analyses factorielles ont été conçues pour étudier un tableau de données unique. Or, les personnes qui analysent des données sont de plus en plus fréquemment confrontées à l'étude simultanée de plusieurs tableaux rectangulaires. Il s'agit le plus souvent :

1. d'une suite de tableaux indicés par le temps ;
2. d'un ensemble de tableaux rectangulaires provenant d'un unique tableau de dimension trois ;
3. d'un tableau initialement unique mais dans lequel on distingue des sous–tableaux (ce cas général inclut le cas particulier dans lequel un ensemble d'individus est décrit à la fois par des variables quantitatives et des variables qualitatives).

Au fil des ans, des méthodologies ont été mises au point. On se ramène généralement à l'analyse d'un tableau complexe formé par la juxtaposition des différents tableaux. Ces méthodes fondées sur les méthodes d'analyse classique, elles-mêmes conçues pour l'étude d'un tableau simple, utilisent largement la technique dite des « éléments supplémentaires ». Mais ces techniques ont leurs limites et les objectifs spécifiques de l'analyse des « tableaux multiples » ne sont pas tous atteints. Aussi, de nouvelles méthodes, utilisant les mêmes principes fondamentaux que les analyses factorielles « classiques » mais prenant en compte le caractère « multiple » des tableaux, ont été mises au point.

### *Esprit du livre*

Cet ouvrage est destiné avant tout aux utilisateurs d'analyse des données. C'est pourquoi il présente des méthodes d'analyse factorielle en tentant de dégager leurs objectifs et les interprétations de leurs résultats. Pour en faciliter la lecture aux non-spécialistes, nous avons pris le parti de séparer le plus possible les aspects intuitifs des méthodes (objectifs, principe général et représentations géométriques), des aspects mathématiques et théoriques. Les aspects intuitifs ne nécessitent qu'un très faible bagage statistique et mathématique et sont donc abordables par beaucoup. Ils sont largement commentés sur quatre exemples.

Les aspects théoriques sont regroupés essentiellement dans deux chapitres. Leur but est de fournir les justifications des méthodes en précisant les critères optimisés et les algorithmes de calcul. La bibliographie est restreinte au minimum : lorsqu'une démonstration risque d'alourdir trop le texte, une note en bas de page renvoie à une référence plus complète.

**Les objectifs.** Devant un jeu de données à analyser, se pose le problème du choix du traitement statistique, c'est-à-dire du choix du couple indissociable codage-méthode. Pour bien choisir, il est nécessaire de connaître les moyens dont on dispose, donc les possibilités des méthodes qui peuvent répondre chacune à un certain nombre d'objectifs précis. La réflexion sur les objectifs d'une étude est fondamentale. Elle est plus efficace si elle se fait dans le cadre des possibilités techniques. Cette réflexion doit toujours intervenir le plus tôt possible car elle influe non seulement sur le traitement statistique mais aussi sur le recueil même des données.

**L'interprétation.** L'analyse effectuée, le travail du statisticien n'est pas terminé : il faut interpréter les résultats. Cette phase, qui peut sembler délicate au néophyte, fait intervenir à la fois la connaissance du problème et celle des méthodes.

### *Contenu du livre*

Ce livre contient à la fois un rappel des méthodes classiques, des exposés des méthodologies d'analyse des tableaux multiples basées sur ces dernières et une introduction aux méthodes d'analyse spécifiques de ces tableaux. Ces dernières ont été conçues par les auteurs et exposées dans le cadre de leurs recherches, mais cet ouvrage est le premier qui en contient une présentation générale destinée aux utilisateurs. L'interprétation des résultats d'une analyse factorielle, qui est avec le codage la phase la plus délicate de l'étude, est illustrée par quatre exemples tout le long du texte ; elle fait aussi l'objet d'une réflexion générale.

La première partie du livre, qui comprend cinq chapitres, présente les méthodes classiques d'analyse factorielle : l'ACP, l'AFC et l'ACM. Le traitement d'un exemple par ACP donne l'occasion de présenter une méthode de classification et son dépouillement conjointement avec celui d'une analyse factorielle. Une présentation formalisée de l'ACP, de l'AFC et de l'ACM, incluant les démonstrations essentielles, est faite dans un cadre commun à ces trois méthodes.

La deuxième partie est consacrée aux tableaux multiples. Les chapitres 6, 7, 8 et 9 concernent l'étude simultanée de plusieurs tableaux croisant les mêmes individus et différents groupes de variables numériques ou qualitatives. Le chapitre 6 commente plusieurs traitements de la même enquête par les méthodes classiques. C'est à la fois une illustration des méthodes présentées dans les premiers chapitres, une réflexion sur les objectifs généraux de l'étude de tableaux comprenant plusieurs groupes de variables, et un bilan sur l'intérêt et les limites des méthodologies basées sur ces méthodes. L'analyse factorielle multiple (AFM), conçue pour ce type de données, est introduite dans le chapitre 7 à partir des résultats issus de son application à un second exemple ; sa présentation complète constitue le chapitre 8 ; une réflexion sur son utilisation constitue le chapitre 9. Le chapitre 10 traite des tableaux de fréquence ternaires et plus généralement de l'étude simultanée de plusieurs tableaux de fréquence binaires. Bien qu'il s'agisse comme dans les quatre chapitres précédents de tableaux multiples, la nature des données (fréquences au lieu de variables) implique des objectifs fondamentalement différents. Ce chapitre tente d'en dégager les principaux et illustre sur un même exemple les méthodologies dérivées de l'AFC et une technique nouvelle, baptisée analyse intra, qui permet d'étudier un aspect spécifique des tableaux de fréquence ternaire : les liaisons conditionnelles.

La dernière partie, réduite à un chapitre, est entièrement consacrée à l'interprétation des résultats en analyse factorielle. Elle est issue en partie des réflexions d'un groupe de

travail[1] réuni par l'ADDAD[2] dans le cadre d'un contrat avec la Société THOMSON. A partir des expériences confrontées et du regroupement de commentaires épars d'applications d'analyse factorielle, nous avons construit un guide. Ce guide propose une démarche générale d'interprétation en analyse factorielle en différenciant ACP, AFC, ACM et AFM.

Il est conseillé aux lecteurs novices en analyse des données de commencer la lecture de cet ouvrage par les deux premières fiches techniques incluses dans le chapitre 12. Ces deux fiches détaillent les représentations géométriques des nuages d'individus et de variables utilisées systématiquement en analyse factorielle. La troisième fiche, plus technique, est destinée plutôt aux lecteurs qui souhaitent approfondir les aspects mathématiques et théoriques développés dans les chapitres 5 et 8.

L'index systématique reprend l'ensemble des notions essentielles.

### *Note sur la quatrième édition*

Pour cette quatrième édition, le texte a été révisé et augmenté notamment sur deux points qui correspondent à une demande croissante des utilisateurs :

1. l'analyse simultanée de variables quantitatives et qualitatives, sans transformer les variables quantitatives ; pour cela, une présentation de l'analyse factorielle sur données mixtes (AFDM) a été incluse ;
2. la prise en compte d'une structure hiérarchique sur les variables dans un tableau individus × variables ; l'exemple classique est celui d'un questionnaire dont les questions sont structurées en thèmes et sous-thèmes ; ce livre contient maintenant une présentation de l'Analyse Factorielle Multiple Hiérarchique (AFMH), prolongement naturel de l'AFM adapté à ce type de données.

Ces méthodes, ainsi que toutes celles décrites dans ce livre, sont désormais disponibles dans FactoMineR, logiciel libre d'analyse des données développé par le laboratoire de mathématiques appliquées d'Agrocampus.

Au terme de ce travail, il est agréable de remercier Radwan JALAM, ingénieur informaticien à Agrocampus, qui a assuré la mise en forme de cette nouvelle édition.



1. Ch. Bastin, Ch. Bourgarit, J. Confais, B. Escofier, B. Gomel, J.P. Fénelon, J.Pagès.
2. L'Association pour le Développement et la Diffusion de l'Analyse des Données diffuse aussi les logiciels correspondants à toutes les méthodes décrites

Chapitre **1**

# Analyse en Composantes Principales

## 1.1 DONNÉES ET OBJECTIFS DE L'ÉTUDE

L'Analyse en Composantes Principales (ACP) s'applique à des tableaux croisant des individus et des variables quantitatives, appelés de façon concise tableaux *Individus*×*Variables quantitatives*.

Selon un usage bien établi, les lignes du tableau représentent les individus et les colonnes représentent les variables. A l'intersection de la ligne $i$ et de la colonne $k$ se trouve la valeur de la variable $k$ pour l'individu $i$. La **figure 1.1** illustre ces notions et complète les notations. Le tableau 2.1 page 32 en est un exemple.

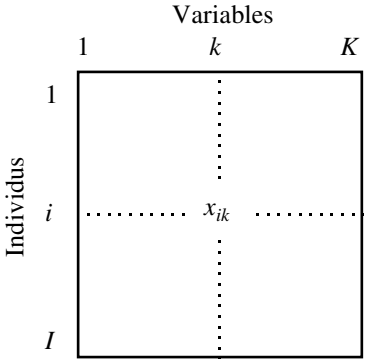



**Figure 1.1** Tableau des données en ACP. $x_{ik}$ : valeur de la variable $k$ pour l'individu i. $I$ : nombre d'individus et ensemble des individus. $K$ : nombre de variables et ensemble des variables.

Les termes *individu* et *variable* recouvrent des notions différentes. Par exemple, dans le tableau étudié au chapitre 6, les individus sont des vins et les variables sont des critères décrivant ces vins (acidité, astringence, etc.). Les questions que l'on se pose sur les individus et celles que l'on se pose sur les variables ne sont pas de même nature.

À propos de deux **individus**, on essaie d'évaluer leur **ressemblance** : deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables. En ACP, la distance d($i$,$l$) entre deux individus $i$ et $l$ est définie par :

$$\mathrm{d}^2(i,l) = \sum_{k \in K} (x_{ik} - x_{lk})^2$$

À propos de deux **variables**, on essaie d'évaluer leur **liaison**. En ACP, la liaison entre deux variables est mesurée par le coefficient de corrélation linéaire (dans de rares situations, on utilise la covariance), noté usuellement $r$. Soit :

$$\begin{aligned} r(k,h) &= \frac{covariance(k,h)}{\sqrt{variance(k) \times variance(h)}} \\ &= \frac{1}{I} \sum_{i \in I} \left( \frac{x_{ik} - \bar{x}_k}{s_k} \right) \left( \frac{x_{ih} - \bar{x}_h}{s_h} \right) \end{aligned}$$

avec $\bar{x}_k$ et $s_k$ la moyenne et l'écart-type de la variable $k$.

Appliquée à un tel tableau, l'objectif général de l'ACP est une étude exploratoire. Les deux voies principales de cette exploration sont :

***Un bilan des ressemblances entre individus***. On cherche alors à répondre à des questions du type suivant : quels sont les individus qui se ressemblent ? Quels sont ceux qui diffèrent ? Plus généralement, on souhaite décrire la variabilité des individus. Pour cela, on cherche à mettre en évidence des groupes homogènes d'individus dans le cadre d'une **typologie des individus**. Selon un autre point de vue, on cherche les **principales dimensions de variablité** des individus.

***Un bilan des liaisons entre variables***. Les questions sont alors : quelles variables sont corrélées positivement entre elles ? Quelles sont celles qui s'opposent (corrélées négativement) ? Existe-t-il des groupes de variables corrélées entre elles ? Peut-on mettre en évidence une **typologie des variables** ?

Un autre aspect de l'étude des liaisons entre variables consiste à résumer l'ensemble des variables par un petit nombre de **variables synthétiques** appelées ici **composantes principales**. Ce point de vue est très lié au précédent : une composante principale peut être considérée comme le représentant (la synthèse) d'un groupe de variables liées entre elles.

Naturellement, ces deux voies ne sont pas indépendantes du fait de la dualité inhérente à l'étude d'un tableau rectangulaire : la structure du tableau peut être analysée à

la fois par l'intermédiaire de la typologie des individus et de la typologie des variables. Aussi, cherche-t-on en général à relier ces deux typologies. Pour cela, on caractérise les classes d'individus par des variables (on sélectionne ainsi les variables pour lesquelles l'ensemble des individus d'une classe possède des valeurs particulièrement grandes ou particulièrement petites). De même, on caractérise un groupe de variables liées entre elles par des individus types (on sélectionne ainsi les individus qui possèdent des valeurs particulièrement grandes ou des valeurs particulièrement petites pour un ensemble de variables liées positivement entre elles). Enfin, dans la situation idéale, les deux typologies peuvent être « superposées » : chaque groupe de variables caractérise un groupe d'individus et chaque groupe d'individus rassemble les individus types d'un groupe de variables. Ajoutons enfin que la notion de principale dimension de variablité des individus rejoint celle de variable synthétique.

### *a) Poids des individus*

Dans la plupart des cas, les individus jouent le même rôle. Nous nous sommes situés implicitement dans cette situation jusqu'ici, en affectant le même poids à chaque individu. Par commodité, on choisit ces poids tels que la masse totale de ces individus soit égale à 1 : à chaque individu on associe alors le poids $1/I$. Toutefois, dans certains cas, on peut souhaiter attribuer des poids différents aux individus. Cette situation se présente notamment lorsque les individus représentent chacun une sous-population ; on affecte alors à un individu un poids proportionnel à l'effectif de la sous-population qu'il représente. Ce poids intervient dans le calcul de la moyenne de chaque variable (c'est-à-dire dans la définition d'un individu théorique moyen), dans le calcul de la variance de chaque variable et dans celui de la mesure de liaison (le coefficient de corrélation) entre les variables. Soit, en appelant $p_i$ le poids affecté à l'individu $i$ ($\sum_i p_i = 1$) :

$$\bar{x}_k = \sum_i p_i x_{ik} \quad s_k^2 = \sum_i p_i (x_{ik} - \bar{x}_k)^2$$

$$r(k,h) = \sum_i p_i \left( \frac{x_{ik} - \bar{x}_k}{s_k} \right) \left( \frac{x_{ih} - \bar{x}_h}{s_h} \right)$$

Les programmes complets d'ACP permettent tous d'introduire des poids d'individus.

### *b) Poids des variables*

Nous avons accordé jusqu'ici la même importance *a priori* aux différentes variables. On est très rarement conduit, dans la pratique, à souhaiter leur affecter des importances différentes. À tel point que les programmes courants d'ACP ne le permettent pas. Cette importance peut être modulée à l'aide d'un coefficient appelé poids de la variable. En appelant $m_k$ le poids de la variable $k$, la distance entre deux individus $i$ et $l$ est définie par :

$$\mathrm{d}^2(i,l) = \sum_{k \in K} m_k (x_{ik} - x_{lk})^2$$

Toutefois, comme nous le verrons dans le chapitre 5 qui contient l'ensemble des résultats techniques concernant les analyses factorielles, ces poids ne modifient en rien les principes généraux de l'analyse. Afin de ne pas alourdir l'exposé de ce chapitre, nous considérons dans la suite que les individus possèdent le même poids ($p_i = 1/I$ quel que soit $i \in I$) ainsi que les variables ($m_k = 1$ quel que soit $k \in K$).

## 1.2 TRANSFORMATION DES DONNÉES

En ACP, le tableau des données est toujours centré (en pratique, le centrage est inclus dans les programmes d'ACP). A chaque valeur numérique, on soustrait la moyenne de la variable en cause. Le tableau obtenu est alors de terme général :

$$x_{ik} - \bar{x}_k$$

Cette transformation n'a aucune incidence sur les définitions de la ressemblance entre individus et de la liaison entre variables. À ce niveau, elle peut être considérée comme un intermédiaire technique qui présente d'intéressantes propriétés mais qui ne change fondamentalement rien à la problématique.

L'ACP peut être réalisée sur des données seulement centrées. Toutefois, ses résultats sont alors très sensibles au choix des unités de mesure. Généralement, ce choix est arbitraire : ainsi, dans l'exemple classique de mensurations d'animaux, la variable *hauteur* peut être exprimée en mètres ou en centimètres. Or ce choix a une grande influence sur la mesure de ressemblance entre individus. Le passage du mètre au centimètre multiplie par $100^2$ l'influence de la variable *hauteur* dans le calcul du carré de la distance entre deux individus.

La façon classique de s'affranchir de l'arbitraire des unités de mesure est de réduire les données. Le tableau obtenu a pour terme général $(x_{ik} - \bar{x}_k)/s_k$. Ce faisant, on utilise comme unité de mesure pour la variable $k$, son écart-type $s_k$. Toutes les variables présentent alors la même variabilité et de ce fait la même influence dans le calcul des distances entre individus.

Dans les études où toutes les variables s'expriment dans la même unité, on peut souhaiter ne pas réduire les variables. En procédant ainsi, on accorde à chaque variable réduite un poids égal à sa variance (*cf.* définition de la distance entre individus). Selon un autre point de vue, la définition de $\mathrm{d}(i,l)$ montre que la variance de la variable $k$ est égale à la contribution moyenne de la variable $k$ au carré de la distance entre individus. Cela se déduit de l'écriture suivante de la variance :

$$s_k^2 = \frac{1}{2I^2} \sum_{i,l} (x_{ik} - x_{lk})^2$$

Un exemple de discussion de l'opportunité de la réduction est donné section 2.1.2 page 32. Dans la suite, sauf mention explicite du contraire, les variables sont toujours supposées centrées et réduites.

## 1.3 NUAGE DES INDIVIDUS

S'intéresser aux individus revient à envisager le tableau en tant que juxtaposition de lignes. À chaque individu est associée une suite de $K$ nombres. Selon ce point de vue, un individu peut être représenté comme un point de l'espace vectoriel à $K$ dimensions, noté $R^K$, dont chaque dimension représente une variable. L'ensemble des individus constitue le nuage $N_I$ dont le centre de gravité G est confondu avec l'origine O des axes du fait du centrage ; G représente l'individu moyen précédemment cité. Ces notations sont rassemblées **figure 1.2**.

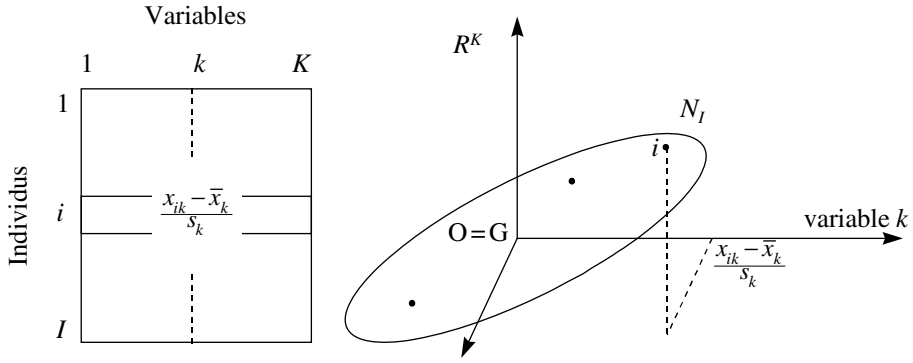



**Figure 1.2** Tableau des données et nuage des individus associé dans l'espace $R^K$. Du fait du centrage, l'origine des axes est confondue avec le centre de gravité du nuage.

Dans l'espace $R^K$, la notion de ressemblance entre deux individus introduite section 1.1 n'est autre que la distance euclidienne usuelle. Cette interprétation géométrique constitue une justification a posteriori décisive du choix de la mesure de ressemblance : le fait qu'elle soit une distance euclidienne lui confère un grand nombre de propriétés mathématiques indispensables pour la suite.

L'ensemble des distances inter-individuelles constitue ce que l'on appelle la forme du nuage $N_I$. Réaliser un bilan de ces distances revient à étudier la forme du nuage $N_I$, c'est-à-dire à y déceler une partition des points (la typologie mentionnée dans les objectifs) ou des directions d'allongement remarquables (les principales dimensions de variabilité).

Dès que $K$ est supérieur à 3, l'étude directe du nuage $N_I$ est impossible du fait de la limitation à trois dimensions de notre sens visuel. D'où l'intérêt des méthodes

factorielles en général, et dans ce cas particulier de l'ACP, qui fournissent des images planes approchant le mieux possible (au sens d'un critère défini et discuté section 1.5) un nuage de points situé dans un espace de grande dimension.

## 1.4 NUAGE DES VARIABLES

S'intéresser aux variables revient à envisager le tableau en tant que juxtaposition de colonnes. À chaque variable, est associée une suite de $I$ nombres. Selon ce point de vue, une variable peut être représentée comme un vecteur de l'espace vectoriel à $I$ dimensions, noté $R^I$, dont chaque dimension représente un individu : par exemple, la variable $k$ est représentée par le vecteur noté lui aussi $k$ et dont la $i^{\text{e}}$ composante est $(x_{ik} - \bar{x}_k)/s_k$. L'ensemble des extrémités des vecteurs représentant les variables constitue le nuage $N_K$. Ces notations sont regroupées dans la **figure 1.3**.

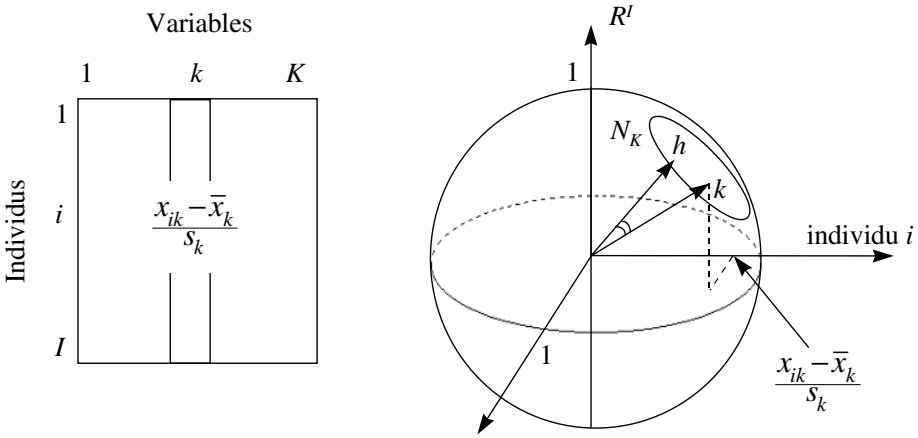



**Figure 1.3** Tableau des données et nuage des variables associé dans l'espace $R^I$.
$\cos(\widehat{Oh}, \widehat{Ok}) = r(h,k)$

Le choix de la distance dans $R^I$ consiste à affecter à chaque dimension un coefficient égal au poids de chaque individu dans le nuage $N_I$ de $R^K$ (on peut avoir l'intuition de ce choix en considérant deux individus absolument identiques que l'on peut remplacer par un seul ayant un poids double). Dans le cas général où ces poids sont identiques, la distance utilisée est, au coefficient $1/I$ près, la distance euclidienne usuelle. Avec cette distance, les vecteurs représentant les variables centrées ont les propriétés suivantes :

1. La norme de chaque vecteur représentant une variable est égale à son écart-type. Soit :

$$\|\text{variable } k\|^2 = \sum_{i=1}^{I} \frac{1}{I} (x_{ik} - \bar{x}_k)^2$$

Ainsi, lorsque les variables sont centrées réduites, chaque variable a pour longueur 1 : le nuage $N_K$ est alors situé sur une sphère de rayon 1 (on dit aussi hypersphère pour rappeler que $R^I$ est de dimension supérieure à 3). Pour cette raison, l'ACP sur données centrées-réduites est dite **ACP normée**. Lorsque les variables sont seulement centrées, leur longueur est égale à leur écart-type et on parle alors d'**ACP non normée**.

2. Le cosinus de l'angle formé par les vecteurs représentant les deux variables $h$ et $k$, obtenu en calculant le produit scalaire noté $\langle h, k\rangle$ entre ces deux vecteurs normés, est égal au coefficient de corrélation entre ces deux variables. Soit :

$$\cos(h,k) = \langle h,k\rangle = \sum_i \frac{1}{I}\left(\frac{x_{ih}-\bar{x}_h}{s_h}\right)\left(\frac{x_{ik}-\bar{x}_k}{s_k}\right) = \text{corrélation}(h,k)$$

L'interprétation d'un coefficient de corrélation comme un cosinus est une propriété très importante puisqu'elle donne un support géométrique, donc visuel, au coefficient de corrélation. Cette propriété nécessite le centrage, ce qui justifie cette transformation présentée section 1.1 comme un intermédiaire technique. Elle justifie aussi le choix de la distance (on dit aussi *métrique*) dans $R^I$ et implique que, dans la représentation des variables, on s'intéresse surtout aux directions déterminées par les variables, c'est-à-dire aux vecteurs plutôt qu'à leurs extrémités.

La longueur des vecteurs représentant les variables étant égale à 1, la coordonnée de la projection d'une variable sur une autre s'interprète comme un coefficient de corrélation.

➤ Conclusion

Réaliser un bilan des coefficients de corrélation entre les variables revient à étudier les angles entre les vecteurs définissant le nuage $N_K$. Cette étude directe est impossible du fait de la dimension de $R^I$. L'intérêt de l'ACP est de fournir des variables synthétiques qui constituent un résumé de l'ensemble des variables initiales et sont la base d'une représentation plane approchée des variables et de leurs angles.

## 1.5 AJUSTEMENT DU NUAGE DES INDIVIDUS

L'objectif est de fournir des images planes approchées du nuage $N_I$ situé dans l'espace $R^K$ (*cf.* section 1.3). Pratiquement, on recherche une suite $\{u_s; s = 1, ..., S\}$ de $S$ directions privilégiées de $R^K$ appelées axes factoriels qui, prises deux à deux, définissent des plans factoriels sur lesquels on projette le nuage $N_I$. Chaque direction $u_s$ est choisie de façon à rendre maximum l'inertie par rapport à l'origine O (confondue avec le centre de gravité G, du fait du centrage) de la projection de $N_I$ sur $u_s$. Dans la recherche d'une suite, on impose à chaque direction d'être orthogonale aux directions déjà trouvées (*cf.* **Figure 1.4**). On peut montrer que le plan engendré par les deux

premiers axes $u_1$ et $u_2$ rend maximum l'inertie projetée sur ce plan. Il en est de même pour le sous-espace engendré par les trois premiers axes, etc.

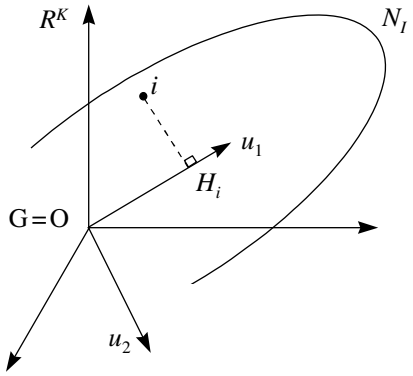



**Figure 1.4** L'ajustement du nuage des individus. L'individu $i$ se projette en $H_i$ sur $u_1$. On cherche d'abord $u_1$ qui rend maximum $\sum_i OH_i^2$. Puis on cherche $u_2$, orthogonal à $u_1$, qui satisfait le même critère et ainsi de suite. Lorsque les individus sont munis de poids $p_i$ différents, le critère consiste à rendre maximum : $\sum_i p_i OH_i^2$.

Il est équivalent de rendre maximum $\sum_i OH_i^2$ ou de rendre minimum $\sum_i iH_i^2$. Cette deuxième écriture, forme classique du critère des moindres carrés, montre que les axes factoriels rendent minimum l'écart entre le nuage des individus et sa projection.

Du fait du centrage, le critère (inertie maximum par rapport au centre de gravité G) permet d'interpréter les axes factoriels comme des directions d'allongement maximum du nuage $N_K$. On parle aussi de **principales dimensions de variabilité**, dans la mesure où ils rendent compte le plus possible de la diversité des individus.

On peut montrer que, toujours du fait du centrage, rendre maximum $\sum_i OH_i^2$ est équivalent à rendre maximum $\sum_i \sum_l (OH_i - OH_l)^2$. Cette dernière forme fait apparaître les distances entre points projetés. La projection ne pouvant que réduire la distance entre points, les axes factoriels apparaissent comme les directions telles que les distances entre points projetés ressemblent le plus possible aux distances entre les points homologues de $N_I$ (*cf.* **Figure 1.5**).

Selon les objectifs d'une analyse, on mettra en avant l'une ou l'autre des interprétations du critère.

➤ Individus supplémentaires (= illustratifs)

Fréquemment, on souhaite que certains individus n'interviennent pas dans la détermination des axes ; par contre, on souhaite connaître la position de leur projection sur les

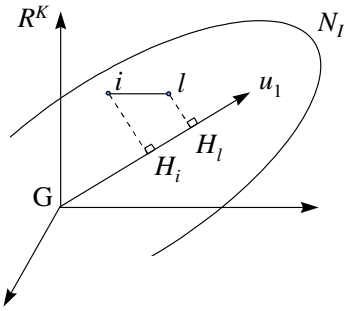



**Figure 1.5** La représentation des distances inter-individuelles. L'axe $u_1$ rend $\sum_i \sum_l (OH_i - OH_l)^2$ maximum, c'est-à-dire est tel que $\sum_i \sum_l \mathrm{d}^2(H_i - H_l)$ est le plus proche possible de $\sum_i \sum_l \mathrm{d}^2(i,l)$.

axes déterminés par les autres individus (dits *actifs*). Tous les programmes prévoient cette situation ce qui revient à mettre un poids nul à certains individus au niveau du critère d'ajustement.

Ces individus sont appelés **individus supplémentaires** (ou illustratifs). On introduit un individu en supplémentaire lorsque l'on souhaite qu'il participe à l'interprétation des plans factoriels mais non à leur construction. C'est le cas lorsque l'on dispose d'individus présentant des caractères exceptionnels, ou suspectés d'avoir été l'objet d'erreurs de mesures, ou enfin n'appartenant pas au champ strict de l'étude mais à un domaine voisin.

## 1.6 AJUSTEMENT DU NUAGE DES VARIABLES

Pour obtenir une suite de $S$ variables synthétiques $\{v_s; s = 1, ..., S\}$ et une représentation approchée des corrélations entre les variables, l'ACP applique au nuage $N_K$ des variables la même démarche qu'au nuage des individus (*cf.* **Figure 1.6**).

Le critère (inertie projetée maximum) satisfait dans le choix des axes est exactement le même que pour le nuage d'individus. Mais il prend une signification différente du fait que le nuage n'est pas centré (son centre de gravité n'est pas à l'origine) et que tous les points sont situés sur la sphère unité : ce sont les angles entre les vecteurs représentant les variables qui sont peu déformés par les projections et non pas les distances entre les points du nuage. En effet, le plan $(v_1, v_2)$, en maximisant l'inertie à l'origine du nuage projeté, rend maximum la somme des cosinus carrés des angles entre les vecteurs et leur projection : il ajuste les vecteurs et déforme donc le moins possible leurs angles.

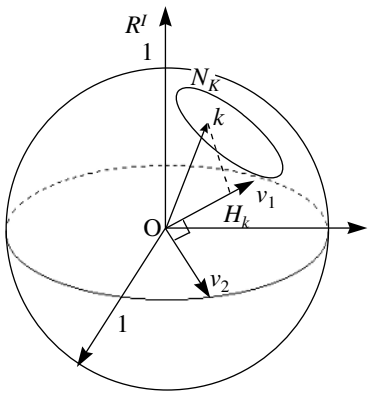



**Figure 1.6** L'ajustement du nuage des variables. $H_k$ : projection du point représentant la variable k sur $v_1$. On cherche d'abord $v_1$ qui rend maximum : $\sum_k O H_k^2$. Puis on cherche $v_2$, orthogonal à $v_1$, qui satisfait le même critère et ainsi de suite.

### ➤ Composantes principales

Le vecteur $v_1$ qui caractérise la direction d'inertie maximum définit une nouvelle variable. Les variables étudiées étant centrées et réduites, leur projection sur $v_1$ est égale à leur coefficient de corrélation avec cette variable (*cf.* section 1.4). De ce fait, rechercher le vecteur $v_1$ qui rend maximum $\sum_k O H_k^2$ équivaut à rechercher la combinaison linéaire la plus liée à l'ensemble des variables (au sens du critère : somme des carrés des corrélations maximum). En ce sens, $v_1$ est la variable qui synthétise le mieux l'ensemble des variables initiales. Les axes factoriels étant orthogonaux deux à deux, on met en évidence une suite de variables synthétiques, les composantes principales, non corrélées entre elles, qui résument au mieux l'ensemble des variables initiales.

### ➤ Variables supplémentaires (= illustratives)

Les variables, comme les individus, peuvent être traitées en éléments supplémentaires. Les variables supplémentaires sont simplement projetées sur les axes déterminés par les autres variables, dites actives. Cela permet de visualiser les corrélations entre n'importe quelle variable, même extérieure au domaine étudié, et les composantes principales.

### ➤ L'effet taille

Si, dans un jeu de données, les variables sont toutes corrélées positivement deux à deux, alors le nuage $N_K$ est loin de l'origine. Le premier axe factoriel rend alors surtout compte de la position de $N_K$ par rapport à l'origine : parallèlement, la forme

du nuage $N_K$ est mal représentée en ce sens que les projections des variables sont proches les unes des autres (*cf.* **Figure 1.7**).

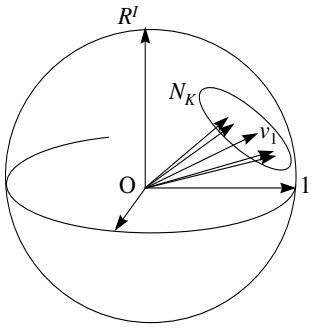



**Figure 1.7** L'effet taille dans $R^I$. Les variables, étant corrélées positivement deux à deux, forment entre elles des angles aigus. Le nuage $N_K$ est concentré sur un petit secteur de la sphère. La projection des variables sur le premier axe factoriel, défini par $v_1$, rend compte principalement de la position de $N_K$ par rapport à O.

Ce cas de figure est couramment appelé « effet taille » : il correspond à la situation dans laquelle certains individus ont des petites valeurs pour l'ensemble des variables, d'autres de grandes valeurs pour l'ensemble des variables, les autres occupant une situation intermédiaire entre ces extrêmes. Il existe donc dans ce cas une structure commune à l'ensemble des variables : c'est ce que traduit la première composante principale.

## 1.7 DUALITÉ ET FORMULES DE TRANSITION EN ACP

Le nuage $N_I$ des individus et le nuage $N_K$ des variables sont deux représentations du même tableau, l'une à travers ses lignes et l'autre à travers ses colonnes. Des relations très fortes, dites relations de dualité (démontrées en section 5.4) lient ces deux nuages.

### 1.7.1 Inerties

Tout d'abord, leur inertie totale est la même ; elle est égale au nombre de variables (lorsque les variables sont réduites) :

$$\text{Inertie totale de } N_I \text{ (ou de } N_K) = \frac{1}{I}\sum_k \sum_i \left(\frac{x_{ik} - \bar{x}_k}{s_k}\right)^2 = K$$

La projection de chacun de ces deux nuages sur une suite d'axes orthogonaux correspond à une décomposition de l'inertie totale. On peut montrer que les deux

décompositions sont identiques : les inerties des nuages $N_I$ et $N_K$ projetés sur les axes factoriels de même rang sont égales (et notées $\lambda_s$). Soit, pour les axes de rang $s$ :

$$\text{Inertie}(N_I/u_s) = \text{Inertie}(N_K/v_s) = \lambda_s$$

### 1.7.2 Facteurs

L'ensemble des projections de tous les points du nuage d'individus$N_I$ sur le $s^{\text{e}}$ axe factoriel $u_s$, appelé $s^{\text{e}}$ facteur sur les individus, constitue une nouvelle variable notée $F_s$. On montre, dans la section 5.4.1, que cette variable se confond, à la norme près, avec la $s^{\text{e}}$ composante principale $v_s$ obtenue dans l'analyse du nuage des variables. Plus précisément, le carré de la norme du facteur $F_s$ (vecteur de $R^I$), étant la somme des carrés de ses coordonnées, vaut $\lambda_s$ ; la relation entre le $s^{\text{e}}$ facteur sur $I$ et le $s^{\text{e}}$ axe factoriel de $R^I$ s'écrit donc :

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

Ces résultats sont illustrés dans la **figure 1.8**.

Ainsi, les projections planes des individus dans $R^K$ sont des représentations graphiques des couples de variables synthétiques obtenues dans $R^I$. Les résultats issus de l'étude de chacun des deux nuages possèdent fondamentalement la même signification, même s'ils s'expriment en termes d'individus pour l'un et en termes de variables pour l'autre.

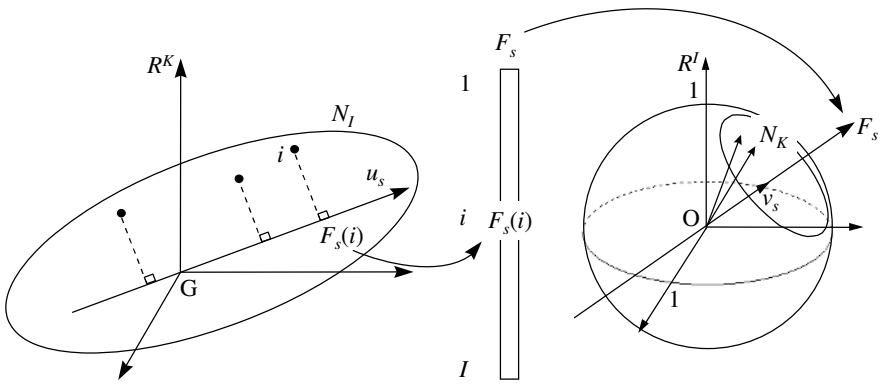



**Figure 1.8** Une des deux formes de la dualité. Les coordonnées de $N_I$ sur $u_s$ ($s^{\text{e}}$ axe factoriel de $N_I$ ) constituent le $s^{\text{e}}$ facteur sur les individus (noté $F_s$). Le vecteur $F_s$ dans $R^I$ est colinéaire à $v_s$ ($s^{\text{e}}$ axe factoriel de $N_K$).

Le rôle du nuage des individus et celui du nuage des variables sont, dans une certaine mesure, symétriques et la dualité se formule de manière analogue en échangeant le rôle des deux nuages : la projection des $K$ variables sur le $s^{\text{e}}$ axe factoriel $v_s$ de leur nuage $N_K$ définit une valeur pour chacune des $K$ variables : ces valeurs constituent le $s^{\text{e}}$ facteur sur les variables (noté $G_s$) qui est en quelque sorte un « individu » nouveau. Cette notion d'individu « type » est moins classique que celle de composante principale (pratiquement, on prend plutôt des individus réels comme individus types). Cependant, dans quelques cas particuliers, comme celui où les individus sont des courbes et les variables leurs valeurs en $K$ points de discrétisation, ces individus sont représentables et de ce fait utilisés.

On montre que le point représentant dans $R^K$ cet individu type est situé sur le $s^{\text{e}}$ axe du nuage des individus. Plus précisément :

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

Cette relation montre que, au coefficient $\sqrt{\lambda_s}$ près, les coordonnées des variables sur $v_s$ sont les coefficients de la combinaison linéaire des variables que constitue l'axe $u_s$ de $R^K$. Ainsi, la coordonnée de la variable $k$ sur $v_s$ s'interprète à la fois comme le coefficient de corrélation entre $k$ et $v_s$ et comme le coefficient de $k$ dans $u_s$ ; cette double interprétation est caractéristique des axes principaux et essentielle dans l'interprétation (à l'inverse, penser aux difficultés d'interprétation des coefficients de la régression multiple quand ils ne sont pas de même signe que les coefficients de corrélation associés). Ce résultat est illustré dans la **figure 1.9**.

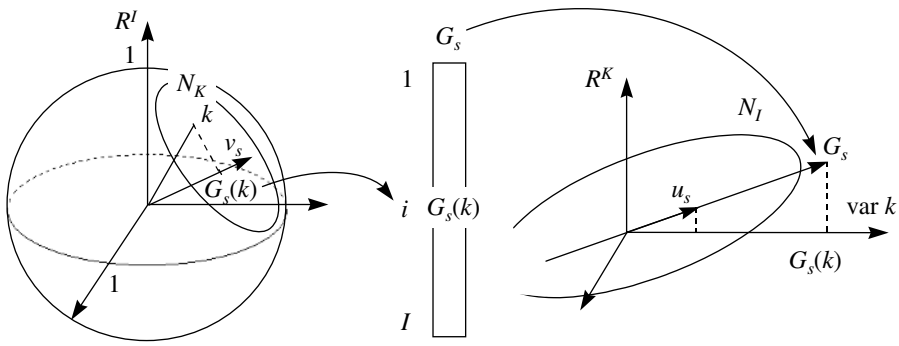



**Figure 1.9** La deuxième forme de la dualité. Les coordonnées de $N_K$ sur $v_s$ ( $s^{\text{e}}$ axe factoriel de $N_K$) constituent le $s^{\text{e}}$ facteur sur les variables (noté $G_s$). Le vecteur $G_s$ dans $R^K$ est colinéaire au $s^{\text{e}}$ axe factoriel $u_s$ de $N_I$.

### 1.7.3 Relations de transition

On appelle relations de transition entre les facteurs de rang $s$, $F_s$ et $G_s$, l'écriture algébrique des propriétés illustrées par les **figures 1.8** et **1.9**. Ces relations s'écrivent, en notant $\lambda_s$ l'inertie projetée de $N_I$ (ou de $N_K$) sur l'axe de rang $s$ :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k \frac{x_{ik} - \bar{x}_k}{s_k} G_s(k)$$

$$G_s(k) = \frac{1}{I} \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{x_{ik} - \bar{x}_k}{s_k} F_s(i)$$

La première relation exprime le fait que la projection $F_s(i)$ d'un individu $i$, est une combinaison linéaire des projections $G_s(k)$ de toutes les variables. Dans cette combinaison linéaire, le coefficient d'une variable $k$ est positif si la valeur $x_{ik}$ de cette variable pour l'individu $i$ dépasse la moyenne $\bar{x}_k$. Dans le cas contraire, ce coefficient est négatif. Ainsi, lorsque l'on regarde simultanément les deux graphiques, un individu est du côté des variables pour lesquelles il a de fortes valeurs **et** à l'opposé des variables pour lesquelles il a de faibles valeurs.

Le graphique des individus est une représentation approchée des distances interindividuelles. Celui des variables peut être considéré en tant qu'élément explicatif de cette représentation : deux individus situés à une même extrémité d'un axe sont proches car ils ont tous deux généralement de fortes valeurs pour les variables situées du même côté qu'eux **et** de faibles valeurs pour les variables situées à l'opposé.

Réciproquement, le graphique des individus peut intervenir en tant qu'aide à l'interprétation du graphique des variables : si deux variables sont très corrélées positivement, elles sont situées du même côté sur un axe. Sur l'axe correspondant du nuage d'individus, les individus qui ont de fortes valeurs pour ces deux variables se situent du même côté qu'elles et ceux qui ont de faibles valeurs se situent à l'opposé. Les individus extrêmes pour ces variables sont loin de l'origine. Les éventuels individus particuliers induisant à eux seuls des corrélations fortes sont ainsi repérés facilement.

Ainsi, en ACP, le graphique des individus et celui des variables sont à la fois optimaux en eux-mêmes (ils représentent le mieux possible l'un les individus l'autre les variables) **et** se servent mutuellement d'aides à l'interprétation. Cette propriété liant les représentations des lignes et des colonnes vaut pour toutes les analyses factorielles et leur est spécifique.

### 1.7.4 Représentation superposée

La nécessité d'une interprétation conjointe des représentations des individus et des variables conduit certains utilisateurs à les superposer. Il importe de souligner que la justification d'une telle représentation simultanée des individus et des variables est

essentiellement pragmatique : la représentation des variables aide l'interprétation de celle des individus et réciproquement. Elle pose toutefois le problème de la représentation sur un même graphique de points de natures différentes, évoluant dans des espaces différents. Cette difficulté n'est pas seulement de principe : la présence simultanée d'individus et de variables sur un même plan engendre des proximités entre individus et variables qui, à leur tour, peuvent suggérer des idées qui ne se vérifient pas dans les données. C'est pourquoi cette représentation est déconseillée. Toutefois, en conservant à l'esprit les points de repère suivants, on pourra utiliser sans danger la représentation simultanée en ACP.

1. Les formules de transition relient la coordonnée sur un axe d'**un** individu avec **l'ensemble** des coordonnées de toutes les variables sur l'axe de même rang. On ne peut interpréter la position d'un individu par rapport à une seule variable (et réciproquement).
2. Fondamentalement, les variables sont des vecteurs et non des points. Ce n'est pas la proximité entre un individu et un ensemble de points représentant des variables qui est importante mais l'éloignement de l'individu dans la direction de cet ensemble de variables.

### 1.7.5 Projection des vecteurs unitaires de la représentation des individus

Une autre idée, en vue de la représentation superposée des individus et des variables, consiste à projeter les vecteurs unitaires de $R^K$ sur les axes $u_s$. On obtient ainsi une représentation superposée plus naturelle que la précédente, en ce sens que les objets représentés proviennent du même espace.

Du fait de la relation entre $u_s$ et $G_s$, et en remarquant que la $k^{\text{e}}$ coordonnée de $u_s$ est égale à la projection sur $u_s$ du vecteur unitaire du $k^{\text{e}}$ axe de $R^K$, cette nouvelle représentation des variables est homothétique de la précédente axe par axe dans le rapport $\sqrt{\lambda_s}$.

Notre préférence va à la 1$^{\text{e}}$ représentation superposée, fondée sur les relations de transition données plus haut, car elle permet d'inclure les variables supplémentaires.

## 1.8 SCHÉMA GÉNÉRAL DE L'ACP

Nous résumons les principaux résultats de ce chapitre dans un schéma général (*cf.* **Figure 1.10**). Les numéros ci-dessous renvoient à ce schéma.

**1.** Les données brutes. Lignes (individus) et colonnes (variables) ne jouent pas des rôles symétriques : les moyennes et les variances n'ont généralement de sens que pour les colonnes.

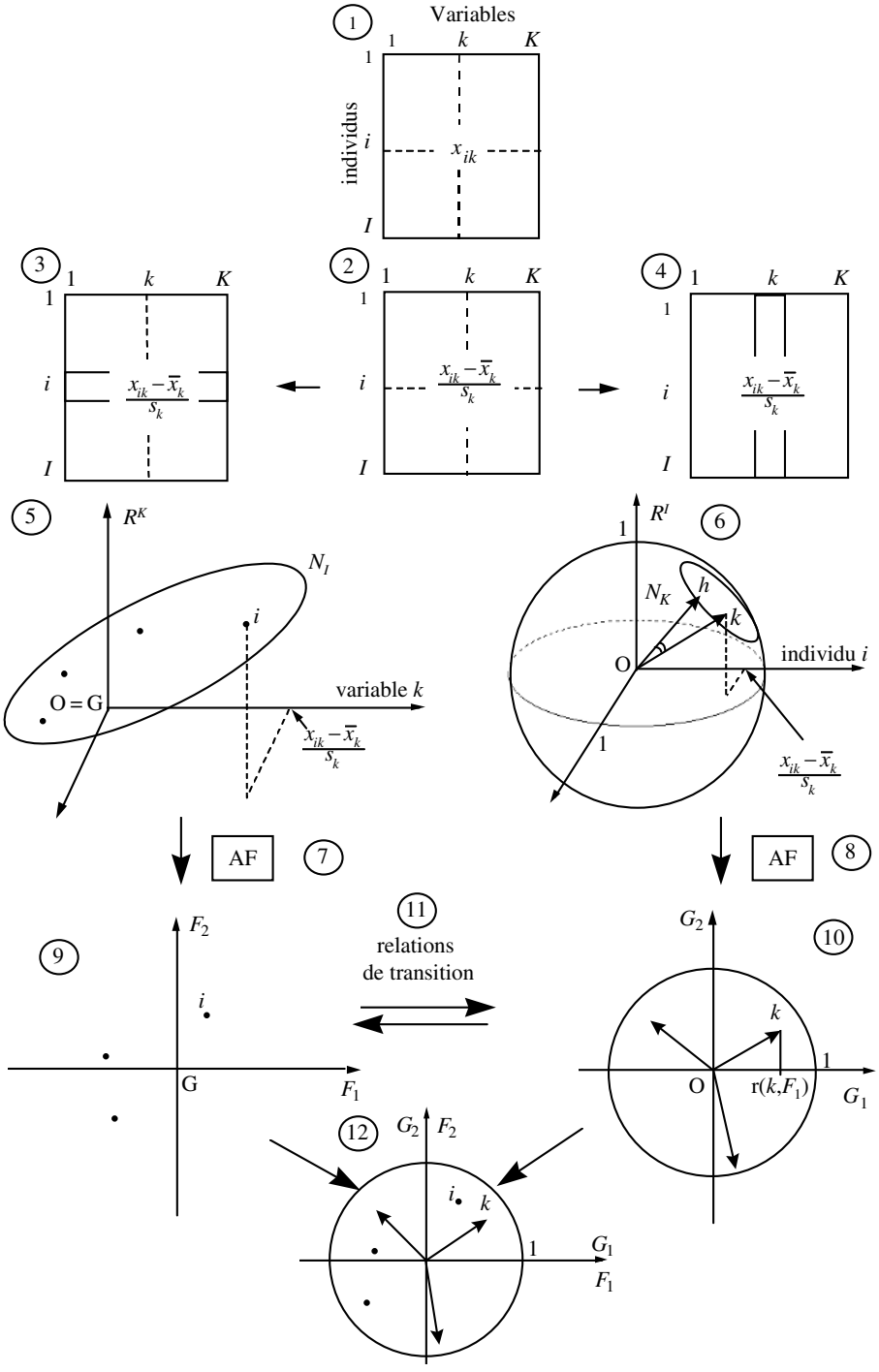



**Figure 1.10** Schéma général de l'ACP.

**2.** Les données centrées et réduites. Que l'on s'intéresse aux individus ou aux variables, le tableau est transformé de la même façon. Le centrage est surtout technique. La réduction permet de s'affranchir de l'arbitraire des unités de mesure.

**3 et 4.** Dans l'étude des individus, le tableau est considéré comme une juxtaposition de lignes. Dans l'étude des variables, le tableau est considéré comme une juxtaposition de colonnes. C'est le même tableau qui est considéré de deux façons différentes.

**5.** Un individu est une suite de $K$ nombres et peut être représenté par un point de $R^K$. Dans le nuage $N_I$, on s'intéresse aux distances inter-individuelles qui s'interprètent comme des ressemblances. Du fait du centrage, l'origine des axes est confondue avec le centre de gravité de $N_I$. Dans la plupart des cas, on affecte à chaque individu le même poids : $1/I$.

**6.** Une variable est une suite de $I$ nombres et peut être représentée par un vecteur de $R^I$. Dans le nuage $N_K$, on s'intéresse surtout aux angles entre variables. Le cosinus d'un angle entre deux variables s'interprète comme le coefficient de corrélation entre les deux variables. Du fait de la réduction, toutes les variables sont équidistantes de l'origine et donc situées sur une hypersphère de rayon 1.

**7 et 8.** L'Analyse Factorielle (AF) d'un nuage consiste à mettre en évidence une suite de directions telles que l'inertie, par rapport à O, de la projection du nuage sur ces directions est maximum. Dans $R^K$, où l'origine O est confondue avec le centre de gravité G, les axes factoriels sont les directions d'allongement maximum de $N_I$. Dans $R^I$, où la projection d'une variable sur une autre s'interprète comme un coefficient de corrélation, les axes factoriels sont les variables synthétiques les plus liées à l'ensemble des variables initiales.

**9.** Le plan factoriel croisant deux facteurs sur les individus -ici $F_1(I)$ et $F_2(I)$- fournit une image approchée de $N_I$ dans $R^K$. La distance entre deux points s'interprète comme une ressemblance.

**10.** Le plan factoriel croisant deux facteurs sur les variables -ici $G_1(K)$ et $G_2(K)$- fournit une image approchée de $N_K$ dans $R^I$. Les coordonnées d'une variable s'interprètent comme des coefficients de corrélation avec les facteurs sur les individus.

**11.** Les relations de transition expriment les résultats d'une AF (par exemple dans $R^I$) en fonction des résultats de l'autre (par exemple dans $R^K$).

**12.** Du fait des relations de transition, les interprétations des axes factoriels doivent être menées simultanément. Il peut être commode de superposer ces deux représentations.

## 1.9 AIDES À L'INTERPRÉTATION

Les axes factoriels fournissent des images approchées d'un nuage de points. Il est donc nécessaire de mesurer la qualité de l'approximation, tant pour chacun des points que pour l'ensemble du nuage. En outre, les plans factoriels représentent les coordonnées des points et non les inerties qui ont présidé à leur détermination. Il est souvent utile de consulter ces inerties. Il en résulte que l'étude d'un plan est toujours réalisée conjointement avec la consultation d'un ensemble d'indicateurs regroupés sous le terme d'aides à l'interprétation. Ce paragraphe définit les principales aides à l'interprétation : le chapitre 2 contient le traitement d'un exemple se référant largement à ces aides ; le chapitre 11 montre comment elles s'insèrent dans une démarche générale d'interprétation.

### 1.9.1 Définitions

*a) Qualité de représentation d'un élément par un axe*

La qualité de représentation de l'élément $i$ (individu ou variable) par l'axe $s$ est mesurée par le rapport :

$$\mathrm{QLT}_s(i) = \frac{[\text{inertie de la projection de l'élément } i \text{ sur l'axe } s]}{[\text{inertie totale de } i]}$$

C'est aussi le cosinus carré de l'angle $\theta$ entre $Oi$ et l'axe $s$ (*cf.* **Figure** 1.11).

$$\mathrm{QLT}_s(i) = \frac{(OH_i^s)^2}{(Oi)^2} = \cos^2\theta$$

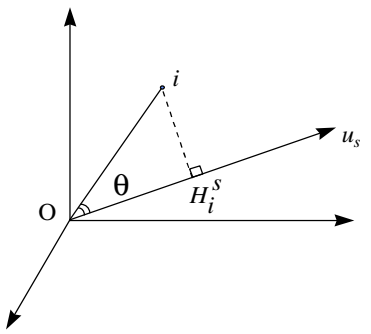



**Figure 1.11** Qualité de représentation d'un élément par un axe. $H_i^s$ : projection de $i$ sur l'axe de rang $s$

Cette définition se généralise au cas d'un plan. En outre, du fait de l'orthogonalité des axes factoriels, la qualité de représentation de l'élément $i$ par le plan (axe $s$, axe $t$) est la somme des qualités de représentation de $i$ par l'axe $s$ et par l'axe $t$. C'est aussi le cosinus carré de l'angle entre le vecteur $Oi$ et le plan de projection. Si la qualité de représentation d'un point sur un axe ou un plan est proche de 1, ce point est très proche de l'axe ou du plan. S'il s'agit d'un individu, sa distance au centre de gravité (qui est le point moyen) est alors bien visible sur la projection. Elle ne l'est pas dans le cas contraire (lorsque sa qualité de représentation est proche de 0). De même, la distance entre deux points sur un plan ne traduit bien leur distance dans le nuage que si ces deux points sont bien représentés. S'il s'agit d'une variable centrée-réduite, le vecteur a pour norme 1 et sa qualité de représentation est le carré de la longueur de sa projection. Sur un plan, elle s'apprécie directement par sa proximité au cercle de rayon 1, trace de l'hypersphère de rayon 1 sur le plan factoriel. Ce cercle est appelé couramment cercle des corrélations.

### *b) Qualité de représentation d'un nuage par un axe*

La définition précédente se généralise à l'ensemble d'un nuage par le rapport :

$$\frac{\text{inertie de la projection du nuage sur l'axe}}{\text{inertie totale du nuage}}$$

Cet indicateur, appelé pourcentage d'inertie associé à un axe, mesure en outre « l'importance » relative d'un axe factoriel dans la variabilité des données.

Comme dans le cas d'un seul élément, ces pourcentages peuvent être cumulés sur plusieurs axes ; on parle alors du pourcentage d'inertie extrait par un plan ou par les $S$ premiers facteurs. Du fait de la dualité (*cf.* section 1.7), il est équivalent de calculer ces pourcentages d'inertie à partir du nuage des individus ou de celui des variables.

### *c) Contribution d'un élément à l'inertie d'un axe*

Un axe factoriel rend maximum (sous contrainte d'orthogonalité avec les axes précédents) l'inertie projetée d'un nuage. Cette inertie projetée du nuage peut être décomposée point par point. Le quotient de l'inertie de la projection de l'élément $i$ (de poids $p_i$) sur l'axe $s$ [soit $p_i(OH_i^s)^2$] par l'inertie de la projection de l'ensemble du nuage sur l'axe $s$ (soit $\lambda_s$) représente la contribution de l'élément $i$ à l'inertie de l'axe $s$. Soit, en notant $\mathrm{CTR}_s(i)$ la contibution de l'élément $i$ à l'axe de rang $s$ :

$$CTR_s(i) = \frac{p_i \left(OH_i^s\right)^2}{\lambda_s}$$

Cet indicateur se généralise à un sous-ensemble d'éléments. La contribution d'un ensemble de points à l'inertie d'un axe est la somme des contributions des points

qui le composent. Ce rapport est précieux pour mettre en évidence le sous-ensemble d'éléments qui ont contribué principalement à la construction de l'axe et sur lequel s'appuiera en premier lieu l'interprétation.

### 1.9.2 Exemple numérique

Nous présentons ici, sur un exemple artificiel, la façon dont les coordonnées des points et les aides à l'interprétation interviennent dans l'analyse d'un facteur. Sept points du plan, munis de poids, sont représentés dans leurs axes principaux (*cf.* **Figure 1.12**).

| | Pour $F_1$ | |
|---|---|---|
| | $QLT_1$ | $CTR_1$ |
| A | 0 | 0 |
| B | 0 | 0 |
| C | 1 | .026 |
| D | 1 | .263 |
| E | 1 | .047 |
| F | .692 | .332 |
| G | .692 | .332 |

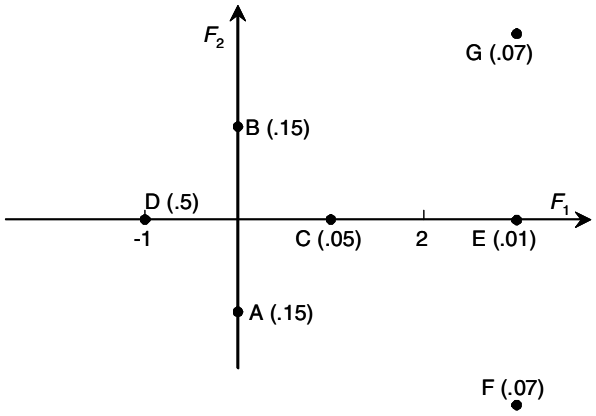



**Figure 1.12** Nuage plan pondéré représenté dans ses axes principaux. Les poids figurent entre parenthèses. $QLT_1$, $CTR_1$ : qualité de représentation et contribution (pour le premier axe).

#### *a) Coordonnées sur $F_1$*

Les points A, B, et C sont moyens ; D, E, F et G sont extrêmes, D étant opposé à E, F et G. Quelle que soit la qualité de représentation de ces points et leur contribution à l'inertie, cette structure traduite par le premier facteur n'est pas à mettre en doute.

#### *b) Qualité de représentation sur $F_1$*

$$\text{Exemple : } QLT_1(G) = \frac{\text{inertie projetée de G}}{\text{inertie totale de G}} = \frac{3^2}{(3^2 + 2^2)} = .692$$

Les points D, C et E, situés sur l'axe, ont une qualité de représentation égale à 1. Leurs distances dans le plan (à l'origine et entre eux) sont complètement traduites dans leur projection sur $F_1$. Les points D et E, à la fois extrêmes et bien représentés, sont caractéristiques de l'axe : l'examen de leurs différences avec la moyenne et entre eux permet de préciser l'opposition traduite par $F_1$. Réciproquement, toute valeur de E et de D qui s'écarte de la moyenne s'interprète par $F_1$.

Les points A et B, situés dans une direction orthogonale à l'axe 1, ont une qualité de représentation sur le premier axe égale à 0 : ni leur écart par rapport à l'origine, ni leur distance dans le plan ne sont visibles sur le premier facteur.

Les points F et G, extrêmes, ont une qualité de représentation moyenne : bien que très marqué pour le facteur $F_1$, leur écart à la moyenne n'est qu'en partie traduit par lui.

➤ Contribution à l'inertie de $F_1$

Exemple : inertie du nuage ($\lambda_1$) : $.5(-1)^2 + .05(1)^2 + .01(3)^2 + .07(3)^2 + .07(3)^2 = 1.9$

$\text{CTR}_1(\text{F})$ = inertie du point F/ inertie du nuage = $(.07 \times 3^2)/1.9 = .332$

Les points A et B ont une coordonnée nulle, donc une contribution nulle. Le point C est proche de O et a un petit poids : sa contribution est extrêmement faible. La suppression de ces trois points ne modifierait pas la direction du premier facteur.

Les points E et F ont la même coordonnée mais E, ayant un poids 7 fois plus faible que F, a une contribution 7 fois plus faible. La suppression de E risque moins de modifier le facteur que celle de F, pourtant moins bien représenté.

Le point D, malgré son poids égal à plus de 7 fois celui de F, a une contribution plus faible car il est situé plus près de l'origine (dans la contribution à l'inertie, la distance intervient par son carré alors que le poids intervient tel quel).

## 1.10 VARIABLES QUALITATIVES ILLUSTRATIVES EN ACP

On est souvent conduit à vouloir relier les résultats d'une ACP à des variables qualitatives définis sur les individus.

**Exemple :** On étudie les notes obtenues à différentes épreuves par un ensemble d'élèves. L'ACP de ce tableau met en évidence les principales dimensions de variabilité des élèves, par exemple une opposition entre les élèves plutôt meilleurs dans les matières scientifiques et ceux plutôt meilleurs dans les matières littéraires. On dispose par ailleurs d'informations sur ces élèves sous forme de variables qualitatives, par exemple leur genre (fille/garçon), la catégorie socio-professionnelle des parents, etc. Il est utile de relier ces variables qualitatives aux axes factoriels, avec en perspective des questions du type : observe-t-on, sur ces données, l'idée souvent émise selon laquelle les filles obtiennent des résultats plutôt meilleurs dans les matières littéraires et les garçons des résultats plutôt meilleurs dans les matières scientifiques ?

Pour cela, on dispose de deux outils graphiques simples et efficaces :

– identification, sur les plans factoriels, des individus par leur modalité à l'aide d'un code, de couleur ou de forme (dans l'exemple on pourra identifier les filles par un point rose et les garçons par un point bleu !) ; cela permet d'étudier finement

la relation entre une variable qualitative et le plan factoriel mais nécessite un graphique par variable ;

– représentation, sur les plans factoriels, des centres de gravité des ensembles d'individus possédant une même modalité (dans l'exemple, le centre de gravité des filles et celui des garçons) ; à la différence de la technique précédente, un seul graphique permet d'examiner plusieurs variables qualitatives simultanément, mais, en revanche, ne donne pas d'informations quant à la variabilité des individus présentant une même modalité.

On peut chercher à traduire la variabilité des individus autour des centres de gravité des variables qualitatives en terme de variabilité des centres de gravité eux-mêmes. Pour cela, on construit, autour de chaque centre de gravité, une ellipse de confiance, analogue bi-dimensionnel de l'intervalle de confiance que l'on calcule usuellement autour d'une moyenne. Pour produire ces ellipses, on procède de la façon suivante :

1. On considère l'ensemble $I$ des $I$ individus observés, comme un échantillon d'une population plus vaste, dite de référence ; dans cette perspective, la variabilité des individus se traduit par une variabilité des centres de gravité induite par le fait que l'ensemble $I$ observé n'est que l'un des ensembles possibles de $I$ individus parmi la population de référence.
2. La variabilité des centres de gravité pourrait être obtenue en extrayant d'autres échantillons de la population de référence mais cela est généralement impossible ; aussi approxime-t-on la population de référence par l'ensemble $I$ et l'on tire, au hasard avec remise, plusieurs échantillons de $I$ individus dans cet ensemble ; cette procédure est appelée « bootstrap ».
3. Pour chaque échantillon « bootstrap », on calcule les centres de gravité des différentes modalités et l'on projette ces centres de gravité (dits bootstrap) en supplémentaire sur les plans de l'ACP (initiale) des $I$.
4. Si l'on effectue $n$ tirages bootstrap, on obtient, pour une modalité donnée, $n$ points ; on pourrait se contenter de représenter ces $n$ points mais les graphiques obtenus sont peu lisibles dès lors que le nombre de modalités étudiées est un tant soit peu grand ; pour simplifier les représentations, on construit l'ellipse centrée sur le centre de gravité initial et contenant 95 % des $n$ centres de gravité bootstrap ; ces ellipses sont dites « ellipses de confiance bootstrap ». L'expérience montre que schématiser la distribution des $n$ points par une ellipse n'est pas gênant (au sens ou, en pratique, l'observation du nuage des $n$ points ne conduit pas à des interprétations plus riches) dès lors que l'effectif par modalité est assez grand (disons une vingtaine d'individus pour fixer les idées).

➤ Remarque

La taille d'une ellipse ainsi obtenue dépend de la variabilité (dans le plan factoriel) des individus présentant la modalité étudiée mais aussi de son effectif.

L'utilisation pratique des ellipses de confiance s'articule autour de deux questions relatives aux modalités.

**La modalité *m* est-elle caractérisée par le plan factoriel** ? Autrement dit, les individus possédant la modalité $m$ occupent-ils (dans l'ensemble) une position excentrée sur le plan ? Pour cela, on examine la position de l'origine des axes, centre de gravité de l'ensemble $I$, par rapport à l'ellipse de confiance de $m$. Si cette ellipse englobe l'origine, on décidera que la modalité $m$ ($i.e.$ les individus possédant cette modalité) n'est pas caractérisée par le plan.

**Les deux modalités *m* et $m'$ sont-elles différenciées par le plan** ? Autrement dit, les individus possédant la modalité $m$ occupent-ils, dans l'ensemble, la même région du plan que ceux possédant la modalité $m'$ ? Pour cela, on examine le recouvrement entre les deux ellipses associées aux modalités $m$ et $m'$. Une absence de recouvrement conduit à décider que le plan différencie les deux modalités et, à l'inverse, un fort recouvrement conduit à décider d'une non différenciation. Un recouvrement faible laisse la place au doute : pour aider sa décision, l'utilisateur peut calculer la probabilité critique du test statistique « $T^2$ de Hotelling » appliqué à la comparaison des deux modalités du point de vue des deux composantes principales étudiées considérées simultanément.

Les questions concernant la position des modalités sur un plan peuvent être posées pour chaque axe. Pour cela, en projetant les ellipses sur chaque axe, on obtient un intervalle de confiance que l'on peut utiliser comme un intervalle de confiance usuel. Il existe aussi un indicateur, appelé valeur-test et introduit initialement dans le logiciel SPAD, qui permet de juger, pour un axe factoriel (et, plus généralement pour n'importe quelle variable), de l'écart entre le centre de gravité d'une classe et le centre de gravité général (*cf.* section 2.4.4 page 54).

Chapitre **2**

# Exemple d'ACP et de CAH

Le commentaire de l'ACP d'un petit tableau permet d'illustrer les règles et la démarche d'interprétation d'une ACP (voir aussi chapitre 11). Nous en présentons un ci-après.

En pratique, le dépouillement des résultats d'une analyse factorielle s'accompagne généralement de celui des résultats d'une classification ascendante hiérarchique (CAH) réalisée sur les mêmes données. L'objet de ce livre, dédié aux analyses factorielles, exclut une présentation générale des méthodes de classification. En revanche, il a paru utile d'accorder quelque place à l'énoncé des principes régissant la méthode de classification ascendante hiérarchique la plus utilisée simultanément aux analyses factorielles (la méthode de Ward) et à l'illustration du dépouillement conjoint des résultats des deux méthodes.

Les données utilisées pour illustrer l'ACP serviront à introduire ces éléments de classification.

## 2.1 DONNÉES ET PROBLÉMATIQUE

### 2.1.1 Description des données

Pour 15 villes de France, on dispose des moyennes des températures mensuelles calculées sur 30 ans (entre 1931 et 1960). Ces données sont extraites du *Quid 1986*, page 507 (Éditions Robert Laffont).

Elles sont rassemblées dans le **tableau 2.1**, qui croise les 15 villes (en lignes) et les 12 mois de l'année (en colonnes). Les quatre colonnes supplémentaires sont commentées par la suite. Les deux dernières lignes, la moyenne et l'écart-type des colonnes, ne sont là que pour information ; elles ne sont pas introduites dans l'analyse.

| | janv | févr | mars | avri | mai | juin | juil | aoû | sept | octo | nove | déce | lati | longi | moy | ampli |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bordeaux | 5.6 | 6.6 | 10.3 | 12.8 | 15.8 | 19.3 | 20.9 | 21.0 | 18.6 | 13.8 | 9.1 | 6.2 | 44.50 | -0.34 | 13.33 | 15.4 |
| Brest | 6.1 | 5.8 | 7.8 | 9.2 | 11.6 | 14.4 | 15.6 | 16.0 | 14.7 | 12.0 | 9.0 | 7.0 | 48.24 | -4.29 | 10.77 | 10.2 |
| Clermont | 2.6 | 3.7 | 7.5 | 10.3 | 13.8 | 17.3 | 19.4 | 19.1 | 16.2 | 11.2 | 6.6 | 3.6 | 45.47 | 3.05 | 10.94 | 16.8 |
| Grenoble | 1.5 | 3.2 | 7.7 | 10.6 | 14.5 | 17.8 | 20.1 | 19.5 | 16.7 | 11.4 | 6.5 | 2.3 | 45.10 | 5.43 | 10.98 | 18.6 |
| Lille | 2.4 | 2.9 | 6.0 | 8.9 | 12.4 | 15.3 | 17.1 | 17.1 | 14.7 | 10.4 | 6.1 | 3.5 | 50.38 | 3.04 | 9.73 | 14.7 |
| Lyon | 2.1 | 3.3 | 7.7 | 10.9 | 14.9 | 18.5 | 20.7 | 20.1 | 16.9 | 11.4 | 6.7 | 3.1 | 45.45 | 4.51 | 11.36 | 18.6 |
| Marseille | 5.5 | 6.6 | 10.0 | 13.0 | 16.8 | 20.8 | 23.3 | 22.8 | 19.9 | 15.0 | 10.2 | 6.9 | 43.18 | 5.24 | 14.23 | 17.8 |
| Montpellier | 5.6 | 6.7 | 9.9 | 12.8 | 16.2 | 20.1 | 22.7 | 22.3 | 19.3 | 14.6 | 10.0 | 6.5 | 43.36 | 3.53 | 13.89 | 17.1 |
| Nantes | 5.0 | 5.3 | 8.4 | 10.8 | 13.9 | 17.2 | 18.8 | 18.6 | 16.4 | 12.2 | 8.2 | 5.5 | 47.13 | -1.33 | 11.69 | 13.8 |
| Nice | 7.5 | 8.5 | 10.8 | 13.3 | 16.7 | 20.1 | 22.7 | 22.5 | 20.3 | 16.0 | 11.5 | 8.2 | 43.42 | 7.15 | 14.84 | 15.2 |
| Paris | 3.4 | 4.1 | 7.6 | 10.7 | 14.3 | 17.5 | 19.1 | 18.7 | 16.0 | 11.4 | 7.1 | 4.3 | 48.52 | 2.20 | 11.18 | 15.7 |
| Rennes | 4.8 | 5.3 | 7.9 | 10.1 | 13.1 | 16.2 | 17.9 | 17.8 | 15.7 | 11.6 | 7.8 | 5.4 | 48.05 | -1.41 | 11.13 | 13.1 |
| Strasbourg | .4 | 1.5 | 5.6 | 9.8 | 14.0 | 17.2 | 19.0 | 18.3 | 15.1 | 9.5 | 4.9 | 1.3 | 48.35 | 7.45 | 9.72 | 18.6 |
| Toulouse | 4.7 | 5.6 | 9.2 | 11.6 | 14.9 | 18.7 | 20.9 | 20.9 | 18.3 | 13.3 | 8.6 | 5.5 | 43.36 | 1.26 | 12.68 | 16.2 |
| Vichy | 2.4 | 3.4 | 7.1 | 9.9 | 13.6 | 17.1 | 19.3 | 18.8 | 16.0 | 11.0 | 6.6 | 3.4 | 46.08 | 3.26 | 10.72 | 16.9 |
| Moyenne | 4.0 | 4.8 | 8.2 | 11.0 | 14.4 | 17.8 | 19.8 | 19.6 | 17.0 | 12.3 | 7.9 | 4.9 | 46.0 | 2.58 | 11.8 | 15.9 |
| Ecart-type | 1.94 | 1.81 | 1.48 | 1.37 | 1.45 | 1.73 | 2.06 | 1.94 | 1.79 | 1.77 | 1.74 | 1.89 | 2.22 | 3.21 | 1.55 | 2.25 |

**Tableau 2.1** Températures moyennes mensuelles de 15 villes de France. La latitude et la longitude (négative à l'ouest du méridien de Greenwich) sont exprimées en degrés. Moy : moyenne des 12 moyennes mensuelles. Ampli : amplitude thermique (moyenne mensuelle maximum-moyenne mensuelle minimum)

### 2.1.2 Problématique

Le but général de l'étude est de comparer les températures mensuelles des différentes villes. Précisons quelques questions auxquelles les résultats de l'ACP permettent de répondre en abordant le tableau successivement à travers ses lignes et à travers ses colonnes.

#### *a) Point de vue des lignes (ou individus : les villes)*

Chaque ville est caractérisée par ses 12 températures moyennes mensuelles. Quelles sont, de ce point de vue, les villes qui se ressemblent ? Quelles sont celles qui diffèrent ? Plus généralement, peut-on faire une typologie des villes mettant en évidence l'ensemble des ressemblances ainsi définies ? En ACP, la dissemblance entre les individus est mesurée par une distance (cf. section 1.1 page 7). Ici, le carré de la distance entre deux villes est la somme des carrés des douze différences entre leurs températures moyennes mensuelles. Cela traduit bien la notion souhaitée de proximité. Cette typologie faite, on peut se demander si ces ressemblances (ou dissemblances) correspondent à des proximités (ou des éloignements) géographiques.

L'étude des individus revient donc à analyser leur variabilité. Un point de vue voisin de celui de typologie consiste à mettre en évidence les principales dimensions de cette variabilité.

### *b) Point de vue des colonnes (ou variables : les mois)*

Chaque mois est vu au travers des températures moyennes mensuelles des 15 villes. Le problème n'est pas de séparer les mois chauds des mois froids pour l'ensemble des 15 villes (ce qui arriverait si nous les considérions comme des individus) mais de comparer la répartition des 15 villes (des plus chaudes aux plus froides) pour deux mois différents *sans tenir compte du fait que d'un mois à l'autre les températures sont globalement plus ou moins élevées* (l'élimination de cet effet de moyenne est assurée par le centrage). Les comparaisons entre mois se font au travers de la notion de liaison, plus précisément de corrélation, entre variables numériques. Deux mois sont d'autant plus corrélés que, pour chacun, on observe la même répartition des 15 villes selon leur température. À l'inverse, ils sont peu corrélés si ce ne sont pas dans les mêmes villes que l'on trouve les températures les plus élevées (ou les plus basses).

Cela posé, les questions sont les suivantes : quels mois sont corrélés entre eux ? Quels sont ceux qui le sont peu ? Plus généralement, peut-on faire un bilan des corrélations entre les 12 mois ? Les températures mensuelles sont-elles liées à la position géographique ? D'autre part, si les mois sont corrélés, l'information donnée par les 12 colonnes est, en un certain sens, redondante. Peut-on la résumer en remplaçant les 12 mois par un petit nombre de variables synthétiques ?

### *c) Ajout de variables supplémentaires (ou illustratives)*

Il apparaît dans la problématique que les températures doivent être analysées en ayant à l'esprit la position géographique des villes. On peut formaliser cette position par la latitude et la longitude, données introduites dans l'analyse en tant que variables supplémentaires. Deux autres variables supplémentaires ont été ajoutées pour des raisons qui apparaissent au cours de l'interprétation.

### *d) Faut-il réduire les données ?*

Lorsque les unités de mesure diffèrent d'une variable à l'autre, le recours à la réduction des variables est systématique (cf. § 1.2 page 10). Ce n'est pas le cas ici et la question mérite d'être posée.

Ne pas réduire revient ici à considérer qu'un écart de 1 degré entre deux villes a la même importance quel que soit le mois au cours duquel il est observé, que ce soit un mois où les écarts entre les températures des 15 villes sont plutôt faibles ou au contraire importants. Selon ce point de vue, dans les distances entre les villes, un mois possède alors d'autant plus d'influence que l'on y observe de grandes différences de températures d'une ville à l'autre (ne pas réduire les variables revient à accorder aux variables réduites un poids égal à leur variance). À l'inverse, en réduisant, on accorde à chaque mois de l'année la même importance *a priori* dans l'analyse.

Sur ce jeu de données, les deux points de vue sont également défendables. Pour cet exemple didactique, nous choisissons de réduire les données ; l'ACP est alors dite normée. Comme les écarts-types varient peu d'un mois à l'autre (minimum : 1.37 et maximum : 2.06), les deux analyses, normée et non normée, conduisent nécessairement à des résultats très proches. Ceci a été vérifié : pour les quatre premiers facteurs, les coefficients de corrélation entre les facteurs de même rang des deux analyses sont tous supérieurs à 0.99.

*Remarque*

En pratique, la réduction est l'option par défaut dans les logiciels.

## 2.2 RÉSULTATS DE L'ACP

### 2.2.1 Indicateurs d'inertie

Dans une ACP normée, l'inertie totale de chacun des nuages (celui des villes et celui des mois) est égale au nombre de variables actives (ici 12). Avec une inertie de 9.58, qui représente 80 % de l'inertie des nuages dans l'espace tout entier, le premier facteur est largement prépondérant. L'inertie du deuxième facteur vaut 2.28 et celle du troisième 0.07 ; les deux premiers facteurs totalisent 98.8 % de l'inertie totale. Les deux nuages de points (individus et variables) sont donc pratiquement bidimensionnels : leur projection sur le premier plan factoriel en donne une représentation quasiment parfaite. On se limite dans l'interprétation à l'étude de ces deux premiers facteurs et du plan qu'ils engendrent.

*Contribution des individus (cf. Tableau 2.2)*

Le premier facteur est dû essentiellement à 5 villes (Lille, Marseille, Montpellier, Nice et Strasbourg) qui totalisent 77.4 % de son inertie. Compte tenu du faible nombre de villes étudiées, cette situation est banale et n'attire pas d'observation particulière.

Le deuxième facteur est dû pour moitié (49.1 %) à la ville de Brest, qui est donc assez particulière du point de vue climatique. Remarquons toutefois que la différence d'inertie entre le deuxième et le troisième facteur ($2.28 - 0.07 = 2.20$) est beaucoup plus grande que l'inertie de Brest le long du deuxième axe ($2.28 \times 0.49 = (-4.093)^2 \times (1/15) = 1.12$). Même sans la ville de Brest, ce deuxième facteur serait donc apparu. Il semble que le cas de Brest est, certes, particulier mais s'inscrit dans une tendance générale, ce qui sera confirmé lors de l'interprétation.

| | Coordonnée | | Contribution | | Qual. de représentation | | | d$(i, O)$ | Inertie |
|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F1 | F2 | F1 | F2 | F1,F2 | | |
| Bordeaux | 3.121 | −0.109 | 6.8 | 0 | .947 | .001 | .948 | 3.207 | 5.7 |
| Brest | −2.268 | −4.093 | 3.6 | 49.1 | .234 | .763 | .998 | 4.685 | 12.2 |
| Clermont | −1.726 | 0.593 | 2.1 | 1 | .88 | .104 | .984 | 1.840 | 1.9 |
| Grenoble | −1.529 | 1.688 | 1.6 | 8.3 | .429 | .523 | .952 | 2.335 | 3 |
| Lille | −4.217 | −0.595 | 12.4 | 1 | .972 | .019 | .991 | 4.278 | 10.2 |
| Lyon | −0.835 | 1.788 | 0.5 | 9.4 | .178 | .817 | .995 | 1.978 | 2.2 |
| Marseille | 4.833 | 0.829 | 16.2 | 2 | .964 | .028 | .993 | 4.922 | 13.5 |
| Montpellier | 4.147 | 0.435 | 12 | 0.6 | .986 | .011 | .997 | 4.177 | 9.7 |
| Nantes | −0.281 | −1.115 | 0.1 | 3.6 | .056 | .886 | .943 | 1.184 | 0.8 |
| Nice | 6.007 | −0.789 | 25.1 | 1.8 | .98 | .017 | .997 | 6.068 | 20.5 |
| Paris | −1.242 | 0.156 | 1.1 | 0.1 | .889 | .014 | .903 | 1.317 | 1 |
| Rennes | −1.439 | −1.671 | 1.4 | 8.2 | .42 | .567 | .986 | 2.220 | 2.7 |
| Strasbourg | −4.106 | 2.172 | 11.7 | 13.8 | .776 | .217 | .993 | 4.662 | 12.1 |
| Toulouse | 1.736 | 0.136 | 2.1 | 0.1 | .953 | .006 | .958 | 1.779 | 1.8 |
| Vichy | −2.201 | 0.575 | 3.4 | 1 | .922 | .063 | .984 | 2.293 | 2.9 |
| Ensemble | 0 | 0 | 100 | 100 | .7985 | .1897 | .9882 | | 100 |

**Tableau 2.2** Aides à l'interprétation des 15 villes pour les 2 premiers facteurs.

| | janv | févr | mars | avri | mai | juin | juil | août | sept | octo | nove | déce | lati | longi | moy | ampl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Facteur 1 | .76 | .88 | .97 | .97 | .87 | .86 | .84 | .90 | .97 | .98 | .90 | .77 | -.84 | .17 | 1.00 | .10 |
| Facteur 2 | -.64 | -.47 | -.16 | .20 | .47 | .50 | .53 | .43 | .21 | -.17 | -.41 | -.62 | -.31 | .79 | -.02 | .99 |

**Tableau 2.3** Coordonnées (=corrélations) des variables actives et supplémentaires pour chacun des 2 premiers facteurs.

### 2.2.2 Interprétation du premier facteur

*a) Coordonnées des variables actives (cf. Tableau 2.3 et Figure 2.1)*

Les 12 variables sont corrélées fortement et positivement au premier facteur. Etant ainsi liées à une même variable, elles sont liées entre elles ; ceci peut être constaté sur la partie haute de la matrice des corrélations (cf. **Tableau 2.4**) dont toutes les valeurs sont positives.

Ce type de facteur est classique et est appelé « effet taille » (cf. § 1.6 page 15). Il exprime que certains individus ont de grandes valeurs pour l'ensemble des variables et d'autres de petites valeurs pour l'ensemble des variables. Dans notre exemple, cela indique que certaines villes sont plus chaudes que d'autres tout au long de l'année.

*b) Coordonnées des individus (cf. Tableau 2.2 et Figure 2.2)*

Compte tenu des relations entre les coordonnées des individus et celles des variables (cf. relations de transition, section 1.7.3 page 20), on s'attend à trouver, le long de l'axe 1, les villes chaudes du côté des coordonnées positives et les villes froides du

| | janv | févr | mars | avri | mai | juin | juil | août | sept | octo | nove | déce | lati | longi | moy | ampl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Janvier** | 1 | | | | | | | | | | | | | | | |
| **Février** | .97 | 1 | | | | | | | | | | | | | | |
| **Mars** | .84 | .93 | 1 | | | | | | | | | | | | | |
| **Avril** | .61 | .76 | .92 | 1 | | | | | | | | | | | | |
| **Mai** | .36 | .55 | .77 | .95 | 1 | | | | | | | | | | | |
| **Juin** | .34 | .52 | .76 | .94 | .99 | 1 | | | | | | | | | | |
| **Juillet** | .30 | .49 | .72 | .91 | .98 | .99 | 1 | | | | | | | | | |
| **Août** | .41 | .59 | .80 | .95 | .98 | .99 | .99 | 1 | | | | | | | | |
| **Septembre** | .60 | .76 | .91 | .98 | .94 | .94 | .93 | .97 | 1 | | | | | | | |
| **Octobre** | .85 | .94 | .97 | .91 | .77 | .76 | .74 | .81 | .93 | 1 | | | | | | |
| **Novembre** | .95 | .99 | .93 | .78 | .59 | .57 | .55 | .64 | .80 | .96 | 1 | | | | | |
| **Décembre** | .99 | .97 | .83 | .62 | .38 | .36 | .32 | .43 | .62 | .87 | .96 | 1 | | | | |
| **Latitude** | -.42 | -.60 | -.81 | -.85 | -.84 | -.87 | -.88 | -.90 | -.90 | -.78 | -.64 | -.44 | 1 | | | |
| **Longitude** | -.39 | -.22 | -.04 | .29 | .54 | .53 | .59 | .50 | .35 | .07 | -.13 | -.35 | -.31 | 1 | | |
| **Moyenne** | .77 | .89 | .97 | .96 | .86 | .85 | .83 | .89 | .97 | .98 | .91 | .79 | -.83 | .16 | 1 | |
| **Amplitude** | -.57 | -.38 | -.06 | .28 | .55 | .58 | .62 | .52 | .31 | -.06 | -.30 | -.54 | -.42 | .83 | .08 | 1 |

**Tableau 2.4** Matrice des corrélations entre toutes les variables.

côté des coordonnées négatives. C'est bien ce que l'on observe, l'axe 1 opposant principalement Nice, Marseille et Montpellier (à droite) à Lille et Strasbourg (à gauche). Cette opposition se retrouve facilement dans les données. Ainsi, quel que soit le mois de l'année, les températures mesurées à Nice, Marseille et Montpellier se situent au-dessus de la moyenne (calculée sur les 15 villes) tandis que celles mesurées à Lille et Strasbourg se situent au-dessous de cette moyenne. Attention, la première formule de transition relie la coordonnée d'une ville à *l'ensemble des coordonnées* des variables. Ainsi, Lille a la plus faible coordonnée sur le premier axe, mais il serait faux d'en conclure qu'elle est, quel que soit le mois, la ville la plus froide. La fausseté de cette affirmation se constate immédiatement sur les données : bien que toujours plus froide que la moyenne, Lille n'est la ville la plus froide que deux mois sur douze (septembre et avril).

La position extrême de Lille provient du fait que cette ville est la plus froide sur *l'ensemble de l'année*. Certains mois de l'année, une autre ville, ou même plusieurs, sont plus froides qu'elle mais elles sont sensiblement moins froides que Lille pendant beaucoup d'autres mois. La position des villes proches de l'origine s'interprète dans le même esprit. La faible coordonnée, sur le premier axe, de Nantes, Lyon ou Paris indique que, sur l'ensemble de l'année, la température de ces villes est moyenne. Mais on ne peut en déduire que les températures y sont toujours moyennes car elles peuvent aussi être tantôt élevées et tantôt basses. Le deuxième facteur est éclairant sur ce point.

*c) Coordonnées des variables supplémentaires (cf. Tableau 2.3)*

Ce premier facteur semble correspondre à la température moyenne annuelle. Pour s'en assurer, on peut faire la moyenne des 12 températures mensuelles pour chacune des 15 villes et calculer le coefficient de corrélation entre cette nouvelle variable et le premier facteur (défini sur les villes). En pratique, il suffit de relancer la même analyse, la température moyenne annuelle étant introduite en variable supplémentaire. Ce coefficient de corrélation vaut 1.00 (aux erreurs d'arrondi près), ce qui achève de justifier l'interprétation du premier facteur comme étant la température moyenne annuelle. Remarquons que, bien que le coefficient de corrélation soit très proche de 1, ce premier facteur n'est pas exactement la moyenne annuelle. Comme toute composante principale, ce facteur est une combinaison linéaire des variables actives dont les coefficients sont proportionnels aux coordonnées des variables (cf. Fig. 1.9 page 19). Si ce facteur coïncidait exactement avec la moyenne, les 12 coefficients de la combinaison linéaire seraient égaux. Or cette combinaison est proportionnelle ici à : 0.76 janvier + 0.88 février + . . . + 0.77 décembre.

Considérer ce premier facteur comme une moyenne annuelle est une interprétation interne aux données traitées. On franchit un nouveau pas dans l'interprétation en le reliant à des données externes comme la position géographique des villes. Le nombre de villes étant faible, on peut constater directement que, parmi les 15 villes, les plus chaudes sont aussi les plus méridionales. La latitude et la longitude ayant été introduites dans l'analyse en tant que variables supplémentaires, on dispose de leur coefficient de corrélation avec le premier facteur. Celui de la latitude vaut 0.84, ce qui exprime que la répartition des 15 villes sur le premier axe correspond à peu près à leur latitude (à peu près seulement : des villes comme Vichy, Clermont, Grenoble et Lyon sont plus froides que ne le laisse attendre leur latitude). La longitude, elle, est très peu liée au premier facteur (corrélation 0.17).

### 2.2.3 Interprétation du deuxième facteur

*a) Coordonnées des variables actives (cf. Tableau 2.3 et Figure 2.1)*

Les mois d'automne et d'hiver sont opposés aux mois de printemps et d'été. Les mois qui encadrent les solstices d'hiver et d'été sont les plus corrélés à ce facteur. Cette opposition montre que, à température moyenne annuelle égale (*i.e.* à premier facteur fixé), certaines villes sont plutôt chaudes en été et plutôt froides en hiver alors que d'autres, à l'inverse, sont plutôt froides en été et plutôt chaudes en hiver. L'amplitude thermique, plus importante pour les premières que pour les secondes, semble correspondre à ce facteur.

*b) Coordonnées des individus (cf. Tableau 2.2 et Figure 2.2)*

Compte tenu des relations de transition, on sait que les coordonnées des villes ayant une forte amplitude thermique sont positives tandis que celles des villes à faible amplitude sont négatives. Ainsi, Brest, dont la coordonnée sur ce facteur est la plus élevée, présente des températures au-dessus de la moyenne depuis novembre jusqu'à février et très au-dessous de la moyenne depuis avril jusqu'à septembre. Cette tendance se retrouve de façon atténuée pour la belle ville de Rennes. À l'opposé, Grenoble subit des températures très en dessous de la moyenne depuis novembre jusqu'à février et presque égales à la moyenne depuis mai jusqu'à août. Brest apparaît donc comme la situation la plus extrême d'une tendance générale.

*c) Coordonnées des variables supplémentaires (cf. Tableau 2.3)*

L'interprétation générale du deuxième facteur est confirmée par sa corrélation avec la variable supplémentaire *amplitude thermique* (température mensuelle maximum – température mensuelle minimum) égale à 0.99. Avec un coefficient de corrélation de 0.79, ce facteur est lié aussi à la longitude (qui, grossièrement, exprime la proximité avec l'océan Atlantique et, encore plus grossièrement, la continentalité). Sur ce deuxième facteur, les villes sont à peu près placées par longitude croissante ; la seule exception notable est Nice qui, très à l'est, a pourtant une amplitude thermique annuelle légèrement inférieure à la moyenne.

### 2.2.4 Premier plan factoriel

Il est toujours intéressant d'étudier globalement un plan factoriel, même si, comme ici, chaque facteur est clairement interprétable.

*a) Remarques sur la représentation des variables (cf. Figure 2.1)*

La projection sur le premier plan factoriel conservant 98.8 % de l'inertie du nuage des mois (construit dans un espace de dimension 15), la déformation des longueurs et des angles des vecteurs représentant ces 12 variables est presque négligeable. Les extrémités des flèches associées aux 12 mois n'atteignent pas le cercle de rayon 1 (appelé cercle de corrélation) mais il s'en faut de très peu. On peut vérifier sur ce plan la représentation géométrique du coefficient de corrélation par le cosinus de l'angle entre les vecteurs représentant les variables. Par exemple, la corrélation entre *janvier* et *juillet* vaut 0.30, ce qui correspond à un angle de 72 degrés, angle que l'on peut mesurer sur le plan.

Insistons sur le fait que cette propriété, toujours vraie dans l'espace complet, ne se vérifie sur les plans factoriels que pour les variables parfaitement bien représentées. Ainsi, l'angle observé dans le plan entre *juillet* et la variable supplémentaire *longitude* vaut 45 degrés, angle dont le cosinus vaut 0.70. Mais la longitude n'est pas très bien

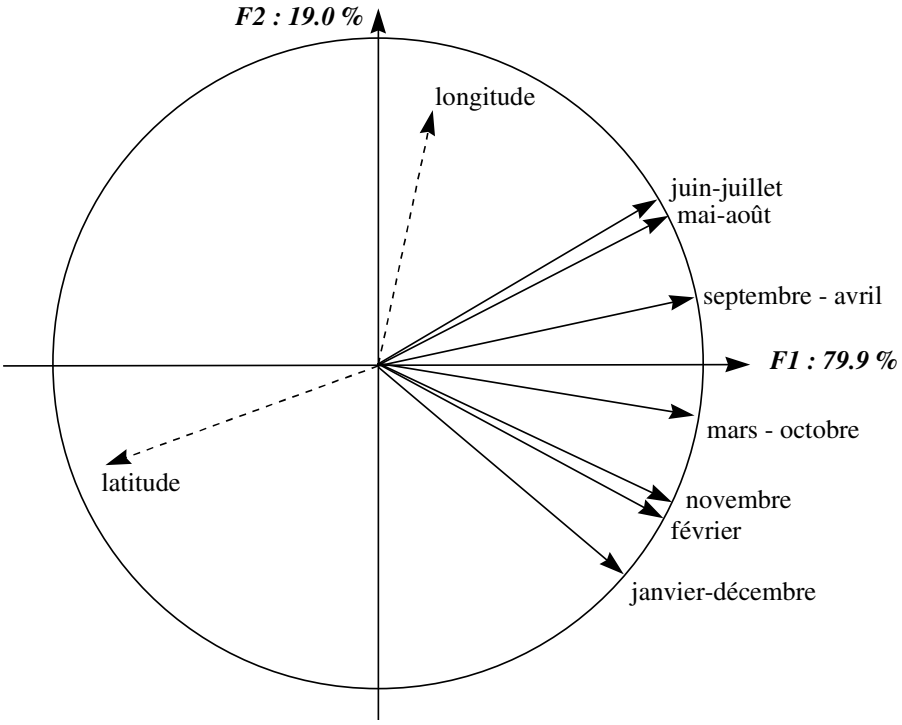



**Figure 2.1** Projection des 12 variables actives et de 2 variables supplémentaires sur le plan des deux premiers facteurs. Deux variables très proches ne sont représentées que par un seul vecteur.

représentée sur ce plan, comme sa distance au cercle de corrélation permet de le constater. Il n'est donc pas étonnant que la corrélation entre *juillet* et *longitude* (0.59) soit inférieure à 0.70 (la projection ne peut que diminuer les angles).

On prendra garde à l'interprétation de la forte corrélation entre 2 mois consécutifs. Dans le calcul du coefficient de corrélation, les variables sont centrées : aussi, le fait que deux mois consécutifs aient des températures moyennes proches n'intervient pas directement dans la forte corrélation. Celle-ci découle du fait que, pour ces deux mois, ce sont les mêmes villes qui sont les plus chaudes et les mêmes villes qui sont les plus froides (plus précisément les différences de températures entre villes sont proportionnelles d'un mois à l'autre).

### *b) Bilan des liaisons entre variables (cf. Figure 2.1)*

Tous les angles entre les vecteurs représentant les variables étant inférieurs à un angle droit, les douze températures mensuelles sont corrélées positivement entre elles. En plus, il apparaît une structure qui correspond au cycle annuel avec deux périodes. De

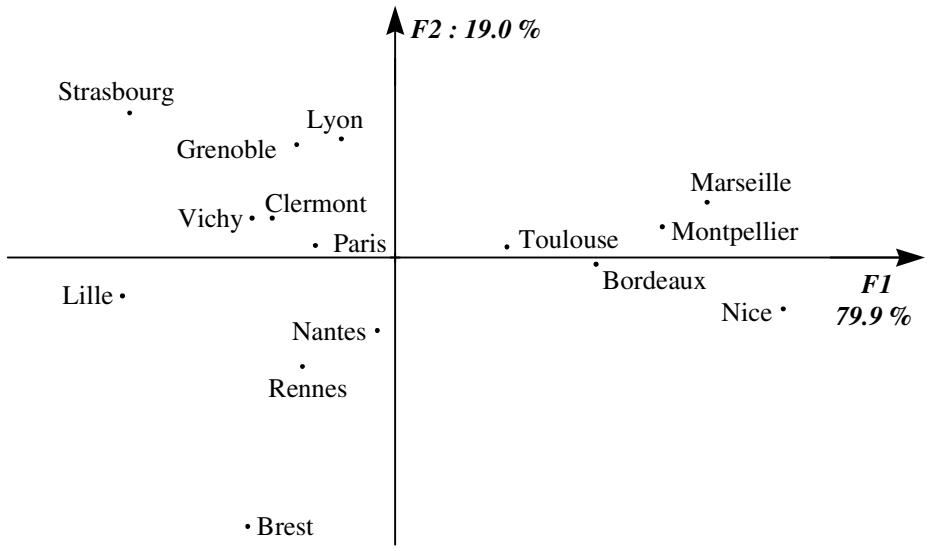



**Figure 2.2** Projection des 15 villes sur le premier plan factoriel.

janvier à juin d'une part et de juillet (très proche de juin) à décembre (très proche de janvier) d'autre part, les mois se répartissent dans l'ordre du calendrier : deux mois proches dans le calendrier sont fortement corrélés entre eux (la corrélation entre deux mois consécutifs n'est jamais inférieure à 0.92) et dans chacune des deux périodes, cette liaison décroît régulièrement avec l'éloignement. D'autre part, les mois des deux périodes se superposent quasiment deux à deux. Finalement, on constate que deux mois sont d'autant plus corrélés qu'ils correspondent à la même durée du jour.

### *c) Variables synthétiques*

Il est clairement apparu que l'évolution thermique annuelle de l'ensemble des 15 villes peut être presque parfaitement synthétisée par deux variables : la température moyenne annuelle et l'amplitude thermique.

### *d) Typologies des villes (cf. Figure 2.2)*

Sur ce plan, les deux axes correspondent aux deux variables synthétiques. Ainsi, plus une ville est froide, plus elle est située à gauche sur le plan ; plus son amplitude thermique est grande, plus elle est située en haut.

Remarquons que les villes « chaudes », situées à droite, sont proches de l'axe horizontal : le deuxième facteur ne les différencie guère. Au contraire, pour les villes « froides », les différences d'amplitude thermique sont importantes.

La répartition sur le plan permet, un peu arbitrairement, de distinguer trois groupes de villes. L'interprétation des deux axes permet de caractériser ces groupes.

1. Les villes à climat chaud : Marseille, Montpellier, Nice, Bordeaux et Toulouse.
2. Les villes à climat froid et continental (été chaud, hiver très froid) : Lille, Strasbourg, Lyon, Grenoble, Vichy, Clermont et Paris.
3. Les villes à climat froid et océanique (été froid, hiver doux) : Brest, Rennes et Nantes.

### *e) Remarques sur la qualité de représentation des villes (cf. Tableau 2.2)*

La qualité de représentation d'un individu (par un axe, un plan ou un sous-espace) est une expression raccourcie de « qualité de représentation de l'écart entre un individu et le point moyen » (par un axe, un plan ou un sous-espace). À la différence de celle des variables (dont la distance à l'origine est constante), la qualité de représentation des individus ne se lit pas directement sur le graphique. Il faut consulter le tableau 2.2 les indiquant.

Toutes les villes sont très bien représentées sur ce plan (ce qui n'est pas étonnant puisque la qualité de représentation de l'ensemble du nuage est de 98.8). La moins bien représentée est Paris avec 0.889+0.014=0.903. La différence entre les températures mensuelles de Paris et les températures mensuelles moyennes des 15 villes n'est pas totalement expliquée sur ce plan ; pour cela il faudrait consulter les facteurs suivants, le quatrième plus que le troisième puisque la qualité de représentation sur ces axes est respectivement de 0.03 et 0.07.

La coordonnée d'un individu est toujours interprétable, même si sa qualité de représentation par cet axe est mauvaise. Ainsi, bien que Paris soit mal représentée par le deuxième axe, sa coordonnée presque nulle indique bien une amplitude thermique moyenne (vérifiable sur les données).

### *f) Autres aides à l'interprétation des individus*

La distance ($d(i, O)$ dans le **tableau 2.2**) calculée dans l'espace complet (ici à 12 dimensions) entre un individu $i$ et le point moyen indique dans quelle mesure l'individu $i$ est extrême – ou particulier – du point de vue de l'ensemble de ses coordonnées. Ici, on remarque que les villes les plus extrêmes du point de vue de l'ensemble de leurs températures mensuelles sont Nice, Marseille, Brest et Strasbourg. Cela n'étonne pas puisque ces villes sont géographiquement les plus excentrées.

Dans cette analyse où les individus sont presque parfaitement représentés sur le premier plan, cet indicateur apporte peu par rapport à l'examen visuel (ces quatre villes sont à la périphérie du nuage). Lorsque ce n'est pas le cas, cet indicateur est précieux pour détecter rapidement des individus particuliers. Remarque : quelques logiciels fournissent le carré de cette distance.

Un autre point de vue pour détecter des individus particuliers consiste à calculer leur inertie, par rapport au point moyen et rapportée à l'inertie totale du nuage. Lorsque les individus ont le même poids, ce qui est le cas ici et est d'ailleurs le cas le plus fréquent, cet indicateur n'apporte qu'une nuance à la distance (en revanche, si les poids diffèrent d'un individu à l'autre, l'information est clairement différente). Dans ces données, on dira que Nice contient 20 % de la variabilité du jeu de données, ou que les quatre villes précédentes (Nice, Marseille, Brest et Strasbourg) en contiennent 58 %.

### 2.2.5 Conclusion

Ce cas est typique d'une ACP car il met en évidence un « effet taille » et une autre structure complémentaire que l'on peut appeler, en opposition à la première, « effet forme ».

En revanche, il présente deux particularités. D'abord, le premier plan factoriel reconstitue presque parfaitement les données, ce qui est en pratique d'autant plus rare que le nombre de variables est grand. Ensuite, chacun des deux facteurs est facilement interprétable, ce qui est précieux pour un exemple à finalité pédagogique, mais l'utilisateur rencontre ordinairement des situations plus complexes.

## 2.3 INTRODUCTION À LA MÉTHODE DE WARD (CLASSIFICATION AUTOMATIQUE)

### 2.3.1 Construction et description d'un arbre

La **figure 2.3** représente l'arbre hiérarchique obtenu en appliquant l'algorithme de Ward aux données *villes*×*températures mensuelles*. La lecture de cet arbre est intuitive et s'apparente à celle d'un arbre généalogique : moins il faut monter haut dans l'arbre pour relier deux villes, et plus la « parenté » entre ces deux villes est grande (c'est-à-dire que leurs températures mensuelles sont proches). Ainsi, l'arbre met en évidence, par exemple, une étroite parenté entre les courbes de température de Lyon et de Grenoble et une grande différence entre celles de Lyon et de Bordeaux.

Le principe général de construction d'un arbre hiérarchique par une méthode ascendante est simple.

On dispose initialement de l'ensemble des $I$ individus à classifier (on distingue usuellement *classifier* - construire une classification - et *classer* - mettre dans des classes préétablies), dits éléments terminaux, soit ici les 15 villes, ainsi que d'une relation de ressemblance entre individus, soit ici la distance euclidienne usuelle (cf. section 2.1 page 31) utilisée en ACP (l'interprétation conjointe d'un arbre hiérarchique et d'un plan factoriel implique que la ressemblance entre deux individus soit définie de la même manière dans les deux méthodes).

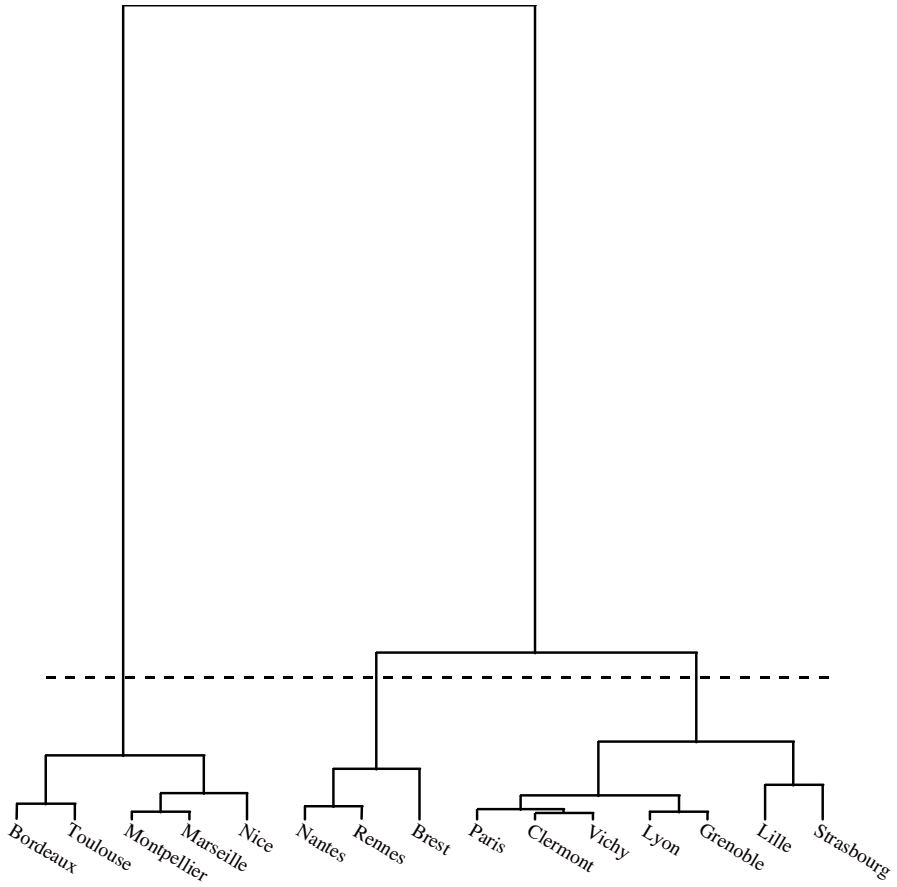



**Figure 2.3** Arbre hiérarchique issu de l'algorithme de Ward appliqué au tableau 2.1.

On commence par regrouper les deux éléments les plus proches. Dans l'exemple, ce sont Vichy et Clermont, ce qui est cohérent avec la position de ces deux villes sur le plan factoriel (cf. **Figure 2.2**). Plus directement, un rapide coup d'œil sur les données montre que ces deux villes ont des températures voisines tout au long de l'année, ce qui n'étonne pas compte tenu de leur proximité géographique. On constitue ainsi le premier nœud de l'arbre. La hauteur à laquelle on relie les éléments correspond à la ressemblance entre les éléments reliés : c'est l'indice de niveau du nœud. La définition de cet indice dans la méthode de Ward est indiquée plus loin.

À l'issue de l'agrégation de Vichy et Clermont, on ne dispose plus que de 14 éléments à classifier : 13 villes et un groupe de 2 villes. Dans l'exemple, l'algorithme regroupe ensuite Marseille et Montpellier. Ces deux villes ont des températures très

voisines, très légèrement moins que les 2 villes précédemment agrégées, ainsi que le montre le calcul de distances réalisé à partir des données initiales centrées-réduites.

$$\mathrm{d}(Vichy, Clermont) = .54 < \mathrm{d}(Marseille, Montpellier) = .86$$

L'indice de niveau du nœud correspondant à cette deuxième agrégation est donc plus élevé que le précédent. Et ainsi de suite, on agrège petit à petit les villes mais aussi les groupes de villes. Ce dernier point pose le problème de la définition de la ressemblance entre groupes de villes. Plusieurs possibilités existent, dont les plus simples sont les suivantes : la distance entre deux groupes A et B peut être définie comme la plus petite (algorithme dit du saut minimum) ou la plus grande (algorithme dit du diamètre) des distances entre deux éléments appartenant l'un à A l'autre à B. La façon dont la méthode de Ward résout ce problème est décrite plus loin.

Si l'on classifie $I$ individus, l'arbre contient $I - 1$ nœuds, qu'il est d'usage de numéroter de $I+1$ à $2J-1$. Les deux éléments réunis par chaque nœud sont quelquefois appelés l'un *aîné*, l'autre *benjamin*.

### 2.3.2 Arbre et partition

Un arbre hiérarchique peut être « coupé » pour faire apparaître une partition. Le niveau de coupure peut être matérialisé par une ligne horizontale. Ainsi, **figure 2.3**, le niveau de coupure (ligne horizontale en pointillé) fait apparaître une partition des villes en 3 classes : les 5 villes méridionales, les 3 villes les plus occidentales et enfin les 7 autres.

En élevant le niveau de coupure, on peut faire apparaître une partition en 2 classes, les 5 villes méridionales et les 10 autres. En abaissant le niveau de coupure, on fait apparaître successivement une partition en 4 classes, 5 classes, etc. Ainsi, en élevant le niveau de coupure à partir de la valeur 0, on met en évidence une **suite de partitions emboîtées**, depuis la partition la plus fine (dans laquelle chaque individu appartient à une classe distincte) jusqu'à la partition la plus grossière (dans laquelle tous les individus appartiennent à la même classe). Du fait de cette suite, un arbre hiérarchique est un outil commode pour raisonner le choix d'une partition.

### 2.3.3 Qualité d'une partition

Intuitivement, une partition d'un ensemble d'individus est bonne si :

1. à l'intérieur de chaque classe, la variabilité est faible, autrement dit si la variance des individus qui composent la classe est faible pour chaque variable ;
2. d'une classe à l'autre, la variabilité est grande, autrement dit si, pour chaque variable, la moyenne des individus qui composent une classe varie beaucoup d'une classe à l'autre.

| | Partition en | | |
|---|---|---|---|
| Inertie | 2 classes | 3 classes | 4 classes |
| **Totale** | **100.00** | **100.00** | **100.00** |
| **Inter-classes** | **65.68** | **78.70** | **84.44** |
| dont | | | |
| classe 1 | 43.79 (5) | 43.79 (5) | 43.79 (5) |
| classe 2 | 21.89 (10) | 11.73 (3) | 11.73 (3) |
| classe 3 | - | 23.19 (7) | 8.90 (5) |
| classe 4 | - | - | 20.03 (2) |
| **Intra-classes** | **34.32** | **21.30** | **15.56** |
| dont | | | |
| classe 1 | 7.29 (5) | 7.29 (5) | 7.29 (5) |
| classe 2 | 27.03 (10) | 3.98 (3) | 3.98 (3) |
| classe 3 | - | 10.03 (7) | 2.07 (5) |
| classe 4 | - | - | 2.21 (2) |

**Tableau 2.5** Décompositions de l'inertie relatives aux trois partitions les moins fines associées à l'arbre hiérarchique de la figure 2.3. Les inerties sont exprimées en % de l'inertie totale ; entre ( ) : effectifs des classes.

Heureusement, ces exigences ne sont pas contradictoires. Le théorème de Huygens exprime la décomposition de l'inertie totale, selon une partition, d'un nuage d'individus (cf. **Figure 12.1 page 297**). Soit :

Inertie totale = Inertie inter-classes + Inertie intra-classes.

L'inertie totale étant fixée par les données, il en résulte qu'il revient au même de rechercher une partition présentant une inertie inter grande ou une inertie intra petite. Cette décomposition suggère de mesurer la qualité globale d'une partition par le rapport *inertie inter/inertie totale* qui peut se voir comme la part d'inertie exprimée par la partition (de façon un peu analogue aux pourcentages d'inertie associés aux axes en ACP). Nous en discutons plus loin l'utilisation.

Le **tableau 2.5** récapitule les décompositions de l'inertie relatives aux partitions en 2, 3 et 4 classes associées à l'arbre.

L'inertie inter-classes peut être décomposée par classes, en considérant l'inertie du centre de gravité de chaque classe (affecté du poids égal à la somme des poids des individus de la classe). Ainsi la classe 1 (Nice, ..., Toulouse), présente dans les 3 partitions, joue un rôle essentiel dans les deux analyses (elle ne s'agrège à d'autres

classes qu'au dernier nœud de la CAH ; elle est clairement isolée par le premier axe de l'ACP) du fait qu'elle exprime, *en tant que classe* (*i.e.* en ne considérant que son centre de gravité), presque la moitié de la variabilité (43,79 %) totale. Remarque : on ne confondra pas cette inertie avec l'inertie totale de la classe ($43.79 + 7.29 = 51.08$) que l'on peut calculer directement à partir de la colonne *inertie* du tableau 2.2.

L'inertie intra-classes peut aussi être décomposée par classes. Ainsi, dans la partition en 2 classes, la seconde classe (les 10 villes du nord) contribue majoritairement à l'inertie intra-classe. Cela a deux origines : d'abord cette classe contient plus de villes ; ensuite elle est plus hétérogène, ce dont on peut se rendre compte en calculant l'inertie intra moyenne (*i.e.* la variance) par classe ($27.03/10 = 2.703 > 7.29/5 = 1.458$).

### 2.3.4 Algorithme de Ward

Remarques préliminaires

1. Au pas $n$, en agrégeant deux éléments (individus et/ou groupes d'individus), on passe d'une partition en $I - n + 1$ classes à une partition en $I - n$ classes.
2. La nouvelle partition (en $I - n$ classes), présente une inertie intra plus grande (éventuellement égale) que celle de la précédente (en $I - n + 1$ classes) : en agrégeant deux classes, on ne peut qu'augmenter l'inertie intra. Cela découle d'une autre forme du théorème de Huygens selon laquelle l'inertie d'un nuage par rapport à un point est minimum lorsque ce point est le centre de gravité du nuage (ce qui fait apparaître aussi que l'inertie intra n'augmente pas dans le seul cas très particulier où les deux classes agrégées ont le même centre de gravité).

L'idée de Ward consiste à choisir à chaque pas le regroupement de classes tel que **l'augmentation de l'inertie intra soit minimum**. Cet algorithme ne fournit évidemment pas des partitions globalement optimales (sauf au premier pas ce qui est sans intérêt pratique) : il faudrait pour cela remettre en cause à chaque pas les regroupements du pas précédent mais cela ferait perdre l'emboîtement des partitions et donc l'arbre hiérarchique.

Si l'on note :

1. $g_i$ (resp. $g_j$) le centre de gravité de la classe $i$ (resp. $j$),
2. $m_i$ (resp. $m_j$) la somme des poids des éléments de la classe $i$ (resp. $j$),

on montre que l'augmentation de l'inertie intra due au regroupement des classes $i$ et $j$ s'écrit :

$$\delta(i, j) = \frac{m_i m_j}{m_i + m_j} \mathrm{d}^2(g_i, g_j)$$

Tel est le critère minimisé à chaque pas et qui définit l'indice de niveau des nœuds de la hiérarchie. Cette écriture fait apparaître que, à chaque pas, on regroupe des classes :

1. proches, *i.e.* telles que $d^2(g_i, g_j)$ soit petit ;
2. de faibles poids, *i.e.* telles que $m_i m_j/(m_i + m_j)$ soit petit.

Ce dernier point montre bien pourquoi l'algorithme de Ward est peu sensible à l'effet de chaîne, fréquent par exemple lorsque l'on utilise l'algorithme du saut minimum, qui conduit à des arbres difficilement exploitables (cf. **Figure 2.4**) : l'algorithme de Ward favorise l'agrégation entre eux des éléments isolés.

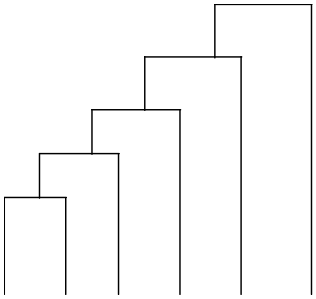


**Figure 2.4** Arbre hiérarchique présentant un effet de chaîne. Les individus s'agrègent un par un au groupe déjà constitué. Les partitions obtenues par coupure d'un tel arbre, mettant toutes en évidence un seul groupe et des individus isolés, sont généralement sans intérêt pratique.

On peut montrer que, lorsque l'algorithme de Ward agrège la classe $k$ à la classe (constituée à une étape antérieure de l'algorithme) réunissant les classes $i$ et $j$, l'augmentation de l'inertie intra est plus grande que celle consécutive à l'agrégation des classes $i$ et $j$. Soit : $\delta(k, \{i, j\}) \geqslant \delta(i, j)$. L'augmentation d'inertie intra étant utilisée comme indice de niveau, cette propriété assure que l'arbre hiérarchique ne présente pas d'inversion (cf. **Figure 2.5**).

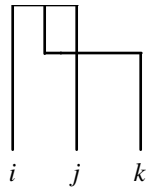



**Figure 2.5** Inversion dans un arbre hiérarchique. k s'agrège au groupe $\{i, j\}$ à un niveau inférieur à celui de l'agrégation entre $i$ et $j$. Ce phénomène est impossible avec les algorithmes usuels.

### 2.3.5 Utilisation des indices d'agrégation

#### *a) Propriété de la somme des indices de niveau*

La somme des indices de niveau, effectuée sur l'ensemble des $I - 1$ nœuds, est égale à l'inertie totale du nuage. Soit, en notant $\delta_n$ l'augmentation d'inertie intra au pas $n$ :

$$\sum_{n=1}^{n=I-1} \delta_n = \text{Inertie totale}$$

Cette propriété est immédiate en remarquant que :

1. la première partition (la plus fine, *i.e.* celle dont chaque classe est réduite à un seul élément) a une inertie intra nulle ;
2. la dernière partition (la plus grossière, *i.e.* celle réduite à une seule classe) a une inertie intra égale à l'inertie totale.

Il en découle que les indices peuvent être exprimés en valeur brute mais aussi en pourcentage de l'inertie totale : l'arbre hiérarchique propose alors une décomposition de l'inertie totale qu'il est intéressant de confronter à la décomposition de l'analyse factorielle sur les mêmes données.

### *b) Interprétation des plus hauts indices*

Dans l'exemple, le plus haut indice vaut 7.88 soit 65.68 % de l'inertie totale (égale à 12 ; cf. section 2.2.1). Ainsi, la partition en deux classes (villes chaudes/villes froides) exprime 65.68 % de la variabilité des données. Autrement dit, en ne considérant que ces deux classes, on a simplifié les données dans une grande proportion (on ne considère plus 15 villes mais 2 points moyens) tout en conservant 65.68 % de la variabilité.

Ce pourcentage est à comparer à celui associé au premier axe de l'ACP : 79.85 %. L'axe exprime plus de variabilité (il distingue, par exemple, Nice et Toulouse, ce que ne permet pas la partition en deux classes) mais est moins synthétique.

Toujours dans l'exemple, le deuxième indice (en partant du haut de l'arbre) vaut 1.56 soit 13.02 % de l'inertie totale. La séparation des 10 villes froides en 3 villes à faible amplitude thermique et 7 villes à forte amplitude thermique exprime donc 13.02 % de la variabilité des données. La comparaison entre ce pourcentage et celui associé au deuxième axe de l'ACP (18.97 %) conduit à un commentaire analogue à celui réalisé pour le premier axe.

En additionnant les pourcentages associés aux deux nœuds les plus élevés, on obtient le rapport *inertie inter/inertie totale* associé à la partition en trois classes : 78.70 % de la variabilité des données est exprimée par cette partition. Ce pourcentage est plus faible que celui associé au premier plan de l'ACP (98.82 %), ce qui correspond au caractère plus synthétique de la partition.

En abaissant encore le niveau de coupure, on augmente le nombre de classes et le rapport *inertie inter/inertie intra*, ce qui montre bien que ce dernier doit toujours être examiné en référence au nombre de classes de la partition et au nombre total d'individus (à la limite, la valeur la plus élevée de ce rapport, 1, est obtenue pour la partition qui contient un et un seul individu par classe, partition sans intérêt pratique).

*c) Diagramme des indices de niveau*

On représente classiquement les niveaux des nœuds (au moins pour les plus élevés lorsqu'il y a beaucoup d'individus) par un diagramme en bâtons (cf. **Figure 2.6**). On illustre ainsi ce que l'on gagne (en inertie inter c'est-à-dire, en quelque sorte, en représentation des données) lorsque l'on passe d'une partition donnée à la partition « immédiatement » plus fine. L'allure de ce diagramme suggère des niveaux de coupure privilégiés, ceux qui précèdent une décroissance rapide du gain en inertie inter.

Pour l'exemple, le diagramme suggère une coupure en 2, 3 ou 6 classes. Dans chacun de ces cas, le gain d'inertie inter obtenu en passant à la partition immédiatement plus fine est « sensiblement » plus petit que celui obtenu en considérant cette partition plutôt que celle immédiatement moins fine (comparer avec le cas des partitions en 4 et 5 classes).

## 2.4 CARACTÉRISATION DIRECTE D'UNE CLASSE D'INDIVIDUS

### 2.4.1 Problématique

L'arbre permet de définir chaque classe par l'énumération des individus qui la composent. Cela est tout à fait approprié dans l'exemple car les individus sont peu nombreux et leurs données sont familières. Mais, même dans ce cas, ce n'est pas suffisant pour connaître avec précision les caractéristiques communes des individus d'une classe.

L'idée la plus simple consiste à calculer, pour chaque variable $X$, la moyenne des individus de chaque classe. Pour une classe $q$ donnée, en comparant pour chaque variable la moyenne de la classe (notée $\bar{x}_q$) à la moyenne générale (notée $\bar{x}$), on peut caractériser la classe. Mais l'indicateur $\bar{x}_q - \bar{x}$ n'est pas suffisant car il doit être relativisé par l'effectif de la classe $q$ (noté $I_q$) et l'écart-type de la variable $X$ (noté $s$), comme l'illustre la figure 2.7.

La **figure 2.7** représente trois cas ayant la même moyenne générale ($\bar{x}$) et la même moyenne pour la classe $q(\bar{x}_q)$. La variable $X$ caractérise mieux la classe $q$ :

1. dans le cas 1 que dans le cas 2 ; $\bar{x}_q - \bar{x}$ doit être apprécié en tenant compte de l'écart-type général ;
2. dans le cas 3 que dans le cas 2 ; $\bar{x}_q - \bar{x}$ doit être apprécié en tenant compte de l'effectif de la classe $q$.

### 2.4.2 Notion de la valeur-test

L'idée de relativiser la quantité $\bar{x}_q - \bar{x}$ par l'écart-type général $s$ apparaît déjà dans le centrage et la réduction, transformation des données utilisée en préalable à l'ACP : on

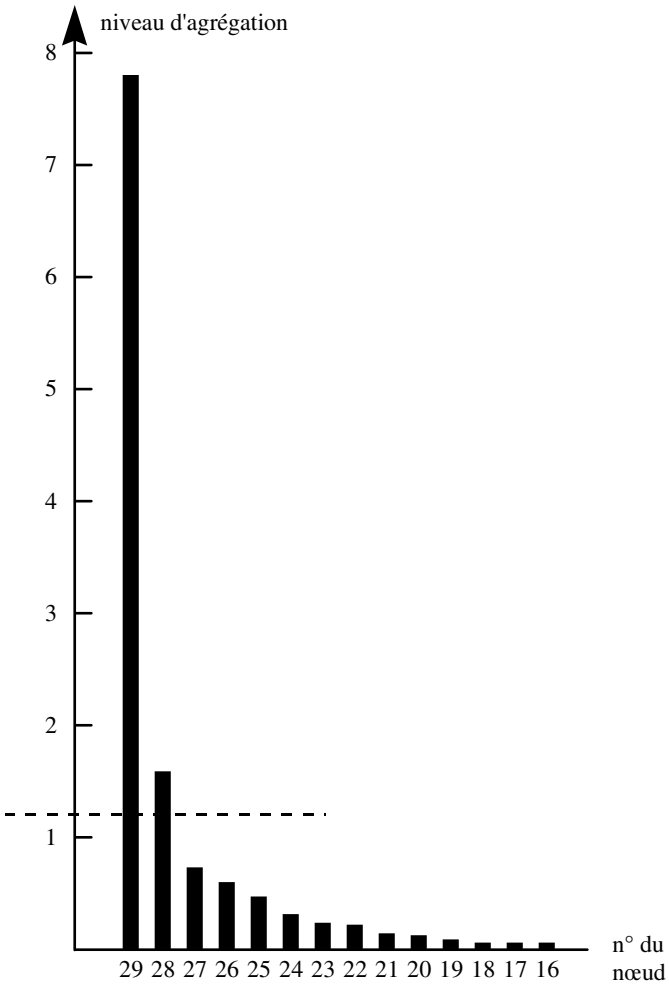



**Figure 2.6** Diagramme des indices de niveau de l'arbre de la figure 2.3. La ligne horizontale en pointillés matérialise le niveau de coupure en 3 classes.

choisit l'écart-type comme unité, ce qui permet de comparer entre elles des valeurs de variables différentes.

L'idée de relativiser par l'effectif de la classe se situe sur un tout autre plan. Empiriquement, on a l'intuition que même pour une variable qui n'a rien à voir avec la partition (ce serait le cas d'une variable supplémentaire « étrangère » aux variables actives), la différence $\bar{x}_q - \bar{x}$ n'est jamais (en pratique) exactement nulle et risque, l'écart-type général $s$ étant fixé, de s'écarter d'autant plus de 0 que l'effectif de la classe est faible.

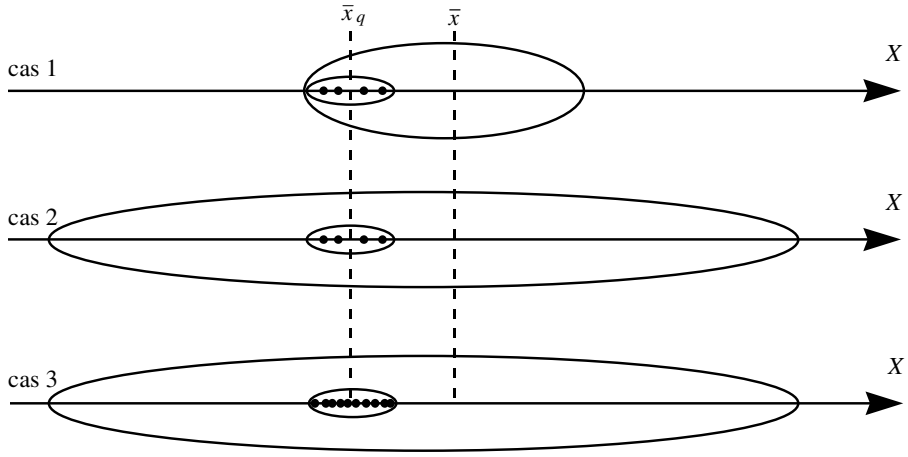



**Figure 2.7** Insuffisance de l'écart entre villes moyennes pour caractériser une classe. La grande ellipse représente l'ensemble des individus ; la petite rassemble les points de la classe $q$.

On peut formaliser la locution « n'a rien à voir avec la partition » par un modèle dans lequel les valeurs de la variable $X$ pour les individus de la classe $q$ sont tirées au hasard parmi les valeurs observées de $X$ sur les $I$ individus. En situant la valeur $\bar{x}_q$ observée par rapport aux valeurs attendues de cette moyenne dans le cadre du modèle de tirage au hasard, on appréhende d'une certaine manière l'écart entre les données et le modèle, ou, dit autrement, le caractère fortuit (*i.e.* imputable au hasard) de $\bar{x}_q$.

On montre facilement que la distribution des valeurs attendues de $\bar{x}_q$ a pour moyenne $\bar{x}$ et pour variance :

$$s^2_{\bar{x}_q} = \frac{s^2}{I_q}\frac{I - I_q}{I - 1}$$

On « situe » $\bar{x}_q$ dans cette distribution en calculant la **valeur-test** :

$$\frac{\bar{x}_q - \bar{x}}{s_{\bar{x}_q}} = \frac{(\bar{x}_q - \bar{x})\sqrt{I_q}}{s}\sqrt{\frac{I - I_q}{I - 1}}$$

La deuxième expression montre comment l'écart $\bar{x}_q - \bar{x}$ est « relativisé » par $s$ et $I_q$. Ainsi construite, la valeur-test, à l'instar des valeurs centrées réduites, est comparable d'une variable à l'autre et d'une classe à l'autre.

Le **tableau 2.6** donne quelques exemples de calculs de valeur-test.

1. **Cas 1 et 2** : Selon l'écart brut $\bar{x}_q - \bar{x}$, la classe {Nice, ..., Toulouse} est plus caractérisée par une température élevée en juillet qu'en avril. Mais la variabilité

**Tableau 2.6** Calcul de quelques valeurs-tests pour caractériser une classe.

| Cas | Classe | Variable | $\bar{x}_q$ | $\bar{x}$ | $\bar{x}_q - \bar{x}$ | $s$ | $s_{\bar{x}_q}$ | valeur-test |
|---|---|---|---|---|---|---|---|---|
| 1 | Nice ... Toulouse | avril | 12.70 | 10.98 | 1.72 | 1.37 | .518 | 3.33 |
| 2 | Nice ... Toulouse | juillet | 22.10 | 19.83 | 2.27 | 2.06 | .779 | 2.92 |
| 3 | Brest ... Nantes | juillet | 17.43 | 19.83 | -2.40 | 2.06 | 1.101 | -2.18 |

des températures est plus grande en juillet qu'en avril. Finalement, selon la valeur-test qui synthétise ces données, cette classe est (légèrement) mieux caractérisée par sa forte température en avril que par sa forte température en juillet.

2. **Cas 2 et 3** : Selon l'écart brut, la température en juillet caractérise moins la classe {Nice, ..., Toulouse} (par des valeurs élevées) que la classe {Brest, ..., Nantes} (par des valeurs basses). Mais l'effectif de la première (5) est plus important que celui de la seconde (3). Finalement, selon la valeur-test qui synthétise ces données, la température en juillet caractérise plus solidement la première classe que la seconde.

On notera ici que c'est la valeur absolue de la valeur-test qui indique le degré de caractérisation d'une classe par une variable ; le signe indique le sens (moyenne de classe plus basse ou plus élevée que la moyenne générale) de cette caractérisation.

### 2.4.3 Synthèse : le tableau de caractérisation (cf. Tableau 2.7)

Pour chaque classe d'individus, on trie l'ensemble des variables par valeurs-tests décroissantes. On fait ainsi apparaître, en haut et en bas de la liste, les variables qui caractérisent le mieux une classe donnée. Éventuellement, lorsque le nombre de variables est important, on ne fait pas apparaître les variables relatives aux valeurs-tests les plus faibles en valeur absolue.

Le **Tableau 2.7** fournit directement la caractérisation des classes. Nous le résumons en 3 points :

1. Les individus de la classe 1 sont caractérisés par une température élevée toute l'année, particulièrement en demi-saison. Ces villes sont méridionales (faible latitude).
2. « À l'opposé », les individus de la classe 3 sont caractérisés par une température faible toute l'année, particulièrement pendant les moins les plus froids.
3. La classe 2 comporte des villes présentant une faible amplitude thermique ; elles sont situées à l'ouest (faible longitude).

**Tableau 2.7** Caractérisation des 3 classes de villes par l'ensemble des variables. Pour chaque classe, les variables sont triées par valeurs-tests décroissantes.

Classe 1 : Nice, Marseille, Montpellier, Bordeaux, Toulouse

| | | Moyennes | | Ecarts-types | | |
|---|---|---|---|---|---|---|
| V. test | Proba | classe | générale | classe | général | Variable |
| 3.40 | 0,001 | 19.28 | 16.99 | 0.75 | 1.79 | septembre |
| 3.39 | 0,001 | 13.79 | 11.81 | 0.74 | 1.55 | moyenne annuelle |
| 3.33 | 0,001 | 12.70 | 10.98 | 0.58 | 1.37 | avril |
| 3.32 | 0,001 | 14.54 | 12.32 | 0.94 | 1.77 | octobre |
| 3.24 | 0,001 | 10.04 | 8.23 | 0.52 | 1.48 | mars |
| 3.18 | 0,001 | 21.90 | 19.57 | 0.79 | 1.94 | août |
| 3.00 | 0,003 | 19.80 | 17.83 | 0.73 | 1.73 | juin |
| 3.00 | 0,003 | 16.08 | 14.43 | 0.69 | 1.45 | mai |
| 2.97 | 0,003 | 9.88 | 7.93 | 1.00 | 1.74 | novembre |
| 2.92 | 0,004 | 22.10 | 19.83 | 1.00 | 2.06 | juillet |
| 2.88 | 0,004 | 6.80 | 4.83 | 0.94 | 1.81 | février |
| 2.54 | 0,011 | 6.66 | 4.85 | 0.90 | 1.89 | décembre |
| 2.46 | 0,014 | 5.78 | 3.97 | 0.92 | 1.94 | janvier |
| 0.65 | 0,516 | 3.37 | 2.58 | 2.68 | 3.21 | longitude |
| 0.50 | 0,617 | 16.34 | 15.91 | 0.99 | 2.25 | amplitude annuelle |
| -2.95 | 0,003 | 43.56 | 46.04 | 0.47 | 2.22 | latitude |

Classe 2 : Brest, Rennes, Nantes

| | | Moyennes | | Ecarts-types | | |
|---|---|---|---|---|---|---|
| V. test | Proba | classe | générale | classe | général | Variable |
| 1.49 | 0,136 | 47.81 | 46.04 | 0.48 | 2.22 | latitude |
| 1.28 | 0,201 | 5.30 | 3.97 | 0.57 | 1.94 | janvier |
| 1.11 | 0,267 | 5.97 | 4.85 | 0.73 | 1.89 | décembre |
| 0.66 | 0,509 | 5.47 | 4.83 | 0.24 | 1.81 | février |
| 0.44 | 0,660 | 8.33 | 7.93 | 0.50 | 1.74 | novembre |
| -0.25 | 0,803 | 8.03 | 8.23 | 0.26 | 1.48 | mars |
| -0.41 | 0,682 | 11.93 | 12.32 | 0.25 | 1.77 | octobre |
| -0.74 | 0,459 | 11.20 | 11.81 | 0.38 | 1.55 | moyenne annuelle |
| -1.30 | 0,194 | 10.03 | 10.98 | 0.65 | 1.37 | avril |
| -1.45 | 0,147 | 15.60 | 16.99 | 0.70 | 1.79 | septembre |
| -2.02 | 0,043 | 12.87 | 14.43 | 0.95 | 1.45 | mai |
| -2.02 | 0,043 | 17.47 | 19.57 | 1.09 | 1.94 | août |
| -2.05 | 0,040 | 15.93 | 17.83 | 1.16 | 1.73 | juin |
| -2.18 | 0,029 | 17.43 | 19.83 | 1.35 | 2.06 | juillet |
| -2.88 | 0,004 | -2.34 | 2.58 | 1.38 | 3.21 | longitude |
| -2.95 | 0,003 | 12.37 | 15.91 | 1.56 | 2.25 | amplitude annuelle |

Classe 3 : Lyon, Grenoble, Strasbourg, Vichy, Clermont, Paris, Lille.

| | | Moyennes | | Ecarts-types | | |
|---|---|---|---|---|---|---|
| V. test | Proba | classe | générale | classe | général | Variable |
| 1.89 | 0,059 | 17.13 | 15.91 | 1.44 | 2.25 | amplitude annuelle |
| 1.69 | 0,091 | 4.13 | 2.58 | 1.68 | 3.21 | longitude |
| 1.60 | 0,110 | 47.05 | 46.04 | 1.88 | 2.22 | latitude |
| -1.00 | 0,317 | 19.24 | 19.83 | 1.04 | 2.06 | juillet |
| -1.19 | 0,234 | 17.24 | 17.83 | 0.91 | 1.73 | juin |
| -1.22 | 0,222 | 13.93 | 14.43 | 0.74 | 1.45 | mai |
| -1.38 | 0,168 | 18.80 | 19.57 | 0.88 | 1.94 | août |
| -2.05 | 0,040 | 15.94 | 16.99 | 0.74 | 1.79 | septembre |
| -2.11 | 0,035 | 10.16 | 10.98 | 0.64 | 1.37 | avril |
| -2.60 | 0,009 | 10.66 | 11.81 | 0.62 | 1.55 | moyenne annuelle |
| -2.81 | 0,005 | 10.90 | 12.32 | 0.66 | 1.77 | octobre |
| -2.85 | 0,004 | 7.03 | 8.23 | 0.81 | 1.48 | mars |
| -3.15 | 0,002 | 6.36 | 7.93 | 0.65 | 1.74 | novembre |
| -3.25 | 0,001 | 3.16 | 4.83 | 0.76 | 1.81 | février |
| -3.28 | 0,001 | 3.07 | 4.85 | 0.91 | 1.89 | décembre |
| -3.36 | 0,001 | 2.11 | 3.97 | 0.88 | 1.94 | janvier |

### 2.4.4 Valeur-test et probabilité associée

La distribution des valeurs possibles de $\bar{x}_q$ dans le cadre du modèle de tirage au hasard a été jusqu'ici caractérisée par sa moyenne et sa variance. En outre, à condition que :

1. $I_q$ soit sensiblement plus petit que $I$ (ce qui rend « à peu près » indépendants les $I_q$ tirages dans l'ensemble des $I$ valeurs de $x$),
2. $I_q$ soit grand (ce qui assure que $\bar{x}_q$ est la moyenne de nombreuses quantités),

cette distribution peut être approchée par une loi Normale, ce qui permet de calculer une probabilité associée à la valeur observée. Cette probabilité a pour objet de situer la valeur observée $\bar{x}_q$ dans la distribution des $\bar{x}_q$ possibles : on la définit comme la probabilité d'obtenir, dans le cadre du modèle de tirage au hasard, une valeur de $\bar{x}_q$ au moins aussi éloignée de $\bar{x}$ que ne l'est la valeur de $\bar{x}_q$ effectivement observée (cf. **Figure 2.8**), soit :

$$P[X \geqslant |\text{valeur-test}|]$$

avec $X$ distribuée selon une loi normale centrée réduite.

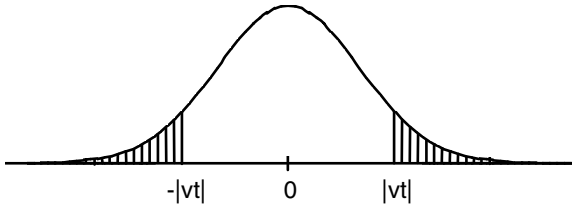



**Figure 2.8** Valeur-test et probabilité associée. L'aire hachurée matérialise la probabilité associée à la valeur-test vt.

L'intérêt de cette probabilité est de fournir un éclairage complémentaire à celui de la valeur-test. Ainsi, une valeur-test de 2 peut être appréciée en disant que l'écart de moyenne correspondant possède environ (« environ » rappelle que la distribution normale n'est qu'une approximation de la distribution réelle) 5 chances sur 100 d'être obtenu ou dépassé dans le cadre du modèle de tirage au hasard.

### 2.4.5 Probabilité associée et test d'hypothèse

Le calcul de probabilité présenté ci-dessus est identique à celui que l'on réalise dans une procédure de test d'hypothèse. Mais le cadre d'analyse ainsi que la problématique sont différents et l'on se gardera d'interpréter ces probabilités associées en termes

de tests d'hypothèse (par exemple on évitera de dire « la température en septembre caractérise *significativement* la classe 1 »).

Ceci est particulièrement flagrant dans le cas des variables actives puisque ce sont elles qui ont servi à définir les classes : l'hypothèse d'absence de différences entre les classes n'est évidemment pas adéquate. Mais, cela l'est encore pour bon nombre de variables supplémentaires : dans l'exemple, c'est bien sûr le cas de la moyenne annuelle et de l'amplitude thermique qui toutes deux combinent les variables actives et ont été calculées parce qu'elles correspondent aux deux premiers facteurs de l'ACP, mais aussi celui de la latitude et de la longitude dont l'introduction en tant que variables supplémentaires a été suggérée par l'analyse du plan factoriel.

Il n'en reste pas moins vrai qu'une classe donnée est souvent bien caractérisée par certaines variables, absolument pas par d'autres et que, plus ou moins explicitement, on ressent la nécessité d'établir une limite, ne serait-ce que pour déterminer la liste des variables à retenir dans la caractérisation des classes. Compte tenu de ce qui a été dit, cette limite ne peut être choisie qu'empiriquement, en s'appuyant sur les principes suivants :

1. les valeurs-tests se servant mutuellement de références, on néglige les valeurs les plus petites ; ainsi on a négligé la longitude et l'amplitude pour caractériser la classe1 ;
2. on peut conserver, en exprimant une nuance, les variables associées à de faibles valeurs-tests mais dont le contenu est cohérent avec les variables les plus caractéristiques ; la classe 3 est « mieux » décrite par « villes froides tout au long de l'année, plus particulièrement de septembre à avril » plutôt que « villes froides de septembre à avril » ;
3. on peut conserver, en exprimant une nuance, les variables associées à une faible valeur-test mais dont l'écart est jugé important (pour les classes de plus faibles effectifs, dont les valeurs-tests ont tendance à être plus faibles) ; ainsi on décrira la classe 2 comme comportant « des villes plutôt froides d'avril à septembre », mois pour lesquels la moyenne de la classe est inférieure à la moyenne générale d'au moins (environ) 1°.

## 2.5 INTERPRÉTATION SIMULTANÉE D'UN PLAN FACTORIEL ET D'UN ARBRE HIÉRARCHIQUE

### 2.5.1 Graphiques

Le principal outil d'examen simultané des résultats des deux méthodes consiste à faire apparaître, sur les plans factoriels, des éléments de la classification, soit :

1. l'appartenance des individus aux classes d'une partition, en représentant chacun par un symbole ou un numéro de classe ; cela est précieux, lorsque les individus sont nombreux (ce n'est pas le cas dans l'exemple, pour lequel il est préférable de délimiter les classes par leur contour, ce qui permet de laisser les identificateurs en clair), pour faire apparaître les dimensions pour lesquelles les classes se séparent et celles pour lesquelles elles se recouvrent ;
2. les centres de gravité des classes définies par une ou plusieurs partitions ; moins riche mais plus synthétique que la précédente, cette représentation est utile dans la confrontation de plusieurs partitions ;
3. le haut de l'arbre hiérarchique, à condition de représenter le plan factoriel en perspective.

L'objectif est d'utiliser les résultats de chaque analyse en tant qu'aides à l'interprétation de l'autre.

Ainsi,

1. les axes factoriels permettent de caractériser de façon synthétique les classes qu'ils séparent ; c'est ce qui a été fait implicitement jusqu'ici, car nous avions à l'esprit l'emplacement des villes sur le plan factoriel lors de la description des classes ;
2. les classes permettent de caractériser les axes sur lesquels elles se séparent ; c'est dans cet esprit que l'on a dit que le premier axe oppose les villes froides et les villes chaudes ; la classification étaie de telles interprétations.

Le faible nombre de points de l'exemple permet de remplacer ces trois représentations par un seul graphique (cf. **Figure 2.9**). La visualisation proposée est une synthèse commode de l'ensemble des résultats.

Sur cet exemple simple, l'apport de la classification ne semble pas décisif : ainsi la partition en 3 classes, par exemple, aurait sans doute pu être construite « à la main », sur la seule vue du plan factoriel. La raison en est que ces données étant quasiment bi-dimensionnelles, l'algorithme de classification s'appuie essentiellement sur les coordonnées pour les deux premiers axes, ce que nous savons faire visuellement. En revanche, lorsque la représentation des données exige plus de deux dimensions, la classification (qui s'appuie sur toutes les dimensions) est irremplaçable : elle « assure » que les points que l'on regroupe sont proches dans l'espace entier et non sur le seul plan factoriel.

### 2.5.2 Indicateurs

Le tableau 2.5 met en évidence la décomposition de l'inertie associée à une partition dans l'espace entier ($R^K$). Chaque terme de cette décomposition peut lui-même être décomposé axe par axe. Cette nouvelle décomposition, appliquée à la partition en

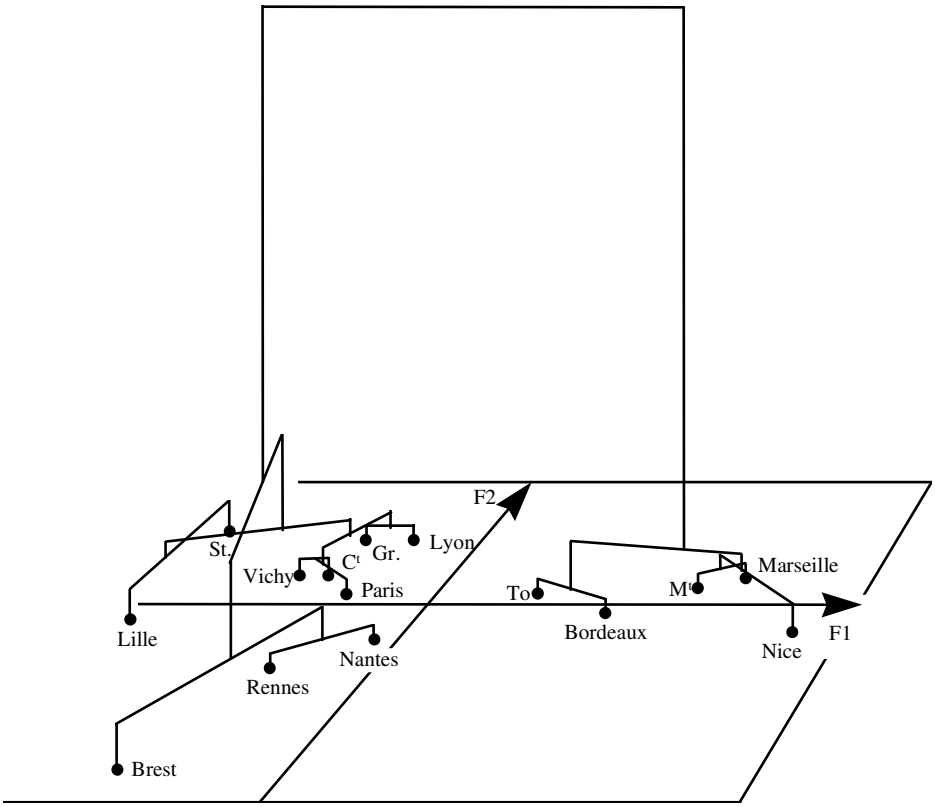



**Figure 2.9** Représentation simultanée d'un arbre hiérarchique (cf. Figure 2.3) et d'un plan factoriel (cf. Figure 2.2).

trois classes, est donnée **tableau 2.8**, dans lequel chaque inertie est exprimée en pourcentages, par rapport :

1. à la somme de sa colonne, c'est-à-dire à l'inertie associée à l'axe correspondant ; ce pourcentage s'interprète comme une contribution à l'axe ; par exemple, la spécificité de la classe 1 (*i.e.* ce qui distingue son centre de gravité du centre gravité général) contribue pour 54.80 % à l'inertie du premier axe ;
2. à la somme de sa ligne (dans cette somme, tous les axes sont pris en compte même si seuls les deux premiers apparaissent dans le tableau), c'est-à-dire à l'inertie de la ligne exprimée dans l'espace complet ; ce pourcentage s'interprète comme une qualité de représentation (au sens du rapport [*inertie projetée* / *inertie totale*]) ; par exemple, la spécificité de la classe 1 est exprimée presque parfaitement (à 99.93 %) par le premier axe.

| | Contributions | | Qualités de représentation | | |
|---|---|---|---|---|---|
| Inertie (effectifs) | F1 | F2 | F1 | F2 | $\sum$(F1, F2) |
| **Totale (15)** | **100.00** | **100.00** | **79.85** | **18.97** | **98.82** |
| **Inter classes** | **83.47** | **63.36** | **84.69** | **15.27** | **99.96** |
| dont | | | | | |
| classe 1 (5) | 54.80 | 1.50 | 99.93 | .06 | 99.99 |
| classe 2 (3) | 3.69 | 46.19 | 25.11 | 74.71 | 99.82 |
| classe 3 (7) | 24.99 | 17.02 | 86.06 | 13.92 | 99.98 |
| **Intra classes** | **16.53** | **36.64** | **61.96** | **32.64** | **94.60** |
| dont | | | | | |
| classe 1 (5) | 7.40 | 4.33 | 81.03 | 11.27 | 92.30 |
| classe 2 (3) | 1.39 | 14.69 | 27.79 | 70.01 | 97.80 |
| classe 3 (7) | 7.74 | 17.62 | 61.65 | 33.34 | 94.99 |

**Tableau 2.8** Inerties associées à la partition en 3 classes, décomposées selon les deux premiers axes factoriels. Contributions : inerties exprimées en % de l'inertie totale de l'axe. Qualités de représentation : inerties exprimées en % de l'inertie dans l'espace complet.

Ce tableau permet de quantifier de façon systématique des faits déjà observés, par exemple l'importance, déjà signalée, de la classe 1 dans l'inertie de l'axe 1 (54.80). Attention, ce pourcentage n'est pas égal à la somme des contributions à l'axe 1 des éléments de la classe 1, calculable à partir du tableau 2.2 : cette dernière somme, qui vaut 62.20 %, est égale à la somme des inerties inter et intra de la classe 1. Cette distinction correspond aux deux points de vue usuels pour prendre en compte une classe : son centre de gravité et l'ensemble de ses éléments (remarque déjà faite à propos du tableau 2.5 dans l'espace complet).

Dans l'exemple, on remarque que, pour une classe donnée, c'est le même axe qui représente bien son inertie inter et son inertie intra. Cela indique que les éléments possèdent les caractéristiques de leur classe à un niveau très variable. Ainsi, pour illustrer une classe par un de ses éléments, on peut opter, selon les cas, pour :

1. l'élément le plus proche du centre de gravité de la classe ; en ce sens, Rennes représente bien la classe 2 ;
2. l'élément le plus éloigné de l'origine dans la direction du centre de gravité de la classe ; en ce sens, Brest illustre bien les caractéristiques de la classe 2 puisqu'elle les possède à un niveau extrême.

Deux autres indicateurs, rassemblés **tableau 2.9** sont très utiles pour analyser une partition.

**Tableau 2.9** Trois indicateurs importants dans l'analyse d'une partition. La distance et les coordonnées sont celles des centres de gravité des classes. La valeur-test, qui prend en compte la coordonnée, l'effectif de la classe et l'inertie de l'axe, est comparable d'un axe à l'autre et d'une classe à l'autre.

| Classe | Distance à l'origine | Coordonnées | | Valeurs-tests | |
|---|---|---|---|---|---|
| | | F1 | F2 | F1 | F2 |
| 1 : Nice, ..., Toulouse | 3.97 | 3.97 | .10 | 3.39 | .18 |
| 2 : Brest, ..., Nantes | 2.65 | -1.33 | -2.29 | -.80 | -2.84 |
| 3 : Lyon, ..., Lille | 2.44 | -2.27 | . 91 | -2.56 | 2.11 |

1. La distance (ou son carré) dans l'espace complet entre le centre de gravité de la classe et le centre de gravité général. En ce sens, la classe 1 est celle qui se différencie le plus, résultat en harmonie avec le tableau 2.5 (l'inertie de cette classe ramenée à son centre de gravité représente 43.79 % de l'inertie totale) et le premier axe de l'ACP.
2. La valeur-test, définie à propos d'une variable initiale, s'applique aux axes factoriels. Ainsi, d'après les coordonnées, la classe 3 se caractérise surtout par l'axe1. D'après sa valeur-test pour l'axe 2, qui prend en compte la beaucoup plus faible variabilité des 15 villes selon cet axe, la classe 3 peut aussi être caractérisée par sa coordonnée pour l'axe 2.

### 2.5.3 Mise en œuvre conjointe d'une ACP et d'une CAH

Beaucoup de logiciels réalisent d'abord une ACP dont ils calculent tous les facteurs sur $I$. La CAH est alors mise en œuvre à partir des facteurs et non des données brutes. Il est clair qu'il revient au même de travailler à partir de tous les facteurs ou à partir des données brutes. L'intérêt de cette démarche est double :

1. en travaillant sur les facteurs d'une autre analyse factorielle (*e.g.* une AFC), le même programme permet d'obtenir une CAH réalisée sur d'autres types de données (*e.g.* les lignes d'un tableau de contingence) ;
2. en sélectionnant les $S$ premiers facteurs, l'analyse factorielle joue le rôle d'un filtre en éliminant les dimensions de très faible inertie assimilables à du « bruit » ; cela est surtout précieux dans le cas de variables qualitatives (donc après une ACM).

## 2.6 CONSTRUCTION ET AMÉLIORATION D'UNE PARTITION

### 2.6.1 Principe

Pour construire directement une partition de $I$ individus en $Q$ classes, plusieurs algorithmes procèdent en tirant au hasard une partition initiale et en améliorant pas à pas cette partition. Le plus simple d'entre eux, dit *agrégation autour des centres mobiles*, améliore à chaque pas la partition ainsi :

1. on calcule le centre de gravité de chaque classe $q$ ;
2. on réaffecte chaque individu $i$ à la classe $q$ dont le centre de gravité est celui qui est le plus proche de $i$ ;
3. si la composition des classes change, alors les centres de gravité aussi et l'on réitère l'opération à partir de 1 ; sinon, le calcul est terminé.

La justification de cet algorithme tient au fait qu'à chaque itération on ne peut que diminuer l'inertie intra-classe. Naturellement cette propriété n'assure pas que l'on obtienne la partition optimale.

**Remarque** : pour initialiser l'algorithme, plutôt que de tirer une partition au hasard, on tire au hasard les centres de gravité des classes. Ces centres sont choisis parmi les individus à partitionner.

Malgré une apparence assez fruste, cet algorithme fournit en pratique des partitions acceptables en un nombre d'itérations faible. Une illustration sommaire en est donnée **Figure 2.10**.

### 2.6.2 Utilisation à l'issue d'une classification hiérarchique.

La partition induite par une coupure d'un arbre hiérarchique n'est en général pas optimale. En l'utilisant comme point de départ de l'algorithme des centres mobiles, on ne peut qu'améliorer (on dit aussi *consolider*) cette partition. En pratique, cette amélioration n'est généralement pas décisive (le rapport *inertie inter/inertie totale* ne progresse que de quelques points). Ainsi, dans l'exemple, aucune des partitions (induites par l'arbre hiérarchique des 15 villes) n'est modifiée par cette procédure.

**Remarque** : une partition ainsi améliorée n'est plus compatible avec l'arbre hiérarchique dont elle est issue. Si l'on souhaite un arbre hiérarchique représentant les classes d'une telle partition, il faut reprendre l'algorithme de Ward à partir de ses classes.

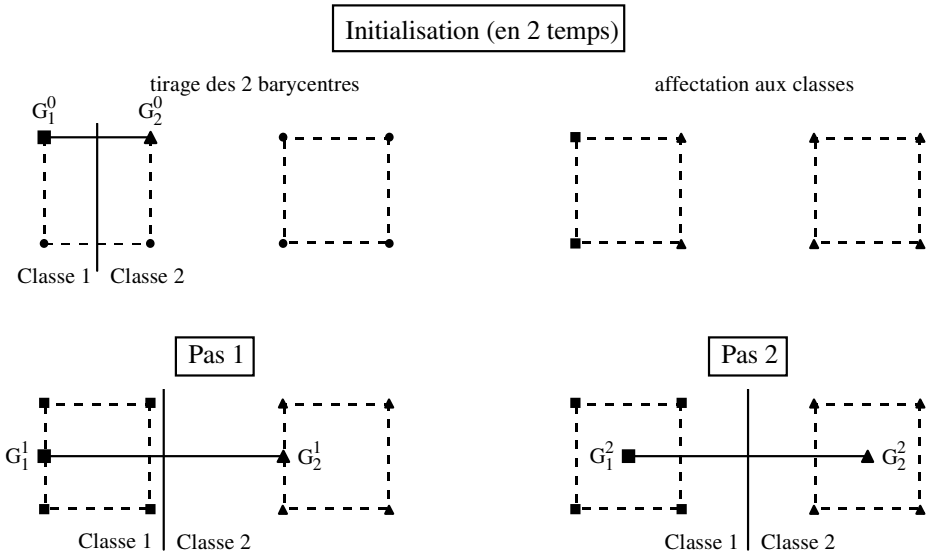



**Figure 2.10** Illustration de l'algorithme d'agrégation autour des centres mobiles. Données : 8 individus situés aux sommets de 2 carrés. Initialisation : le tirage au hasard a conduit aux barycentres $G_1^0$ et $G_2^0$ ; la médiatrice du segment $G_1^0$ $G_2^0$ permet de définir l'affectation des individus aux classes : chaque individu est affecté à la classe correspondant au barycentre dont il est le plus proche. Pas 1 : on calcule les barycentres $G_1^1$ et $G_2^1$ des classes du pas précédent ; la médiatrice du segment $G_1^1$ $G_2^1$ permet de définir une nouvelle affectation des individus aux classes. Pas 2 : on calcule les barycentres $G_1^2$ et $G_2^2$ des classes du pas précédent ; l'affectation aux classes 1 et 2 induite par la médiatrice de $G_1^2$ $G_2^2$ est identique à celle du pas précédent. L'algorithme a convergé.

### 2.6.3 Utilisation en amont d'une classification hiérarchique

Lorsque le nombre d'individus à classifier est très grand, (disons au-delà de 1 000 pour fixer les idées), la construction d'un arbre hiérarchique peut ne pas être réalisable pour des raisons de capacité machine. Une pratique commode consiste alors à construire, à l'aide de l'algorithme des centres mobiles, une partition en un grand nombre de classes (disons une centaine pour fixer les idées) et à construire ensuite, à partir de ces classes, un arbre hiérarchique. L'arbre obtenu par cette procédure (dite de classification mixte) s'utilise exactement comme le haut d'un arbre issu de l'algorithme usuel.

Chapitre **3**

# Analyse Factorielle des Correspondances

## 3.1 DONNÉES, NOTATIONS, HYPOTHÈSE D'INDÉPENDANCE

À l'origine, l'Analyse Factorielle des Correspondances (AFC) a été conçue pour étudier des tableaux appelés couramment tableaux de contingence (ou tableaux croisés). Il s'agit de tableaux d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de $n$ individus. Dans l'exemple commenté au chapitre 10, la population est constituée par l'ensemble des individus qui ont quitté le système scolaire français en 1972 et qui occupent un emploi en 1973 ; pour chaque individu, on connaît son niveau de diplôme et sa catégorie d'emploi. La **figure 3.1** résume les principales notations.

On parle indifféremment de la modalité $i$ (par exemple le baccalauréat) ou de la classe $i$, c'est-à-dire de la classe des individus qui possèdent la modalité $i$ (par exemple les bacheliers).

Dans ce chapitre, nous nous limitons à l'étude d'un tableau de contingence. Cependant, la plupart des notions introduites et des résultats présentés peuvent être généralisés à des tableaux qui ne sont pas strictement de ce type. Le cas très important du tableau disjonctif complet fait l'objet d'un chapitre particulier : l'Analyse des Correspondances Multiples. La conclusion du présent chapitre donne quelques points de repère sur l'application de l'AFC à d'autres tableaux que les tableaux de contingence.

On considère souvent le tableau des fréquences relatives $F$, obtenu en divisant chaque effectif $k_{ij}$ par l'effectif total $n$. Ce nouveau tableau définit une mesure de probabilité sur l'ensemble produit $I \times J$. Ses marges, ou probabilités marginales,

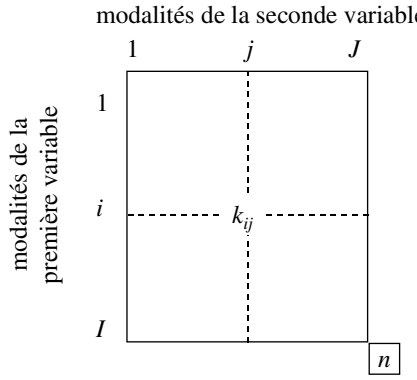
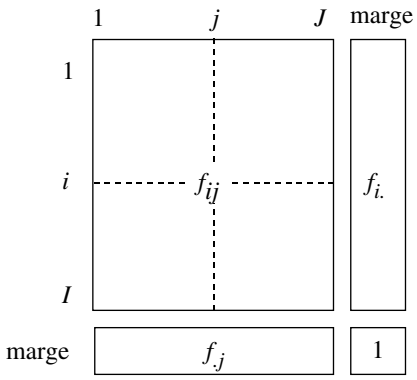



**Figure 3.1** Tableau des données brutes. $I$ : ensemble des lignes et nombre de lignes (8 niveaux de diplôme). $J$ : ensemble des colonnes et nombre de colonnes (9 catégories d'emploi). $k_{ij}$ : nombre d'individus possédant à la fois la modalité $i$ de la première variable et la modalité $j$ de la seconde (i.e. qui ont le niveau de diplôme $i$ et qui occupent un emploi de la catégorie $j$). $\sum_i \sum_j k_{ij} = n$ (nombre total d'individus).




$$
\begin{aligned}
f_{ij} &= k_{ij}/n \\
f_{i.} &= \sum_j f_{ij} \\
f_{.j} &= \sum_i f_{ij} \\
\sum_i f_{i.} &= \sum_j f_{.j} = \sum_i \sum_j f_{ij} = 1
\end{aligned}
$$

**Figure 3.2** Tableau F des fréquences relatives et ses marges.

ont pour terme général $f_{i.}$ pour la marge-colonne et $f_{.j}$ pour la marge-ligne (*cf.* **Figure 3.2**).

Un tableau de contingence exprime la liaison entre deux variables qualitatives. Classiquement, pour une mesure de probabilité, on dit qu'il y a **indépendance** entre les deux variables lorsque, pour tout $i$ et pour tout $j$, on a l'égalité :

$$f_{ij} = f_{i.} f_{.j}$$

Il y a **liaison** entre les deux variables dès que certaines cases du tableau $f_{ij}$ diffèrent du produit $f_{i.} f_{.j}$. Si $f_{ij}$ est supérieur à ce produit, les modalités $i$ et $j$ s'associent plus qu'elles ne le font dans l'hypothèse d'indépendance : on dit que $i$ et $j$ s'attirent. Au contraire, si $f_{ij}$ est inférieur au produit des marges, $i$ et $j$ s'associent moins que dans l'hypothèse d'indépendance : on dit qu'il y a répulsion entre ces deux modalités.

L'**indépendance** s'exprime aussi en considérant le tableau comme un ensemble de lignes. En effet, l'égalité ci-dessus est équivalente à l'égalité :

$$\frac{f_{ij}}{f_{i.}} = f_{.j}$$

La quantité $f_{.j}$ représente le pourcentage de la population totale qui possède la modalité $j$ tandis que $f_{ij}/f_{i.}$ représente ce même pourcentage dans la sous-population possédant la modalité $i$. Lorsqu'il y a indépendance, les $I$ sous-populations caractérisées par les modalités $i$ de la première variable se répartissent selon les $J$ modalités $j$ de la deuxième variable avec les mêmes pourcentages. Toutes les lignes sont alors proportionnelles. La réciproque est vraie : lorsque toutes les lignes sont proportionnelles, elles sont proportionnelles à la marge $f_{.j}$ et les deux variables sont indépendantes. Il y a donc **liaison** dès lors que les lignes ne sont pas toutes proportionnelles à la marge, c'est-à-dire lorsqu'elles ne sont pas identiques du point de vue de leur association avec l'ensemble des colonnes.

Remarquons enfin que, dans un tableau de contingence, les lignes et les colonnes jouent un rôle absolument symétrique : l'indépendance s'exprime de la même façon sur l'ensemble des colonnes. Les deux égalités ci-dessus sont en effet équivalentes à la suivante :

$$\frac{f_{ij}}{f_{.j}} = f_{\mathrm{i.}}$$

Il y a indépendance lorsque tous les pourcentages en colonnes sont égaux à la marge $f_{i.}$ c'est-à-dire lorsque les colonnes sont proportionnelles. Il y a liaison lorsqu'elles ne le sont pas.

## 3.2 OBJECTIFS

Bien que le tableau étudié soit de nature très différente de celui étudié en ACP, les objectifs de l'AFC peuvent s'exprimer de manière analogue à ceux de l'ACP : on cherche à obtenir une typologie des lignes, une typologie des colonnes et à relier ces deux typologies entre elles ; mais la notion de ressemblance entre deux lignes, ou entre deux colonnes, est différente de celle de l'ACP.

Dans un tableau de contingence, la ressemblance, entre deux lignes d'une part et entre deux colonnes d'autre part, s'exprime de manière totalement symétrique. Deux lignes sont considérées comme proches si elles s'associent de la même façon à l'ensemble des colonnes, c'est-à-dire si elles s'associent trop (ou trop peu) aux mêmes colonnes ; les termes « trop » et « trop peu » sont pris en référence à la situation d'indépendance. Symétriquement, deux colonnes sont proches si elles s'associent de la même façon à l'ensemble des lignes.

Schématiquement, l'étude de l'ensemble des lignes revient à mettre en évidence une typologie dans laquelle on cherche les lignes dont la répartition s'écarte le plus de celle de l'ensemble de la population, celles qui se ressemblent entre elles (dans le sens précisé ci-dessus) et celles qui s'opposent. Pour mettre en relation la typologie des lignes avec l'ensemble des colonnes, on caractérise chaque groupe de lignes par les colonnes auxquelles ce groupe s'associe trop ou trop peu.

L'étude de l'ensemble des colonnes est absolument analogue.

Cette approche, grâce à la notion de ressemblance utilisée, permet d'étudier la liaison entre les deux variables, c'est-à-dire l'écart du tableau à l'hypothèse d'indépendance. L'analyse de cette liaison est l'objectif fondamental de l'AFC.

Une approche complémentaire de la précédente, fait intervenir conjointement l'ensemble des lignes et celui des colonnes en ne privilégiant ni l'un ni l'autre. Prenons l'exemple du tableau croisant les catégories d'emploi et les niveaux de diplôme. L'ensemble des diplômes est ordonné par la longueur des études tandis que celui des catégories d'emploi l'est par le salaire moyen. La relation entre ces deux ordres (un salaire élevé correspond généralement à un diplôme élevé) explique clairement une bonne part de la liaison entre emplois et diplômes. Mais ce lien ne se restreint peut-être pas à cet unique aspect ; il peut exister d'autres phénomènes comme l'association presque exclusive de certains diplômes avec certains emplois. L'objectif de l'AFC est de décomposer la liaison entre deux variables en une somme (ou une superposition) de tendances simples et interprétables comme celles qui viennent d'être évoquées et de mesurer leur importance relative afin de les ordonner.

Enfin, bien qu'il y soit fait peu référence par la suite, il faut signaler que l'AFC, comme toute Analyse Factorielle, est utilisée aussi dans le but de réduire la dimension des données en conservant le plus d'information possible. Ceci en vue d'un traitement statistique ultérieur (classification, régression, analyse discriminante, etc.) ou d'une transmission d'information.

## 3.3 TRANSFORMATIONS DES DONNÉES EN PROFILS

En AFC, le tableau brut n'est pas analysé directement. Dans l'étude des lignes, le tableau des données est transformé en divisant chaque terme $f_{ij}$ de la ligne $i$ par la marge $f_{i.}$ de cette ligne $i$. La nouvelle ligne est appelée profil-ligne (*cf.* **Figure 3.3**).

Cette transformation découle de l'objectif qui vise à étudier la liaison entre les deux variables au travers de l'écart entre les pourcentages en lignes. Elle se justifie aussi de façon directe puisque la comparaison de deux lignes du tableau brut risque d'être influencée principalement par leurs effectifs marginaux. Ainsi, dans le tableau croisant emplois et diplômes, la différence entre les lignes brutes *Bac technique* et *Bac général* traduit essentiellement une différence entre les effectifs globaux de ces deux diplômes.

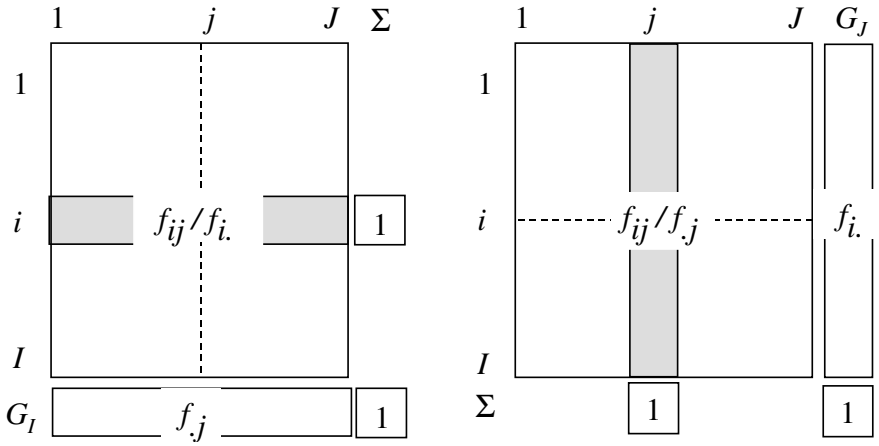



**Figure 3.3** Profil-ligne (à gauche) et profil-colonne (à droite). $G_I$ et $G_J$ : profils marginaux.

Le nombre $f_{ij}/f_{i.}$ représente, dans notre exemple, la probabilité d'occuper un emploi de la catégorie $j$ sachant que l'on détient le niveau de diplôme $i$. Le profil-ligne $i$ n'est rien d'autre que la loi de probabilité conditionnelle définie par $i$ sur l'ensemble des colonnes. Pour analyser l'écart à l'indépendance, on confronte ces profils au profil ligne marginal (= établi sur l'ensemble de la population) de terme général $f_{.j}$ et noté $G_I$.

Du fait du rôle symétrique joué par les lignes et les colonnes, un raisonnement analogue peut être mené à propos des colonnes. Il conduit à la notion de profil-colonne (*cf.* **Figure 3.3**).

Ainsi, en AFC, selon que l'on s'intéresse aux lignes ou aux colonnes, on ne considère pas le même tableau transformé. Toutefois, les deux transformations en profils possèdent la même signification vis-à-vis des objets qu'elles concernent. Ces transformations sont intéressantes en elles-mêmes indépendamment de tout contexte d'analyse factorielle. Lorsqu'un tableau croisé est commenté, il est presque toujours présenté sous la forme de pourcentages, par rapport aux lignes ou aux colonnes selon les aspects que l'on cherche à mettre en évidence.

## 3.4 RESSEMBLANCE ENTRE PROFILS : DISTANCE DU $\chi^2$

En AFC, la ressemblance entre deux lignes ou entre deux colonnes est définie par une distance entre leurs profils connue sous le nom de distance du $\chi^2$. Elle est définie de façon symétrique pour les lignes et pour les colonnes. Soit :

$$\begin{aligned} \mathrm{d}\chi^2(\text{profil-ligne } i, \text{profil-ligne } l) &= \sum_j \frac{1}{f_{.j}} \left( \frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2 \\ \mathrm{d}\chi^2(\text{profil-colonne } j, \text{profil-colonne } k) &= \sum_i \frac{1}{f_{i.}} \left( \frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2 \end{aligned}$$

Dans ces relations, la distance entre deux lignes dépend essentiellement des différences terme à terme entre les deux profils dont elle fait une somme des carrés pondérés. La pondération $1/f_{.j}$ équilibre l'influence des colonnes sur la distance entre les lignes : elle augmente les termes, *a priori* plus faibles, concernant les modalités rares ; elle joue, jusqu'à un certain point, un rôle analogue à celui de la division par l'écart-type dans le cas des variables numériques.

La distance du $\chi^2$ jouit d'une propriété fondamentale appelée **équivalence distributionnelle**. Selon cette propriété, si deux colonnes proportionnelles d'un tableau sont cumulées en une seule, la distance entre les profils-lignes est inchangée. Le cas d'une proportionnalité parfaite entre deux colonnes ne se rencontre guère en pratique mais constitue une situation limite dont on peut être assez proche. La propriété mathématique est alors utilisée sous la forme d'une règle pragmatique : remplacer, par leur somme, deux colonnes ou deux lignes presque proportionnelles ne modifie pas sensiblement les résultats d'une AFC. On se réfère surtout à cette règle lorsque plusieurs ensembles de modalités sont possibles pour définir une même variable. Ainsi, la variable *catégorie d'emploi* peut être plus ou moins détaillée : par exemple, on peut se demander si les catégories *ouvrier qualifié* et *ouvrier non qualifié* peuvent être regroupées en une seule catégorie. Du fait de l'équivalence distributionnelle, si ces deux catégories ont des profils voisins, le choix entre les deux solutions n'est pas fondamental puisque les AFC des deux tableaux aboutissent à des résultats analogues.

## 3.5 LES DEUX NUAGES

### 3.5.1 Nuage des profils-lignes

S'intéresser aux modalités de la première variable revient à considérer les données comme une juxtaposition de profils-lignes. Chaque profil-ligne est une suite de $J$ valeurs numériques et peut être représenté par un point de l'espace $R^J$ dont chacune des $J$ dimensions est associée à une modalité de la seconde variable. La distance

du $\chi^2$ définissant la ressemblance entre profils-lignes (*cf.* section 3.4) possède les propriétés d'une distance euclidienne et confère à $R^J$ la structure d'espace euclidien. Cette distance revient à affecter le poids $1/f_{.j}$ à la $j^{\text{e}}$ dimension de $R^J$. La somme des coordonnées de chaque profil-ligne vaut 1 ; il en résulte que le nuage $N_I$ appartient à un hyperplan, noté $H_I$ (*cf.* **Figure 3.4**).

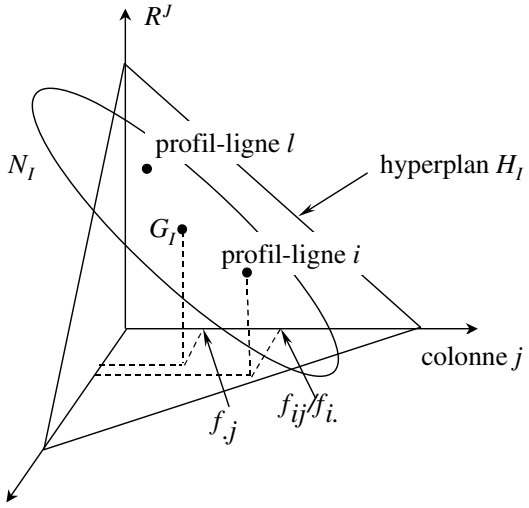



**Figure 3.4** Le nuage $N_I$ des profils-lignes dans $R^J$. Le point $i$ a pour coordonnée sur l'axe $j$ : $f_{ij}/f_{i.}$ ; son poids est $f_{i.}$ ; la distance entre deux profils est la distance du $\chi^2$ ; Le barycentre $G_I$ du nuage $N_I$ a pour coordonnée sur l'axe $j$ la fréquence marginale $f_{.j}$ ; le nuage $N_I$ appartient à un hyperplan noté $H_I$.

En AFC, les poids affectés à chaque point du nuage sont imposés. Le point $i$ a un poids égal à la fréquence marginale $f_{i.}$ (ce poids est proportionnel à l'effectif de la classe d'individus représentée par le point $i$).

Le **barycentre** des points de $N_I$ munis de ces poids est noté $G_I$. Sa $j^{\text{e}}$ coordonnée est égale à la fréquence marginale $f_{.j}$ .

$$f_{.j} = \sum_i f_{i.} \frac{f_{ij}}{f_{i.}}$$

Il s'interprète comme un profil moyen. Dans l'exemple du tableau qui croise les niveaux de diplôme et les catégories d'emploi, $G_I$ est le profil d'emplois de l'ensemble de la population, tous les diplômes étant cumulés. Il sert constamment de référence dans l'étude des lignes du tableau ; ainsi, étudier dans quelle mesure et de quelle façon une classe d'individus $i$ diffère de l'ensemble de la population revient à étudier l'écart entre le profil de cette classe $i$ et le profil moyen. Étudier la dispersion du nuage

autour de son barycentre revient à étudier l'écart entre les profils des lignes et le profil marginal, et donc la liaison entre les deux variables (*cf.* section 3.1).

### 3.5.2 Nuage des profils-colonnes

Compte tenu du rôle symétrique joué par les lignes et les colonnes en AFC, la construction du nuage des profils-colonnes s'effectue selon une démarche strictement identique à celle du nuage des profils-lignes. Il est toutefois utile de la décrire, ne serait-ce que pour fixer les notations.

S'intéresser aux modalités de la seconde variable revient à considérer les données comme une juxtaposition de profils-colonnes. Chaque profil-colonne est une suite de $I$ valeurs numériques et peut être représenté par un point de l'espace $R^I$ dont chacune des dimensions est associée à une modalité de la première variable. $R^I$ est muni d'une structure euclidienne par la distance du $\chi^2$ : à la $i^{\text{e}}$ dimension on affecte le poids $1/f_{i.}$ (*cf.* **Figure 3.5**).

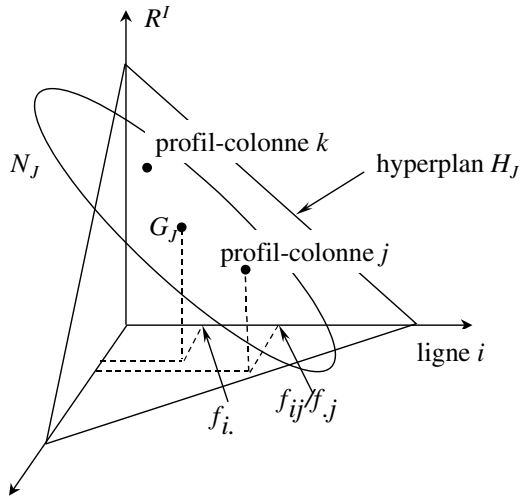



**Figure 3.5** Le nuage $N_J$ des profils-colonnes dans $R^I$. Le point $j$ a pour coordonnée sur l'axe $i$ : $f_{ij}/f_{.j}$ ; son poids est $f_{.j}$ ; la distance entre deux profils est la distance du $\chi^2$ ; le barycentre $G_J$ du nuage $N_J$ a pour coordonnée sur l'axe $i$ la fréquence marginale $f_{i.}$ ; le nuage $N_J$ appartient à un hyperplan noté $H_J$.

Le point $G_J$ représente la marge $\{f_{i.} | i = 1, ..., I\}$ ; c'est le barycentre de $N_J$ lorsque l'on munit chaque profil-colonne $j$ du poids $f_{.j}$ ; en tant que profil moyen, il sert constamment de référence dans l'étude de $N_J$.

## 3.6 AJUSTEMENT DES DEUX NUAGES

### 3.6.1 Ajustement du nuage des profils-lignes

Dans $R^J$, l'ajustement vise à obtenir une suite d'images planes approchées du nuage $N_I$. De la même façon que l'ACP, l'AFC procède en recherchant une suite d'axes orthogonaux sur lesquels le nuage $N_I$ est projeté. Chaque axe possède la propriété de rendre maximum l'inertie projetée du nuage $N_I$ avec la contrainte d'être orthogonal aux axes déjà trouvés.

Les images planes de $N_I$ doivent être telles que les distances entre les points de l'image ressemblent le plus possible aux distances entre les points de $N_I$. Cet objectif est tout à fait analogue à celui de l'ajustement du nuage des individus en ACP : pratiquement, il implique que le nuage analysé soit centré, c'est-à-dire que son barycentre soit choisi comme origine des axes (*cf.* section 3.5).

Dans le nuage centré, la classe définie par la modalité $i$ est représentée par un point dont la coordonnée sur le $j^{\text{e}}$ axe vaut : $f_{ij}/f_{i.} - f_{.j}$. La position de ce point exprime la différence entre la répartition, sur l'ensemble des modalités de la seconde variable, des individus de la classe $i$ et celle de la population totale. Ainsi, rechercher les directions d'inertie maximum du nuage centré revient à mettre en évidence les classes qui s'écartent le plus du profil de l'ensemble de la population.

Chaque profil est muni d'un poids égal à sa fréquence marginale $f_{i.}$. Ce poids intervient en premier lieu dans le calcul du barycentre du nuage. Il intervient aussi dans l'inertie et donc dans le critère d'ajustement satisfait par les axes (*cf.* **Figure 3.6**).

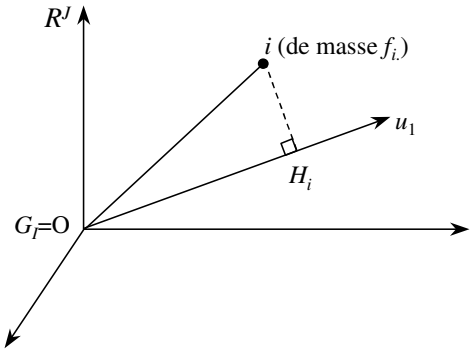



**Figure 3.6** Ajustement dans $R^J$ du nuage des profils-lignes. $i$ : point associé au profil-ligne $i$. $u_1$ : vecteur unitaire du premier axe factoriel. $H_i$ : projection de $i$ sur $u_1$. $u_1$ rend maximum $\sum_i f_{i.} \text{O}H_i^2$.

Du fait de l'introduction des poids $f_{i.}$ dans le critère d'ajustement, chaque modalité possède un poids proportionnel à la population qu'elle représente. Ainsi, à disparité de profil égale, les axes factoriels mettent plutôt en évidence des phénomènes concernant

une fraction importante de la population totale. Selon un autre point de vue, les modalités d'effectif faible, pour lesquelles les profils risquent d'être moins fiables, interviennent moins dans la construction des axes.

En résumé, l'ajustement du nuage $N_I$ en AFC est analogue à celui du nuage des individus en ACP. Il en diffère par trois points :

1. les lignes interviennent au travers de leur profil ;
2. la distance entre les profils est celle du $\chi^2$ ;
3. chaque ligne $i$ est affectée du poids $f_{i.}$ .

### 3.6.2 Ajustement du nuage des profils-colonnes.

Du fait du rôle symétrique joué par les lignes et les colonnes en AFC, l'ajustement de $N_J$ dans $R^I$ se pose dans les mêmes termes et possède les mêmes propriétés que l'ajustement de $N_I$ dans $R^J$. Nous les résumons ci-dessous.

1. Les images planes de $N_J$ doivent être telles que les distances entre les profils projetés ressemblent le plus possible aux distances entre les profils dans $R^I$. Il en résulte la nécessité d'analyser le nuage $N_J$ par rapport à son barycentre $G_J$. L'inertie totale de $N_J$ par rapport à $G_J$ provient des différences entre les profils des différentes classes $j$ et le profil de l'ensemble de la population.
2. Chaque colonne $j$ est affectée d'un poids égal à sa fréquence marginale $f_{.j}$. Avec des notations analogues à celles de la **figure 3.6**, $H_j$ étant la projection sur $v_1$ (vecteur unitaire du premier axe factoriel dans $R^I$) du point $j$ associé au profil-colonne $j$, $v_1$ rend maximum la quantité : $\sum_j f_{.j}(\mathrm{O}H_j)^2$. La justification de ce poids $f_{.j}$ est strictement analogue à celle développée à propos des profils-lignes.

### 3.6.3 Un aspect technique du centrage en AFC

Du point de vue technique, on peut montrer (*cf.* **section 5.5 page 121**) qu'il n'est pas nécessaire de centrer explicitement le nuage $N_I$ avant de l'analyser. En effet, mis à part le premier facteur, l'analyse du nuage par rapport à O sans centrage conduit aux mêmes facteurs que l'analyse du nuage centré.

Lorsque l'on réalise l'AFC du nuage $N_I$ non centré (c'est-à-dire par rapport à l'origine O sans centrage), le premier axe factoriel possède les propriétés suivantes (*cf.* **Figure 3.7**) :

1. il relie l'origine O au barycentre $G_I$ du nuage $N_I$ ;
2. cet axe est orthogonal, au sens de la distance utilisée (i.e. distance du $\chi^2$), à l'hyperplan $H_I$ contenant le nuage $N_I$;
3. l'inertie projetée de $N_I$ dans cette direction vaut 1.

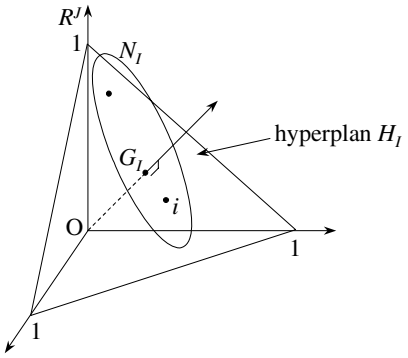



**Figure 3.7** Le premier axe factoriel du nuage $N_I$ non centré est le facteur trivial O$G_I$ orthogonal à $H_I$. L'inertie projetée de $N_I$ sur O$G_I$ vaut 1.

Naturellement, cet axe ne présente pas d'intérêt en lui-même : la projection sur O$G_I$ de chaque point de $N_I$ est confondue avec $G_I$. Cette projection de $N_I$ sur l'axe O$G_I$ est appelée **facteur trivial** ou facteur constant.

L'orthogonalité du premier axe O$G_I$ avec l'hyperplan $H_I$ présente une conséquence importante. Les axes suivants étant par définition orthogonaux à O$G_I$, l'analyse peut être poursuivie indifféremment par rapport à O ou à $G_I$ (*cf.* **Figure 3.8**).

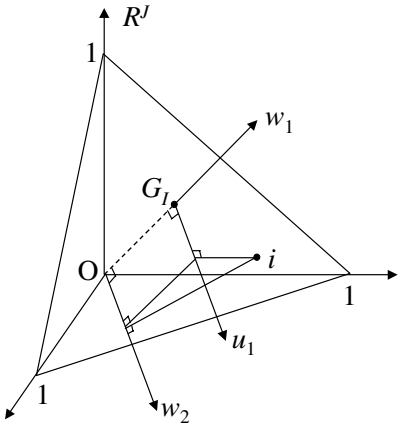



**Figure 3.8** Analyse par rapport au barycentre et par rapport à l'origine. $w_1$ : premier axe factoriel du nuage $N_I$ lorsque l'origine des axes est en O. $w_2$ : deuxième axe factoriel du nuage $N_I$ lorsque l'origine des axes est en O (orthogonal à $u_1$). $u_1$ : premier axe factoriel du nuage $N_I$ lorsque l'origine des axes est en $G_I$. Les projections de $N_I$ sur $w_2$ et $u_1$ sont identiques.

## 3.7 LA DUALITÉ

Les deux nuages $N_I$ et $N_J$ constituent deux représentations d'un même tableau, l'une à travers ses profils-lignes, l'autre à travers ses profils-colonnes. Il s'ensuit que les analyses de ces deux nuages ne sont pas indépendantes : les relations entre ces deux analyses sont communément regroupées sous le terme de dualité.

Cette dualité est plus fondamentale et plus riche en AFC qu'en ACP car les lignes et les colonnes représentent des objets de même nature, ce qui n'est pas le cas en ACP.

### 3.7.1 Statistique du $\chi^2$ et inertie des deux nuages $N_I$ et $N_J$

Lorsque l'on étudie un tableau de contingence, c'est-à-dire une population de $n$ individus au travers de deux variables qualitatives, il est classique de mesurer la significativité de la liaison entre ces deux variables à l'aide de la statistique $\chi^2$. Appliquée à un tableau d'effectifs, cette statistique mesure l'écart entre les effectifs observés et les effectifs théoriques que l'on obtiendrait en moyenne si les deux variables étaient indépendantes. Elle s'écrit :

$$\chi^2 = \sum_{ij} \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}} = \sum_{ij} \frac{(n\ f_{ij} - n\ f_{i.} f_{.j})^2}{n\ f_{i.} f_{.j}}$$

La statistique $\chi^2$ est égale, au coefficient $n$ près, à l'inertie totale par rapport à leur barycentre de l'un ou l'autre des nuages $N_I$ et $N_J$. En effet, dans $R^I$, l'inertie totale de $N_I$ par rapport à $G_I$ s'écrit :

$$\text{Inertie}(N_I) = \sum_i \text{Inertie}(i) = \sum_i f_{i.} d^2(i, G_I) = \sum_i f_{i.} \sum_j \frac{1}{f_{.j}} \left( \frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$$

Soit :

$$\chi^2 = n[\text{Inertie}(N_I)] = n[\text{Inertie}(N_J)]$$

Cette double égalité montre que l'inertie totale de chacun des deux nuages $N_I$ et $N_J$ représente, sous deux formes différentes, la liaison entre les deux variables.

> ***Remarque :*** La quantité $\chi^2/n$ , notée $\Phi^2$, mesure l'intensité de la liaison entre deux variables qualitatives (cette liaison est d'autant plus intense que les modalités de l'une s'associent exclusivement aux modalités de l'autre) et non sa significativité (elle ne dépend pas de l'effectif total) ; l'indicateur $\chi^2$ , lui, mesure la significativité (une liaison forte peut ne pas être significative si elle est observée sur très peu d'individus ; une liaison faible peut être significative si elle est observée sur beaucoup d'individus).

### 3.7.2 Dualité entre les facteurs sur *I* et les facteurs sur *J*

De même qu'en ACP, on appelle *facteur* l'ensemble des coordonnées des projections des points d'un nuage sur l'un de ses axes factoriels ; les facteurs sur les lignes sont les projections de $N_I$ et les facteurs sur les colonnes les projections de $N_J$. Le rang d'un facteur est le rang de l'axe factoriel correspondant. Outre leur inertie totale identique, les nuages $N_I$ et $N_J$ possèdent une propriété remarquable : leur ajustement conduit à deux suites de facteurs « duaux ». Plus précisément, nous montrons au chapitre 5 que :

1. les inerties associées aux axes de même rang dans chacun des nuages sont égales ;
2. les facteurs (de même rang) sur les lignes et ceux sur les colonnes sont liés par des relations dites de transition (elles permettent de transiter de $R^I$ dans $R^J$ et inversement).

Les deux paragraphes suivants détaillent cette dualité dont la conséquence essentielle est la suivante : les facteurs sur $I$ et sur $J$ de même rang doivent être interprétés conjointement car ils mettent en évidence la même part de liaison, exprimée pour l'un en termes de profils-lignes et pour l'autre en termes de profils-colonnes.

#### *a) Relations de transition*

Les formules de transition précisent les relations entre les points représentant d'une part les lignes et d'autre part les colonnes. Avec les notations suivantes :

1. $F_s(i)$ : projection de la ligne $i$ sur l'axe de rang $s$ de $N_I$,
2. $G_s(j)$ : projection de la colonne $j$ sur l'axe de rang $s$ de $N_J$,
3. $\lambda_s$ : valeur commune de l'inertie associée à chacun de ces deux axes,

les deux relations de transition s'écrivent :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{.j}} F_s(i)$$

Ces deux propriétés, qui expriment les résultats de l'analyse d'un nuage en fonction des résultats de l'analyse de l'autre nuage, conduisent à une économie de calcul. Mais surtout, elles donnent un sens à une représentation simultanée des lignes et des colonnes.

#### *b) Représentation simultanée des lignes et des colonnes ; relations barycentriques*

La représentation simultanée s'obtient en superposant les projections de chacun des deux nuages $N_I$ et $N_J$ sur des plans engendrés par des axes de même rang pour

les deux nuages. Sur les graphiques ainsi obtenus, les rapports entre la position des points lignes et des points colonnes dus aux relations de transition peuvent être décrits ainsi : au coefficient $1/\sqrt{\lambda_s}$ près, la projection, notée $F_s(i)$, de la ligne $i$ sur l'axe de rang $s$ (dans $R^J$) est le barycentre des projections, notées $G_s(j)$, des colonnes $j$ sur l'axe de rang $s$ (dans $R^I$), chaque colonne $j$ étant affectée du poids $f_{ij}/f_{i.}$ (cette expression d'une formule de transition est appelée propriété barycentrique). Les éléments « lourds » attirant le barycentre, une colonne $j$ attire d'autant plus une ligne $i$ que la valeur de $f_{ij}/f_{i.}$ est élevée. Sur les plans factoriels, les points éloignés de l'origine retiennent particulièrement l'attention car ce sont les profils les plus différents du profil moyen. On trouve donc, pour un facteur, du même côté qu'une ligne $i$ les colonnes $j$ auxquelles elle s'associe le plus et, à l'opposé, celles auxquelles elle s'associe le moins. Il est ainsi possible d'interpréter la position d'**une ligne** par rapport à **l'ensemble des colonnes**, ce qui justifie l'intérêt pratique de la représentation simultanée.

La formulation symétrique vaut, en inversant les rôles joués par les lignes et les colonnes. D'où le nom de double propriété barycentrique donnée à ce qui est **la principale règle d'interprétation des graphiques de l'AFC**. Cette double propriété est non seulement spécifique de l'AFC, mais la caractérise : on démontre que l'on retrouve les facteurs de l'AFC en cherchant à construire des fonctions définies sur les lignes et les colonnes d'un tableau de contingence telles que la double propriété barycentrique soit vérifiée.

La représentation simultanée en AFC est universellement adoptée, ce qui n'est pas le cas de celle de l'ACP. On peut citer deux arguments importants en faveur de cette superposition.

1. Alors qu'en ACP les lignes et les colonnes représentent des objets de nature bien différentes (individus et variables), les lignes et les colonnes, dans l'AFC d'un tableau de contingence, sont de même nature, à savoir des classes d'individus. Selon ce simple point de vue, cela ne pose aucun problème de figurer toutes ces classes sur un même graphique.

2. Il existe d'autres présentations de l'AFC dans lesquelles les classes d'individus que constituent les lignes et les colonnes d'un tableau de contingence sont situées dans un même espace : leur représentation simultanée est alors naturelle.

En résumé, sur les graphiques de la représentation simultanée des lignes et des colonnes, la position relative de deux points d'un même ensemble (lignes ou colonnes) s'interprète en tant que distance tandis que la position d'**un** point d'un ensemble par rapport à celle de **tous** les points de l'autre ensemble s'interprète en tant que barycentre. Toute association entre **une** ligne et **une** colonne suggérée par une proximité sur le graphique doit être contrôlée sur le tableau de données.

### 3.7.3 Interprétation de l'inertie des axes

L'inertie d'un point (ou d'un nuage de points) dans un espace euclidien se décompose sur toute base orthogonale : c'est la somme de ses inerties sur chacun des axes de cette base.

L'ajustement des nuages $N_I$ et $N_J$ décompose leur inertie selon des directions privilégiées : du fait de l'orthogonalité des axes, la somme des inerties d'un nuage sur chacun des axes est égale à l'inertie totale du nuage.

Contrairement au cas de l'ACP, dans laquelle l'inertie des nuages est égale au nombre de variables, cette inertie en AFC traduit la structure du tableau : l'inertie de chacun des deux nuages, des profils-lignes et des profils-colonnes, est égale à la statistique $\Phi^2$ (*cf.* section 3.7.1). L'AFC propose donc une décomposition de cette statistique et chaque facteur représente une part de la liaison entre les variables. L'inertie d'un facteur a donc une signification en absolu, et pas seulement en pourcentage de l'inertie totale du nuage : elle mesure en absolu l'importance de la part de liaison qu'il représente. Nous donnons l'interprétation des deux valeurs limites entre lesquelles elle se situe.

Lorsqu'un tableau vérifie les relations d'indépendance, les nuages sont concentrés en un point (leur barycentre) ; tous les profils-lignes sont identiques et égaux à la marge ligne $\{f_{.j}; j = 1, \ldots, J\}$ et tous les profils-colonnes sont identiques et égaux à la marge-colonne $\{f_{i.}; i = 1, \ldots, I\}$. L'inertie des nuages $N_I$ et $N_J$ relativement à leur centre de gravité est nulle et l'AFC ne donne aucun facteur (ou plutôt toute direction est associée à une inertie projetée nulle).

Il découle de la double propriété barycentrique que l'inertie associée à un axe factoriel vaut au maximum 1. Lorsque cette inertie vaut 1, l'axe factoriel met en évidence une situation « d'extrême dépendance » au sens suivant : l'ensemble des lignes d'une part, et celui des colonnes d'autre part, peuvent être divisés en au moins deux groupes, chaque groupe de lignes ne s'associant qu'à un groupe de colonnes (et réciproquement) selon le schéma de la **figure 3.9**. Dans ce cas, les facteurs définis par ces axes ont la même valeur pour chaque élément d'un même groupe de lignes ainsi que pour chaque élément du groupe de colonnes qui s'y associe. Une inertie proche de 1 indique que la structure du tableau est proche de cette situation limite : il existe une partition de $I$ et de $J$ telle que chaque classe de $I$ s'associe presque exclusivement à une classe de $J$ et réciproquement.

Lorsque deux axes factoriels ont une inertie égale à 1, les lignes d'une part et les colonnes d'autre part peuvent être divisées en au moins trois groupes qui ne s'associent qu'à un seul groupe de l'autre ensemble, etc. La situation de plus extrême dépendance entre deux variables qualitatives présentant le même nombre de modalités est celle où chaque modalité de l'une des variables ne s'associe qu'à l'une des modalités de l'autre. En ce cas, le tableau de contingence ne possède des effectifs non nuls que sur

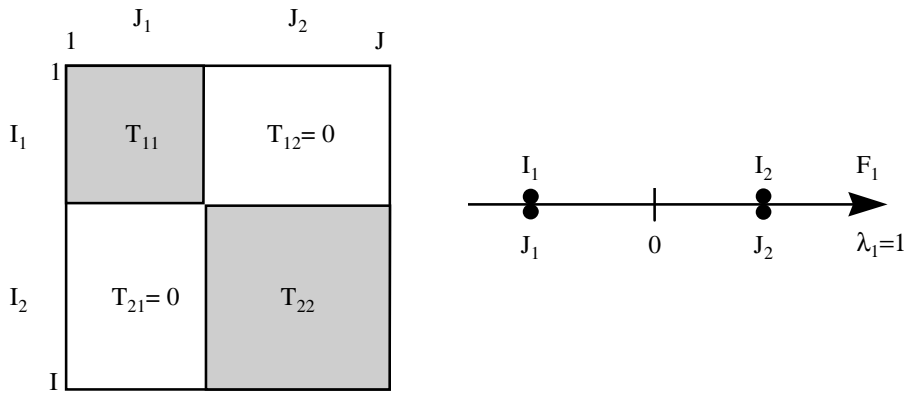



**Figure 3.9** Cas d'une inertie associée à un axe égale à 1. Partitions, des lignes d'une part et des colonnes d'autre part, mises en évidence par un axe factoriel associé à une inertie égale à 1. Tous les effectifs des sous-tableaux $T_{12}$ et $T_{21}$ sont nuls.

la diagonale. Il résulte de ce qui précède que, dans ce cas, chaque axe de l'AFC est associé à une inertie de 1.

### 3.7.4 Formule de reconstitution des données

À la décomposition de l'inertie, on peut associer une décomposition du tableau lui-même. En effet, on peut montrer (*cf.* section 5.6) que :

$$f_{ij} - f_{i.}f_{.j} = f_{i.}f_{.j}\sum_{s} F_s(i)G_s(j)/\sqrt{\lambda_s}$$

Cette formule, appelée formule de reconstitution des données, permet de recalculer les valeurs du tableau initial en fonction des marges et des facteurs. Lorsque l'on dépouille les résultats d'une AFC, on limite généralement l'interprétation aux premiers facteurs. Cela revient à considérer non pas le tableau des données mais son approximation obtenue à l'aide des premiers termes de la somme ci-dessus.

Cette relation met en évidence une décomposition de l'écart du tableau relativement à l'hypothèse d'indépendance en une somme de tableaux dont chacun ne dépend que d'un couple de facteurs $(F_s, G_s)$ de même rang. Elle formalise l'aspect de l'objectif annoncé : décomposition de la liaison en éléments simples. En effet, chaque tableau de terme général $f_{i.}f_{.j}F_s(i)G_s(j)$ exprime une liaison simple puisque le terme de la case $(i, j)$ ne dépend que de la ligne $i$ et de la colonne $j$. Si les valeurs de $F_s(i)$ et de $G_s(j)$ sont de même signe, cette case exprime une attirance entre $i$ et $j$ ; dans le cas contraire, il exprime une répulsion d'autant plus importante que $F_s(i)$ et $G_s(j)$ sont grands en valeur absolue.

Nous illustrons cette décomposition dans la section 10.3.1.a, page 231, à propos d'un exemple.

## 3.8 NOMBRE D'AXES ET INERTIE TOTALE

Dans l'espace $R^J$, le nuage $N_I$ est contenu dans un sous-espace de dimension $J-1$ ; dans cet espace, on peut donc trouver au maximum $J-1$ dimensions orthogonales d'inertie non nulle. De même, dans l'espace $R^I$ , on peut trouver au maximum $I-1$ dimensions orthogonales d'inertie non nulle. Compte tenu de la dualité (même inertie sur les axes de même rang dans les deux espaces), en AFC on peut trouver au maximum $\min\{I-1, J-1\}$ axes d'inertie non nulle.

L'inertie associée à un axe étant au maximum égale à 1, l'inertie totale en AFC est donc comprise entre 0 (indépendance) et $\min\{I-1, J-1\}$ (liaison d'intensité maximum = association stricte entre les modalités des deux variables mises en correspondances).

## 3.9 AIDES À L'INTERPRÉTATION ET ÉLÉMENTS SUPPLÉMENTAIRES

Les indices d'aide à l'interprétation (qualité de représentation d'un élément par un axe ou un plan et contribution d'un élément à l'inertie d'un axe) définis en ACP (*cf.* section 1.9) sont valables pour un nuage quelconque. Ils s'appliquent donc en AFC. Notons que, si en ACP les poids de tous les éléments sont en général égaux, ce n'est pas le cas en AFC ; or ces poids interviennent dans la contribution d'un point à l'inertie d'un axe.

En AFC, comme en ACP, on utilise presque systématiquement la technique des éléments supplémentaires, qui consiste à projeter sur les axes factoriels des profils de lignes ou de colonnes qui n'interviennent pas dans le calcul de ces axes. Une ligne supplémentaire est reliée aux colonnes actives par la formule barycentrique. De même, une colonne supplémentaire est reliée aux lignes actives par la formule barycentrique. Ces éléments servent très souvent, eux aussi, d'aides à l'interprétation ; dans les tableaux de grande dimension, par exemple, il est très pratique de connaître la position et la qualité de représentation du barycentre de plusieurs lignes ou de plusieurs colonnes.

## 3.10 SCHÉMA GÉNÉRAL DE L'AFC

Nous résumons les principaux résultats de l'AFC dans un schéma général (*cf.* **Figure 3.10**). Les numéros ci-dessous renvoient à ce schéma.

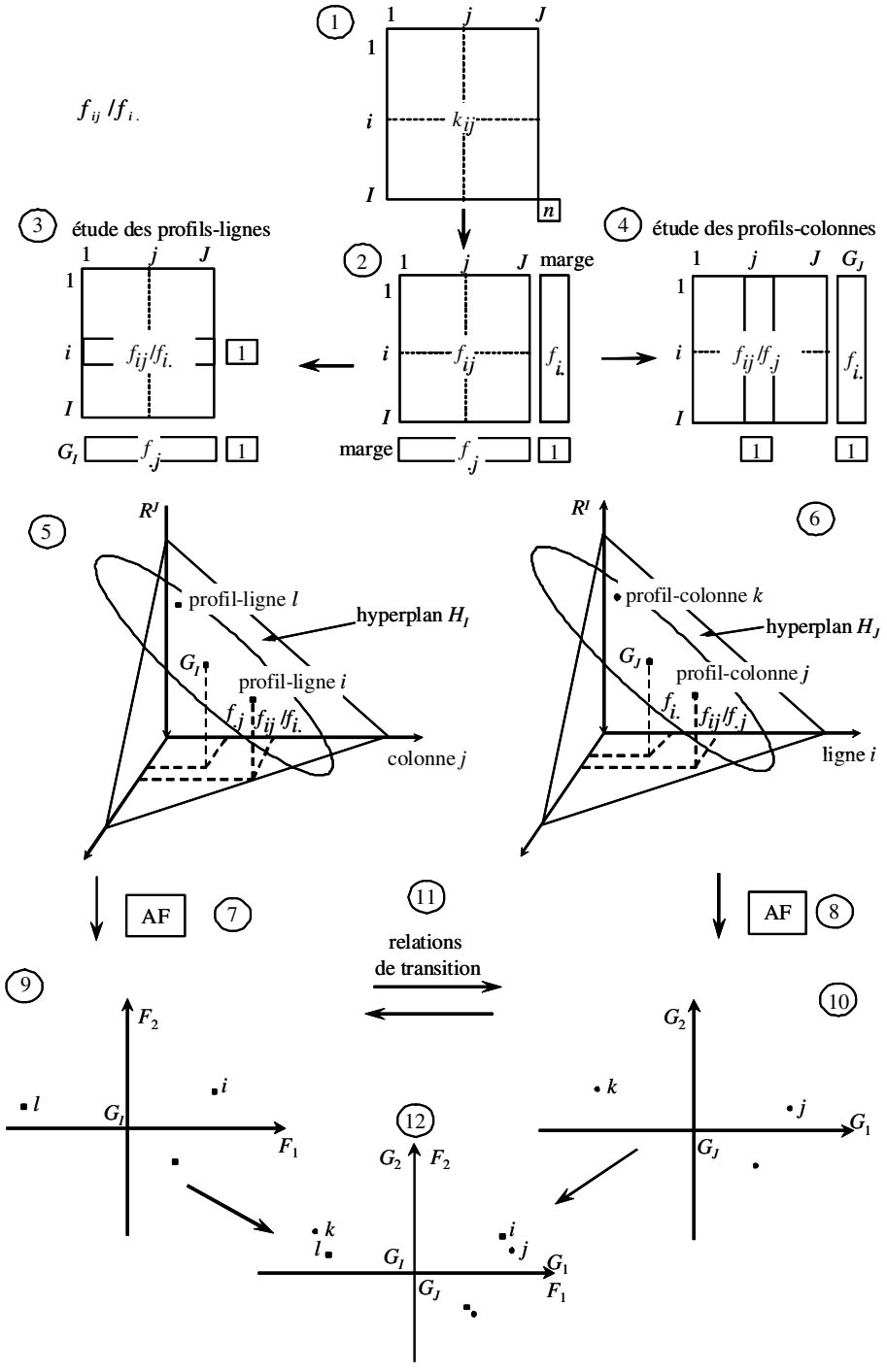



**Figure 3.10** Schéma général de l'AFC.

**1.** Les données brutes. Lignes et colonnes jouent des rôles symétriques : ce sont des modalités de variables. La somme de tous les termes $k_{ij}$ du tableau est $n$.

**2.** Ce tableau intermédiaire fait apparaître les données sous forme de loi de probabilité : $f_{ij} = k_{ij}/n$. Les probabilités marginales sont $\{f_{i.} | i \in I\}$ et $\{f_{.j} | j \in J\}$.

**3.** et **4.** Pour étudier les lignes du tableau, on les transforme en profils-lignes. Pour étudier les colonnes, on les transforme en profils-colonnes. On dispose donc de deux tableaux. Un profil s'interprète comme une probabilité conditionnelle. Les profils moyens $G_I$ et $G_J$ sont les distributions marginales associées au tableau **2**.

**5.** Un profil-ligne est une suite de $J$ nombres et peut être représenté par un point de $R^J$. Le nuage $N_I$ des profils-lignes appartient à l'hyperplan $H_I$ des vecteurs dont la somme des coordonnées vaut 1. Chaque profil-ligne $i$ est affecté du poids $f_{i.}$ ; le nuage $N_I$ ainsi pondéré a pour barycentre le profil moyen $G_I$. Dans le nuage $N_I$, on s'intéresse à la ressemblance entre les profils mesurée au travers de la distance du $\chi^2$.

**6.** La représentation des profils-colonnes dans $R^I$ appelle des commentaires strictement symétriques à ceux de la représentation des profils-lignes dans $R^J$.

**7.** L'Analyse Factorielle (AF) d'un nuage consiste à mettre en évidence une suite de directions orthogonales telles que l'inertie, par rapport à O, de la projection du nuage sur ces directions est maximum. Appliquée à $N_I$, l'AF fournit une première direction – dite triviale – reliant O à $G_I$ et orthogonale à $H_I$. Pour les directions suivantes, $G_I$ se projette à l'origine des axes : ces directions suivantes sont les directions d'allongement maximum de $N_I$. Il est donc équivalent de réaliser l'analyse par rapport à O ou par rapport à $G_I$.

**8**. On peut reprendre point par point le commentaire de **7** en le transposant aux colonnes.

**9.** et **10.** Les plans factoriels, croisant deux facteurs, sur les lignes ou sur les colonnes, fournissent des images approchées des nuages $N_I$ et $N_J$. Sur ces plans, la distance entre deux points s'interprète comme une ressemblance entre les profils de ces points. L'origine des axes est confondue avec le profil moyen.

**11.** Les relations de transition expriment les résultats d'une AF (par exemple dans $R^I$) en fonction des résultats de l'autre (par exemple dans $R^J$).

**12.** Du fait des relations de transition, les interprétations des plans factoriels représentant $N_I$ et $N_J$ doivent être menées simultanément. Il est commode de superposer ces représentations. L'interprétation de cette représentation simultanée est régie par la double propriété barycentrique.

## 3.11 CONCLUSION

Dans ce chapitre, l'AFC est introduite comme une méthode particulièrement bien adaptée à l'étude d'un tableau de contingence. D'un point de vue historique, elle a d'ailleurs été imaginée pour traiter ce type de tableau. Toutefois, les remarquables propriétés de cette méthode ont très tôt incité à l'appliquer à d'autres tableaux : aujourd'hui, la pratique courante de l'AFC dépasse largement le cadre des tableaux de contingence.

Dès l'instant que l'on étudie un tableau qui n'est pas un tableau de contingence, l'objectif de l'AFC ne peut plus être formulé en terme de liaison entre deux variables qualitatives. En revanche, il existe des tableaux dont l'étude nécessite une typologie des lignes d'une part et des colonnes d'autre part, à travers leur profil.

Pour établir l'intérêt de l'AFC dans la réalisation de telles typologies, il convient de s'assurer que les différentes notions mises en jeu par cette méthode (transformation en profils, distance du $\chi^2$ , poids des éléments) sont en accord avec le point de vue que l'on veut avoir sur les données étudiées. Les formules barycentriques, qui relient les projections des lignes et des colonnes et qui permettent à elles seules de caractériser les facteurs, peuvent aussi justifier l'application de l'AFC.

Nous illustrons ces situations à l'aide de deux exemples.

**Premier exemple** : Dans l'étude de la liaison entre le diplôme obtenu et l'emploi occupé, on peut s'intéresser à deux tableaux de même structure établis l'un en se limitant aux hommes et l'autre en se limitant aux femmes. Le chapitre 10 propose une série d'analyses pour ce couple de tableaux. Dès maintenant, on peut se rendre compte de l'intérêt de l'AFC sur une juxtaposition « en ligne » de plusieurs tableaux. En réalité, ce tableau est encore un tableau de contingence dont l'une des deux variables est obtenue par croisement des deux variables *emploi* et *sexe*.

**Second exemple** : Les lignes sont les entreprises d'un secteur économique. Les colonnes sont les postes d'actif du bilan. À l'intersection de la ligne $i$ et de la colonne $j$, se trouve la valeur du poste $j$ pour l'entreprise $i$. Un tel tableau peut être analysé à l'aide d'une ACP. En ce cas, les postes sont des variables centrées et réduites ; chaque poste est affecté du même poids ainsi que chaque entreprise. Généralement, les entreprises diffèrent assez sensiblement par leur total d'actif, ce qui induit presque toujours un effet taille en tant que premier facteur (*cf.* section 1.6).

Mais ce tableau peut aussi être analysé à l'aide d'une AFC. Tout d'abord, ses marges (qui servent de référence) ont une signification claire : la somme des termes de la $i^{\text{e}}$ ligne est le total des actifs de l'entreprise $i$ ; la somme des termes de la $j^{\text{e}}$ colonne est la valeur du poste $j$ pour l'entreprise fictive que constitue l'ensemble du secteur. Sans entrer dans les détails, les principales caractéristiques impliquées par l'AFC de ce tableau sont les suivantes.

1. Chaque entreprise est analysée au travers de son profil : chacun de ses postes est exprimé par rapport au total des actifs. Un éventuel effet taille est éliminé.
2. Chaque entreprise a un poids proportionnel à son total d'actif.
3. Chaque poste de bilan a un poids proportionnel à son importance pour l'ensemble du secteur.
4. Les postes du bilan sont transformés en profil ; cette harmonisation des données n'est pas très différente du couple centrage-réduction en ACP. À la différence de l'ACP, le nuage des postes est analysé à partir de son barycentre : on étudie les différences entre postes. Ce qui est commun à l'ensemble des postes est éliminé : on ne peut observer d'effet taille.

Ce second exemple montre que certains tableaux peuvent être analysés par ACP ou AFC. Ces deux analyses ne sont pas équivalentes et peuvent fournir des éclairages assez différents. On examinera les pondérations induites par l'AFC aussi bien pour choisir entre les deux méthodes que pour interpréter conjointement leurs résultats.

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.

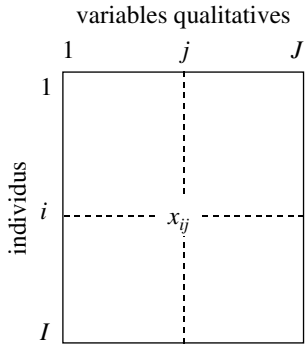



**Figure 4.1** Tableau des données sous forme de codage condensé. $I$ : nombre et ensemble des individus. $J$ : nombre et ensemble des variables qualitatives. $x_{ij}$ : codage condensé de la valeur de l'individu $i$ pour la variable $j$ (numéro ou chaîne de caractère).

### 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.

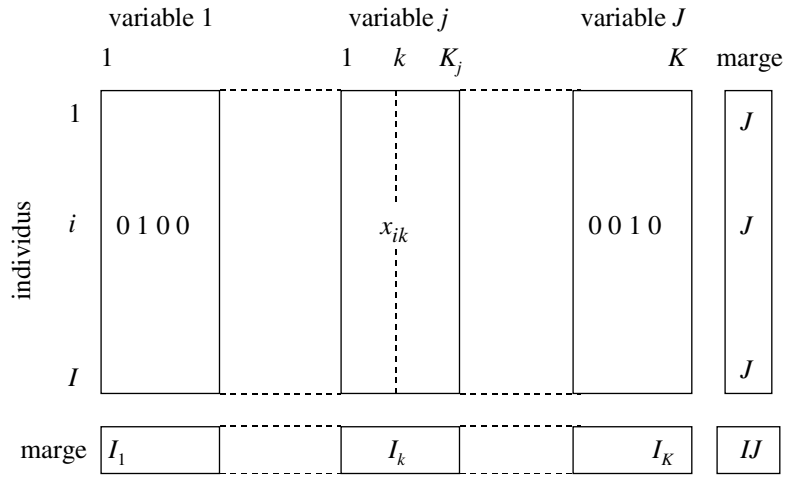



**Figure 4.2** Tableau des données sous forme disjonctive complète. $K_j$ = nombre et ensemble des modalités de la variable $j$. $K = \sum_{j=1}^{j=J} K_j$ = nombre et ensemble des modalités toutes variables confondues. $x_{ik} = 1$ si l'individu $i$ possède la modalité $k$ et 0 sinon $\sum_{k=1}^{k=K_j}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

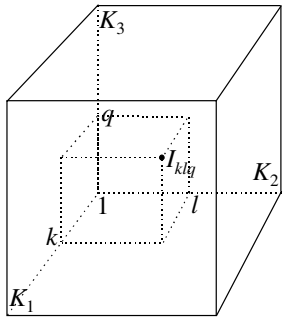



**Figure 4.3** L'hypertableau de contingence associé à 3 variables qualitatives. $K_1$ : nombre de modalités de la première variable. $I_{klq}$ : nombre d'individus possédant les modalités $k$ (de la variable 1), l (de la variable 2) et $q$ (de la variable 3).

### 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.

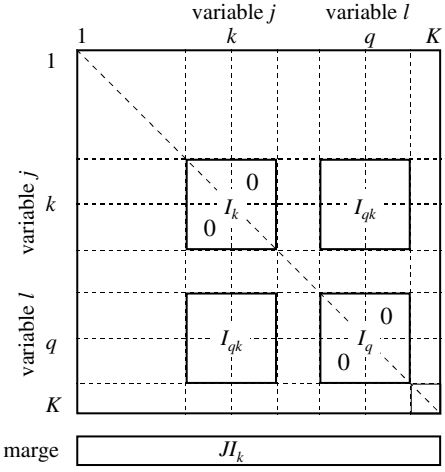



**Figure 4.4** Tableau de Burt. Le tableau est symétrique. Les tableaux $J$ situés sur la diagonale sont diagonaux. $I_{qk}$ : nombre d'individus possédant à la fois la modalité $q$ (de la variable l) et la modalité $k$ (de la variable $j$). $I_k$ : nombre d'individus possédant la modalité $k$ (de la 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*×*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 : $d^2(k, G_K) = (I/I_k) - 1$ ; 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 :

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

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{x_{ik}}{I_k} F_s(i)$$

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.

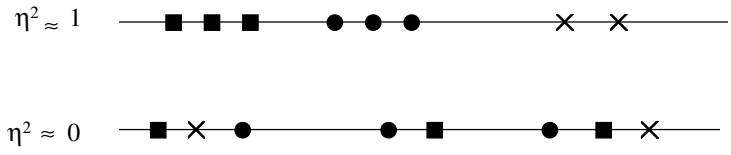



**Figure 4.5** Illustration des deux valeurs extrêmes du rapport de corrélation. 8 individus, représentés par un symbole différent selon leur modalité pour une variable qualitative, figurent sur un axe représentant une variable numérique.

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.

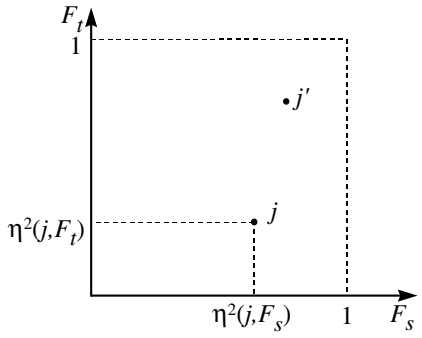



**Figure 4.6** Représentation des variables en ACM (carré des liaisons). $\eta^2(j, F_s)$ : rapport de corrélation entre la variable qualitative $j$ et le facteur $F_s$. Par construction, pour tout $j$ et tout $s$ : $0 \leqslant \eta^2(j, F_s) \leqslant 1$. Ce graphique montre que les variables $j$ et $j'$ 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 $k^{\text{e}}$ 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 $\chi^2$

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 $\chi^2$ 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 $\chi^2$ 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 $\chi^2$. En ce sens, cette AFC est une étude simultanée des liaisons binaires.

Dans cette somme de $\chi^2$, 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 $\chi^2$ 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 $\eta^2$ 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.

Chapitre **5**

# Calculs et dualité en Analyse Factorielle

## 5.1 INTRODUCTION

Les méthodes d'analyse factorielle présentées dans les premiers chapitres sont fondées sur des principes communs : à partir d'un tableau de données, on construit deux nuages de points représentant respectivement les lignes et les colonnes ; ces deux nuages sont projetés chacun sur une suite d'axes orthogonaux maximisant l'inertie projetée ; sur chacun de ces axes, les deux nuages ont la même inertie projetée et les projections des points sont liées d'un nuage à l'autre par les relations dites de transition.

Dans ce chapitre, nous indiquons comment calculer ces facteurs, montrons la dualité des deux nuages et donnons des démonstrations des formules de transition. Le cadre dans lequel nous nous plaçons est assez général. Non seulement il recouvre l'ACP et l'AFC, mais il permet d'introduire et de calculer les facteurs d'analyses factorielles fondées sur d'autres distances et d'autres poids.

## 5.2 CALCUL DES AXES D'INERTIE ET DES FACTEURS D'UN NUAGE DE POINTS

Le problème est posé en ces termes : étant donné un nuage de $I$ points noté $N_I$ dans un espace euclidien de dimension $J$, on cherche une suite d'axes orthonormés (pour la métrique de l'espace) telle que l'inertie du nuage projeté sur ces axes soit maximum.

L'ensemble des coordonnées des $I$ points du nuage sur un de ces axes définit une fonction numérique sur $I$, appelée facteur sur $I$. Dans les résultats d'une analyse, seuls les facteurs apparaissent, les axes n'étant que des intermédiaires de calcul. Pour

obtenir les facteurs et leur inertie, nous utilisons des techniques simples de calcul matriciel.

### 5.2.1 Notations : les matrices *X*, *M* et *D*

Les coordonnées $x_{ij}$ des $I$ points du nuage $N_I$ dans l'espace $R^J$ forment un tableau, ou une matrice, de dimensions $(I, J)$, notée $X$. L'espace $R^J$ est muni d'une métrique euclidienne qui peut être différente de la métrique canonique (ou usuelle). Cette métrique dérive d'un produit scalaire dont la matrice, de dimensions $(J, J)$, est notée $M$. Nous nous restreignons à des métriques associées à des matrices diagonales car elles seules sont facilement interprétables en termes de données initiales. En effet, lorsque $M$ est diagonale, la distance $d_M$ entre deux points $i$ et $l$ de $N_I$ s'écrit, en notant $m_j$ les éléments diagonaux de $M$ :

$$d_M^2(i,l) = \sum_j (x_{ij} - x_{lj})^2 m_j$$

Les coefficients $m_j$ pondèrent l'influence de chaque colonne $j$ dans les distances entre éléments ; cette propriété justifie leur nom de « poids des colonnes ». Or, lorsque $M$ n'est pas diagonale, ses termes apparaissent comme des poids associés à des couples de colonnes, ce qui n'a pas de résonance concrète.

Le produit scalaire (associé à $d_M$) entre deux vecteurs $u$ et $v$ s'écrit :

$$\langle u, v \rangle_M = u'Mv = v'Mu$$

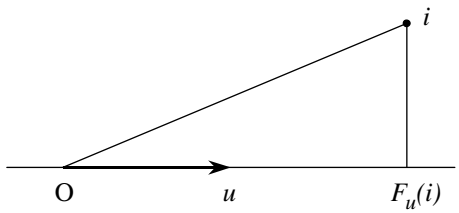
où $u'$ et $v'$ désignent les transposés des vecteurs colonnes $u$ et $v$.

Les coordonnées des points de $N_I$ et la métrique de l'espace $R^J$ définissent entièrement la forme du nuage mais, dans le calcul des axes d'inertie, le poids des points de $N_I$ intervient. Ces poids, notés $p_i$, sont rangés dans une matrice diagonale, de dimension $I$, notée $D$. Toute l'information nécessaire pour calculer les facteurs est contenue dans les trois matrices $X$, $M$, $D$.

**Matrice et application linéaire**. Dans ce chapitre, nous serons conduit à considérer l'application linéaire associée à une matrice ; nous utilisons la même notation pour ces deux objets. Nous précisons qu'il s'agit d'un endomorphisme lorsque l'application associe à un vecteur d'un espace vectoriel E un autre vecteur de cet espace.

### 5.2.2 Projection d'un nuage sur un axe

Soit $u$ un vecteur unitaire (pour la métrique $M$, *i.e.* vérifiant $u'Mu = 1$) d'un axe quelconque de $R^J$. L'ensemble des coordonnées des projections des $I$ points du nuage $N_I$ sur l'axe $u$ constitue un vecteur de dimension $I$, que nous notons $F_u$. Pour tout




$$\begin{aligned} F_u(i) &= \langle x_i, u \rangle_M = x_i' M u \\ F_u &= X M_u \\ \sum_i p_i [F_u(i)]^2 &= F_u' D F_u \end{aligned}$$

**Figure 5.1** Projection $F_u(i)$ du point $i$ sur l'axe défini par le vecteur unitaire $u$.

point $i$ du nuage $N_I$, $F_u(i) = x_i' M u$ où $x_i$ est le vecteur de $R^J$ dont les coordonnées sont celles de $i$ : $x_i'$ n'est autre que la ligne $i$ de la matrice $X$. De cette égalité, on déduit la relation matricielle $F_u = XMu$.

### 5.2.3 Inertie du nuage projeté

L'inertie du nuage projeté sur $u$ est égale à $\sum_i p_i [F_u(i)]^2$. Cette quantité s'écrit matriciellement en fonction de la matrice diagonale $D$ et du vecteur $F_u$ sous la forme $F_u' D F_u$. Comme $F_u = XMu$, l'inertie vaut $u'MX'DXMu$.

Chercher un axe de $R^J$ tel que l'inertie du nuage projeté soit maximum revient donc à chercher un vecteur $u$, unitaire pour la métrique $M$ (*i.e.* $u'Mu = 1$), rendant maximum la quantité $u'MX'DXMu$.

### 5.2.4 Calcul des axes d'inertie maximum ; cas de la métrique identité

Lorsque le produit scalaire sur $R^J$ est le produit scalaire canonique, la matrice $M$ est la matrice identité et l'écriture des expressions ci-dessus s'allège : on cherche un vecteur $u$, vérifiant $u'u = 1$ et rendant maximum $u'X'DXu$.

La matrice $X'DX$ est symétrique, donc diagonalisable, et ses vecteurs propres forment une base orthonormée de $R^J$. Soient $\lambda_1, ..., \lambda_s, ..., \lambda_J$ les valeurs propres de $X'DX$ rangées par ordre décroissant et $\{e_s ; s = 1, ..., J\}$ une base orthonormée de vecteurs propres associés ($X'DXe_s = \lambda_s e_s$). Décomposons le vecteur $u$ sur cette base. On a :

$$\begin{aligned} u &= \sum_s u_s e_s \text{ avec } \sum_s u_s^2 = 1 \\ X'DXu &= X'DX \sum_s u_s e_s = \sum_s \lambda_s u_s e_s \end{aligned}$$

L'inertie projetée sur $u$ s'écrit donc :

$$u'X'DXu = \sum_s \lambda_s u_s^2 \leqslant \lambda_1 \sum_s u_s^2 = \lambda_1$$

Ainsi, avec la contrainte $\sum_s u_s^2 = 1$, cette inertie est majorée par $\lambda_1$. Ce maximum est atteint lorsque la première composante $u_1$ de $u$ vaut 1 ou $-1$ et que les autres sont nulles c'est-à-dire lorsque $u = \pm e_1$. L'inertie du nuage projeté sur un axe est donc maximum lorsque cet axe est colinéaire aux vecteurs propres de $X'DX$ associés à sa plus grande valeur propre $\lambda_1$ . Elle vaut alors $\lambda_1$ .

Les vecteurs propres de la matrice symétrique $X'DX$ étant orthogonaux deux à deux, le même raisonnement montre que la direction orthogonale à $u_1$ qui maximise l'inertie du nuage projeté est celle d'un vecteur propre associé à la deuxième valeur propre $\lambda_2$ de $X'DX$ ; cette inertie vaut alors $\lambda_2$. La suite d'axes orthogonaux maximisant l'inertie projetée est donc définie par une suite de vecteurs propres de $X'DX$ rangés par valeurs propres décroissantes (les valeurs propres sont supposées distinctes ce qui toujours le cas en pratique).

### 5.2.5 Calcul des axes d'inertie maximum pour une métrique quelconque

Si la métrique $M$ n'est pas la métrique identité, le raisonnement ci-dessus s'applique sans changement majeur. En effet, $X'DXM$ définit un endormophisme de $R^J$ symétrique pour la métrique $M$. Rappelons que la $M$-symétrie d'un endomorphisme $A$ est définie par l'égalité, pour tout couple de vecteurs $u$ et $v$, des deux expressions :

$$\langle u, Av\rangle_M = \langle Au, v\rangle_M$$

Matriciellement : $A'M = MA$ ; on retrouve la notion usuelle de matrice symétrique si $M$ est la matrice identité. Cette égalité est vérifiée pour $X'DXM$ :

$$\langle u, X'DXMv\rangle_M = u'MX'DXMv = \langle X'DXMu, v\rangle_M$$

L'endomorphisme $X'DXM$, étant $M$-symétrique, est diagonalisable et admet une base $M$-orthonormée de vecteurs propres. Comme au paragraphe précédent, la décomposition d'un vecteur $u$ quelconque sur cette base montre que la solution est donnée par les vecteurs propres de $X'DXM$ rangés par valeurs propres décroissantes.

### 5.2.6 Calcul des facteurs et de leur inertie

Notons $F_s$ le facteur de rang $s$ défini par la projection du nuage sur le $s^{\text{e}}$ axe d'inertie. Pour calculer les facteurs $F_s$, on peut diagonaliser la matrice $X'DXM$, calculer une

suite de vecteurs propres $M$-normés $u_s$ associés aux valeurs propres $\lambda_s$, et appliquer aux vecteurs $u_s$ la matrice $XM$, soit $F_s = XMu_s$.

Il est possible aussi d'obtenir directement les facteurs $F_s$ et leur inertie en diagonalisant la matrice $XMX'D$ de dimension $I$. En effet, les égalités ci-dessous montrent que si $u_s$ est vecteur propre de $X'DXM$ associé à $\lambda_s$ , alors $F_s = XMu_s$ est un vecteur propre de $XMX'D$ associé à la même valeur propre $\lambda_s$ :

$$\begin{aligned} X'DXMu_s &= \lambda_s u_s \\ (XM)(X'DXMu_s) &= \lambda_s (XM)u_s \\ XMX'DF_s &= \lambda_s F_s \end{aligned}$$

L'inertie du nuage $N_I$ projetée sur $u_s$ est la somme des carrés des termes de $F_s$ pondérés par les poids des éléments $i$ soit :

$$\sum_i p_i F_s(i)^2 = F_s' D F_s = \lambda_s$$

### 5.2.7 Définition du nuage des colonnes de X

Le nuage des colonnes $N_J$ comprend $J$ points situés dans un espace de dimension $I$, noté $R^I$. Les coordonnées $x_{ij}$ de ces points sont contenues dans les colonnes de $X$ (qui sont d'ailleurs les lignes de la transposée $X'$ ). Pour qu'il y ait dualité entre le nuage des lignes $N_I$ et le nuage des colonnes $N_J$, il est nécessaire que ces deux nuages représentent la même information et soient construits de façon symétrique. Tout d'abord, il est logique d'affecter à chaque colonne $j$ le poids $m_j$ (terme général de la matrice $M$ déjà interprété comme un poids de colonne : *cf.* section 5.2.1 ; rappelons que nous nous sommes limités aux métriques associées à une matrice diagonale). Ainsi, le choix des poids des éléments du nuage $N_J$ et le choix de la métrique dans $R^J$ sont liés.

En outre, la construction symétrique des deux nuages implique que le poids des individus du nuage $N_I$ induise la métrique dont $R^I$ est muni. De façon directe, on peut remarquer qu'il revient au même de dupliquer un élément $i$ ou de doubler son poids. Dans $R^I$, la distance entre deux points est la même dans ces deux cas à condition d'adopter la métrique $D$.

Le **tableau 5.1 page suivante** résume les poids et les métriques mis en jeu. Les nuages $N_I$ et $N_J$ ainsi construits sont dits duaux en ce sens qu'ils représentent tous deux les mêmes données $\{X, M, D\}$.

**Tableau 5.1** Les deux nuages duaux

| | Espace | Métrique | Poids | Coordonnées du point $k$ |
|---|---|---|---|---|
| Nuage des lignes $N_I$ | $R^J$ | $M$ | $D$ | $k^{\text{e}}$ ligne de $X$ |
| Nuage des colonnes $N_J$ | $R^I$ | $D$ | $M$ | $k^{\text{e}}$ ligne de $X'$ |

## 5.3 NUAGES DES LIGNES ET DES COLONNES EN ACP ET EN AFC

Le cadre général choisi pour démontrer les principaux résultats d'analyse factorielle suppose que l'on peut définir de manière totalement symétrique, à partir du triplet $\{X, M, D\}$, le nuage des lignes et celui des colonnes. En Analyse des Correspondances, comme en Analyse en Composantes Principales, il est possible de calculer des matrices $\{X, M, D\}$ permettant cette construction symétrique. Nous les précisons dans les paragraphes suivants.

### 5.3.1 Matrices X, M, D en ACP

En Analyse en Composantes Principales, la matrice $X$ est le tableau des données centrées et généralement réduites. Dans certains cas assez rares, on souhaite conserver l'échelle de chaque variable : la matrice $X$ est alors la matrice des variables centrées non réduites.

La matrice diagonale $D$ contient les poids des individus. Dans la plupart des cas, tous les individus ont le même poids $1/I$ mais il est possible de leur affecter des poids différents. Notons que les poids $p_i$ des individus doivent avoir pour somme 1 afin que le cosinus dans $R^I$ traduise exactement la corrélation.

Les variables ont presque toujours un poids égal à 1, mais il est possible, là encore, de modifier ces poids pour moduler l'influence respective des variables.

Si l'on ne centre pas les variables, l'analyse factorielle est techniquement possible : ses résultats s'interprètent alors comme les projections duales du nuage des lignes et du nuage des colonnes mais il ne s'agit plus alors véritablement d'une ACP en ce sens qu'elle n'a pas les mêmes propriétés. Ainsi, c'est le centrage qui permet d'interpréter les axes factoriels dans $R^J$ comme les directions de plus grande variabilité de $N_I$ ; en l'absence de centrage, ces axes sont influencés non seulement par la forme du nuage $N_I$ mais aussi par sa position par rapport à l'origine. Par ailleurs, le centrage permet d'interpréter le cosinus de l'angle entre deux vecteurs représentant deux colonnes dans $R^I$ comme un coefficient de corrélation.

Remarquons que la matrice $X'DX$ est, dans le cas de données centrées-réduites, la matrice des corrélations (et la matrice des covariances lorsque les données sont seulement centrées). Le calcul des axes factoriels ne dépendant des données $X$ qu'au

travers de cette matrice, il apparaît clairement ici que ces axes ne dépendent que des liaisons linéaires entre variables.

### 5.3.2 Matrices X, M, D en AFC

La présentation de l'AFC (chapitre 3) met l'accent sur l'analyse des nuages des profils des lignes et des colonnes du tableau de données. Ainsi, les deux matrices qui contiennent les coordonnées des profils des lignes et des colonnes du tableau de données correspondent à deux transformations différentes de ce tableau et, d'autre part, les métriques employées dans l'analyse d'un nuage ne sont pas les poids de l'autre nuage mais leur inverse. Cela laisserait à penser que l'AFC n'entre pas dans le cadre général défini au début de ce chapitre. Nous introduisons ici une autre définition de l'AFC avec des matrices $X$, $M$, $D$ qui respectent les conditions de la section 5.2.7.

#### *a) Une autre définition de l'AFC*

En reprenant les notations du chapitre 3, le terme général de la matrice $X$ s'écrit :

$$x_{ij} = \frac{f_{ij}}{f_{i.}f_{.j}} - 1 = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}}$$

Cette matrice contient les écarts (rapportés au produit $f_{i.}f_{.j}$) entre le tableau des données $f_{ij}$ et le tableau de terme général $f_{i.}f_{.j}$ qui correspond à l'hypothèse d'indépendance. Cette présentation des données correspond bien aux objectifs de l'AFC décrits au chapitre 3.

Les matrices $M$ et $D$ sont diagonales de coefficients $f_{.j}$ et $f_{i.}$ respectivement.

Les poids des lignes sont donc égaux aux $f_{i.}$ et ceux des colonnes sont égaux aux $f_{.j}$.

#### *b) Équivalence entre les deux définitions*

Pour montrer qu'avec ces matrices on obtient les résultats de l'AFC présentée au chapitre 3, il faut montrer que les nuages de lignes et de colonnes obtenus par les deux approches sont isomorphes. Le nuage des lignes de $X$ est, comme le nuage des profils-lignes du tableau de données, situé dans un espace de dimension $J$. Les coordonnées des points sont différentes et les deux espaces ne sont pas munis de la même métrique. L'un est muni de la métrique $M$ et l'autre de la métrique du $\chi^2$ qui n'est autre que l'inverse $M^{-1}$ de $M$.

On peut vérifier directement que les distances entre les couples de points homologues sont les mêmes. Mais cette égalité découle d'un isomorphisme induit par $M$ que l'on peut utiliser dans toute analyse factorielle et qui a une signification intéressante en AFC. En effet, la métrique $M$ de l'espace $R^J$ définit un isomorphisme de $R^J$ dans son dual noté $R^{J*}$. Si l'on munit $R^{J*}$de la métrique $M^{-1}$, l'application $M$ est

un isomorphisme d'espaces euclidiens : les distances et les formes sont conservées ($\langle u, v\rangle_M = \langle Mu, Mv\rangle_{M^{-1}}$).

Le dual, noté $E^*$, d'un espace vectoriel $E$ est l'espace des formes linéaires : $f : E \to R$. Projeter, dans $E$ et au sens de la métrique $M$, le vecteur $v$ sur $u$ revient à appliquer à $v$ la forme linéaire $Mu$. Cette forme linéaire est l'élément de $E^*$ associé à $u$ de $E$ par l'application $M$. La figure 5.2 résume les relations entre un espace euclidien et son dual.

$$\begin{array}{ll} u & (R^J, M) \\ & M \downarrow \quad \uparrow M^{-1} \\ Mu & (R^{J*}, M^{-1}) \end{array}$$

**Figure 5.2** Relations entre l'espace euclidien $R^J$ et son dual.

Or le nuage des profils-lignes dans $R^{J*}$, noté ici $N_I^*$, est l'image, par cet isomorphisme $M$, du nuage $N_I$ défini en 5.3.2. En effet, si l'on applique $M$ au point $i$ de $N_I$ , sa $j^{\text{e}}$ coordonnée devient :

$$x_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \xrightarrow{M} x_{ij}^* = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}} = \frac{f_{ij}}{f_{i.}} - f_{.j}$$

On retrouve la coordonnée, après centrage, du point $i$ dans le nuage des lignes de l'AFC du chapitre 3. L'AFC présentée dans ce chapitre est l'analyse factorielle de $N_I$ dans $(R^J, M)$. L'AFC présentée au chapitre 3 est celle de $N_I^*$ dans $(R^{J*}, M^{-1})$.

L'isomorphisme entre ces deux nuages assure la même décomposition sur les axes d'inertie, donc l'égalité des facteurs de rangs homologues. Notons que les axes d'inertie sont situés dans des espaces différents et, par conséquent, sont différents.

Parallèlement, les colonnes de $X$ (défini en a) peuvent être représentées par un nuage $N_J$ situé dans un espace de dimension $I$ (noté $R^I$) muni de la métrique $D$. L'application $D$ associe à ce nuage $N_J$ le nuage $N_J^*$ situé dans le dual de $R^I$ muni de la métrique $D^{-1}$. Le nuage $N_J^*$ n'est rien d'autre que le nuage des profils-colonnes analysé au chapitre 3. Les nuages $N_J$ et $N_J^*$ sont isomorphes et les facteurs sur $J$ de l'AFC peuvent être obtenus par l'analyse factorielle de l'un ou l'autre de ces nuages.

### 5.3.3 Matrices X, M, D en ACM

Les facteurs de l'ACM pouvant être obtenus en réalisant les calculs de l'AFC sur le tableau disjonctif complet, appliquons les formules du paragraphe précédent à un

TDC. Notons $y_{ij}$ le terme général de ce tableau, $Q$ le nombre de variables qualitatives, $I_j$ le nombre d'individus possédant la modalité $j$ et $I$ le nombre total d'individus.

Le terme général $f_{ij}$ du tableau dont la somme des termes est égale à 1 est $y_{ij}/IQ$. La marge sur $J$ a pour terme général $I_j/IQ$. La marge sur $I$ est constante et égale à $1/I$ : la matrice $D$ est donc, à $1/I$ près, la matrice identité et le terme général de la matrice $X$ s'écrit :

$$x_{ij} = \frac{f_{ij}}{f_{i.}f_{.j}} - 1 = \frac{Iy_{ij}}{I_j} - 1$$

Comme il s'agit d'un tableau *Individus*×*Variables*, on peut souhaiter imposer des poids différents aux individus. Par exemple, tripler le poids d'un individu est équivalent à tripler la ligne concernant cet individu : la structure disjonctive complète est donc conservée lorsque des poids $p_i$ sont affectés aux individus.

## 5.4 DUALITÉ

### 5.4.1 Relations entre les axes d'inertie et les facteurs des deux nuages

Le calcul des axes d'inertie et des facteurs du nuage des colonnes est absolument identique à celui du nuage des lignes. Tous les résultats concernant le nuage des colonnes se déduisent de ceux obtenus pour le nuage des lignes, en remplaçant $X$ par sa transposée $X'$ et en échangeant les matrices $M$ et $D$.

Ainsi, dans l'espace $R^I$, on cherche une suite de vecteurs $\{v_s; s = 1, ..., I\}$, chacun rendant maximum la quantité $v_s' XMX'Dv_s$ sous la double contrainte d'être unitaire ($v_s' Dv_s = 1$) et orthogonal aux vecteurs déjà trouvés ($v_s' Dv_t = 0$ pour tout $t < s$). La solution est donnée par l'équation :

$$XMX'Dv_s = \mu_s v_s$$

qui exprime que $v_s$ est vecteur propre unitaire de $XMX'D$ associé à la valeur propre $\mu_s$ de rang $s$. La comparaison de cette équation avec l'équation aux facteurs de la section 5.2.6 ($XMX'DF_s = \lambda_s F_s$) conduit aux deux résultats suivants.

1. $\mu_s = \lambda_s$ : les inerties projetées des nuages $N_I$ et $N_J$ sur leurs axes principaux de même rang sont identiques. Ces valeurs propres étant positives ou nulles, les inerties totales des deux nuages sont égales à $\sum_s \lambda_s$. Lorsque les matrices $X'DXM$ et $XMX'D$ ne sont pas de même dimension et admettent des nombres différents de valeurs propres, les valeurs propres non communes aux deux matrices sont nulles.
2. Les facteurs $F_s$ et les axes $v_s$ sont vecteurs propres, de la même matrice $XMX'D$, associés à la même valeur propre. Or, les équations aux vecteurs propres caractérisent ces vecteurs à la norme près (sauf en cas d'égalité de plusieurs valeurs

propres, cas particulier ne se produisant jamais avec des données réelles). Le facteur $F_s$ et l'axe $v_s$ sont donc deux vecteurs colinéaires de $R^I$.

Le vecteur $v_s$ étant unitaire et la norme de $F_s$ étant donnée par :

$$\|F_s\|_D^2 = u_s' M X' D X M u_s = \lambda_s$$

il en résulte la relation très importante :

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

Un raisonnement analogue, comparant l'équation aux facteurs $G_s$ sur les colonnes et l'équation aux axes $u_s$, conduit à la relation symétrique de la précédente :

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

Ces deux dernières relations sont illustrées dans la **figure 5.3**. L'ensemble des résultats est présenté schématiquement dans le **tableau 5.2**.

**Tableau 5.2** Les deux nuages, leurs axes d'inertie et leurs facteurs.

| | Nuage $N_I$ | Nuage $N_J$ |
|---|---|---|
| Espace | $R^J$ | $R^I$ |
| Métrique | $M$ | $D$ |
| Coordonnées | $X$ | $X'$ |
| Poids | $D$ | $M$ |
| Axe d'inertie | $u_s$ | $v_s$ |
| Equation | $X'DXMu_s = \lambda_s u_s$ | $XMX'Dv_s = \lambda_s v_s$ |
| Norme | $\|u_s\|_M = 1$ | $\|v_s\|_D = 1$ |
| Orthogonalité | $\langle u_s, u_t \rangle_M = 0$ si $s \neq t$ | $\langle v_s, v_t \rangle_D = 0$ si $s \neq t$ |
| Facteur | $F_s = XMu_s$ | $G_s = X'Dv_s$ |
| Equation | $XMX'DF_s = \lambda_s F_s$ | $X'DXMG_s = \lambda_s G_s$ |
| Norme | $\|F_s\|_D = \sqrt{\lambda_s}$ | $\|G_s\|_M = \sqrt{\lambda_s}$ |
| Orthogonalité | $\sum_s F_s(i)F_t(i)p_i = 0$ si $s \neq t$ | $\sum_s G_s(j)G_t(j)m_j = 0$ si $s \neq t$ |
| Inertie sur l'axe $s$ | $\lambda_s$ | $\lambda_s$ |
| Inertie totale | $\sum_s \lambda_s = \sum_i \sum_j p_i p_j x_{ij}^2$ | $\sum_s \lambda_s = \sum_i \sum_j p_i p_j x_{ij}^2$ |

### 5.4.2 Le schéma de dualité

La méthode factorielle consiste à analyser simultanément d'une part dans $(R^J, M)$ le nuage $N_I$ affecté des poids contenus dans $D$ et d'autre part dans $(R^I, D)$ le nuage $N_J$ affecté des poids contenus dans $M$.

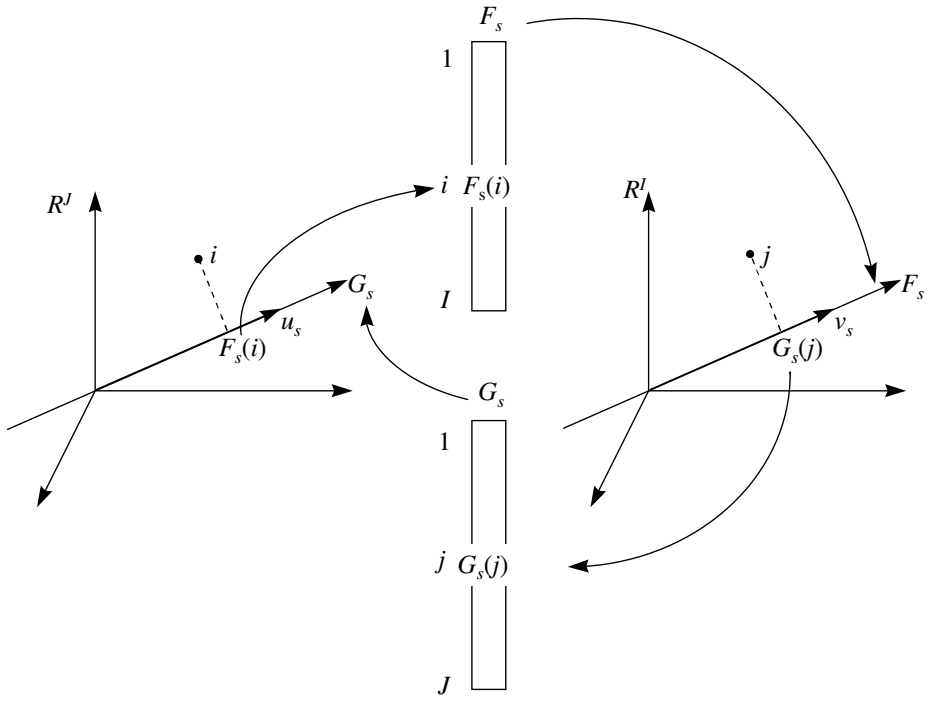



**Figure 5.3** Relations entre les axes d'inertie d'un nuage et les facteurs de l'autre nuage.

Les matrices $XM$ et $X'D$ définissent des applications de $R^I$ dans $R^J$ et de $R^J$ dans $R^I$ qui lient les facteurs et les axes des deux nuages.

L'application $M$ a déjà été considérée comme un isomorphisme de $R^J$ dans son dual $R^{J*}$ (*cf.* section 5.3.2). De même, $D$ définit un isomorphisme de $R^I$ dans son dual $R^{I*}$. L'analyse des nuages $N_I$ et $N_J$ est équivalente à celle de leurs images $N_I^*$ et $N_J^*$ par $M$ et $D$.

La matrice $X$ définit donc une application $X$ de $R^{J*}$ dans $R^I$. De façon analogue, la matrice $X'$ définit une application de $R^{I*}$ dans $R^J$. La **figure 5.4**, appelée « le schéma de dualité », récapitule ces applications et les relations qui permettent de passer des axes (ou des facteurs) d'un nuage aux axes (et aux facteurs) des autres nuages.

Si, par exemple, on applique au vecteur $u_s$ de $R^J$ successivement $M$, $X$, $D$ et $X'$, on obtient $u_s$ au coefficient $\lambda_s$ près. L'écriture de cette propriété pour n'importe quel axe principal ou n'importe quel facteur fournit l'équation qui le caractérise, c'est-à-dire la matrice dont il est vecteur propre. Ainsi, par exemple, les axes principaux $u_s^*$ de $N_I^*$ vérifient : $MX'DXu_s^* = \lambda_s u_s^*$.

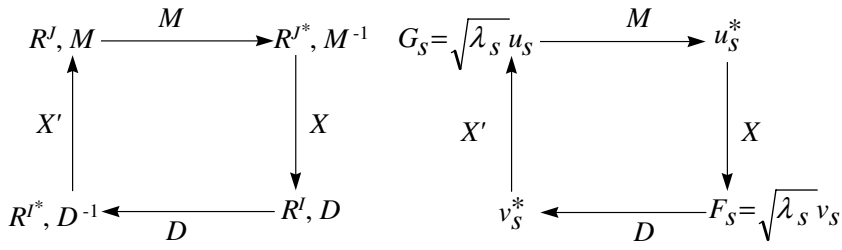



**Figure 5.4** Le schéma de dualité. $M, D, X$ et $X'$ désignent ici les applications associées aux matrices de même nom. À gauche, les espaces en jeu et leur métrique ; à droite, les résultats de l'analyse factorielle dans ces espaces.

Appliqué à l'AFC, ce diagramme schématise bien les deux présentations de la méthode. Si l'on met en évidence les écarts à l'indépendance de chaque case du tableau de données, on analyse les nuages $N_I$ et $N_J$, et la matrice des poids pour un nuage est la métrique pour l'autre. Si l'on met en évidence les profils (*cf.* chapitre 3), on analyse les nuages $N_I^*$ et $N_J^*$, et la matrice des poids pour un nuage est l'inverse de la métrique pour l'autre.

### 5.4.3 Dualité des objectifs en ACP

Toute projection $F_u$ du nuage $N_I$ sur un axe $u$ de $R^J$ est une combinaison linéaire des colonnes $x_j$ de $X$ puisqu'elle s'écrit $F_u = XMu = \sum_j m_j u_j x_j$. Réciproquement, toute combinaison linéaire $\sum_j y_j x_j$ des colonnes $x_j$ est colinéaire à une projection de $N_I$ sur un axe de $R^J$, l'axe défini par le vecteur de composantes $y_j / m_j$ :

$$\sum_j y_j x_j = \sum_j m_j \frac{y_j}{m_j} x_j$$

En ACP, où les colonnes de $X$ sont les variables initiales, nous avions proposé deux objectifs : la recherche de projections du nuage des individus et la recherche de variables synthétiques, combinaisons linéaires des variables initiales. Les critères d'ajustement choisis, inertie projetée maximum du nuage d'individus et variable maximisant la somme des carrés des corrélations avec les autres variables (= inertie projetée du nuage des variables), aboutissent au même résultat. L'identité entre projection du nuage d'individus et combinaison linéaire des variables montre que ces deux objectifs sont deux expressions d'un même problème exprimé à travers les individus d'une part et à travers les variables d'autre part.

### 5.4.4 Formules de transition

Dans le paragraphe 5.4.1, un aspect de la liaison entre les analyses de chacun des deux nuages a été exprimé à l'aide des relations suivantes :

$$\begin{aligned} u_s &= \frac{1}{\sqrt{\lambda_s}} G_s \\ v_s &= \frac{1}{\sqrt{\lambda_s}} F_s \end{aligned}$$

Elles indiquent que, dans l'espace $R^I$, la représentation des colonnes $(G_s)$ sert de base $(u_s)$ à la représentation des lignes et réciproquement. La liaison entre les facteurs des deux nuages est donc une liaison fondamentale et il est nécessaire de les interpréter conjointement.

Les formules de transition permettent de calculer les projections de l'un des deux nuages en fonction des facteurs sur l'autre nuage. Elles dérivent directement des relations entre axes et facteurs et s'écrivent :

$$\begin{aligned} F_s &= \frac{1}{\sqrt{\lambda_s}} XMG_s \\ G_s &= \frac{1}{\sqrt{\lambda_s}} X'DF_s \end{aligned}$$

Ce qui donne, point par point :

$$\begin{aligned} F_s(i) &= \frac{1}{\sqrt{\lambda_s}} \sum_j x_{ij} m_j G_s(j) \\ G_s(j) &= \frac{1}{\sqrt{\lambda_s}} \sum_i x_{ij} p_i F_s(i) \end{aligned}$$

Ces formules montrent comment, de façon concrète, les facteurs des deux nuages doivent s'interpréter conjointement, c'est-à-dire comment chacun des ensembles peut servir de support et d'aide à l'interprétation des facteurs de l'autre ensemble. Dans une représentation superposant les projections des lignes et des colonnes (pour les facteurs de même rang), la relation entre la position d'un élément d'un ensemble et celles de tous les éléments de l'autre ensemble peut s'exprimer ainsi : si $x_{ij}$ est positif, il y a attirance entre $i$ et $j$, si $x_{ij}$ est négatif il y a répulsion. Les poids $m_j$ et $p_i$ pondèrent cette influence. Un élément $i$ (resp. $j$) est donc situé du côté des éléments $j$ (resp. $i$) pour lesquels les valeurs de $x_{ij}$ sont les plus grandes.

Appliquée à l'ACP normée, la seconde formule de transition montre que la coordonnée de la variable centrée-réduite $j$ sur l'axe de rang $s$ est égale au coefficient de

corrélation entre la variable $j$ et le facteur $F_s$. Si l'on applique ces formules à l'AFC, on obtient :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}} f_{.j} G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j) - \frac{1}{\sqrt{\lambda_s}} \sum_j f_{.j} G_s(j)$$

La première formule indique que l'élément $i$ est situé du côté des $j$ auxquels il s'associe plus que dans l'hypothèse d'indépendance et est à l'opposé de ceux auxquels il s'associe moins que s'il y avait indépendance. La deuxième formule, où le second terme est nul du fait du centrage de $G_s$, se réduit à la formule barycentrique déjà commentée dans la présentation de l'AFC (*cf.* section a page 75).

### 5.4.5 Analyse factorielle de tableaux de distances ou de similarités

Les vecteurs propres de $XMX'D$ coïncident, à un facteur près, avec les facteurs sur les individus. Il est donc possible d'obtenir, uniquement à partir de la matrice $W = XMX'$ des produits scalaires entre individus et de la matrice $D$ des poids des individus, une projection du nuage $N_I$, sans faire appel aux coordonnées des individus dans $R^K$.

Concrètement, on dispose rarement d'une matrice de produits scalaires entre individus sans disposer en même temps des coordonnées des individus. En revanche, il peut arriver qu'un ensemble de données soit constitué uniquement de l'ensemble des poids des individus et de leur distances. En notant $p_i$ le poids d'un individu $i$, d$(i,l)$ la distance entre deux individus $i$ et $l$ et en posant :

$$d^2(i,.) = \sum_l p_l d^2(i,l)$$

$$d^2(.,.) = \sum_i p_i d^2(i,.) = \sum_{i,l} p_i p_l d^2(i,l)$$

on peut définir une matrice $W$ de « produits scalaires » de terme général :

$$\langle i,l \rangle = \frac{1}{2} \left[ d^2(i,.) + d^2(.,l) - d^2(i,l) - d^2(.,.) \right]$$

Cette relation est dite « formule de Torgerson ». On appelle Analyse Factorielle sur Tableau de Distances (AFTD) la technique qui, à partir d'un tableau de distances entre individus, calcule la matrice $W$ associée et construit la représentation des individus déduite des premiers vecteurs propres de $WD$ (cette idée dérive de la propriété selon laquelle $F_s$ est vecteur propre de $WD$ ; *cf.* tableau 5.2).

On peut montrer que si la distance d est une distance euclidienne, $W$ correspond au produit scalaire dont dérive d. Si d n'est pas une distance euclidienne, les valeurs

propres de $WD$ ne sont pas toutes positives. Dans ce cas, on se limite aux vecteurs propres associés aux valeurs propres positives, c'est-à-dire à une approximation euclidienne des données.

Si les données ne sont pas des distances mais des similarités, on se ramène au cas précédent en les transformant en distances. Par exemple, on peut définir la distance par la différence entre la borne supérieure des similarités et chaque similarité.

## 5.5 MISE EN ŒUVRE DES CALCULS

Les formules de transition, outre l'intérêt fondamental qu'elles présentent pour l'interprétation conjointe des facteurs des lignes et des colonnes, permettent des économies de calcul très substantielles. En effet, les facteurs de l'un des ensembles se déduisant des facteurs de l'autre ensemble, il suffit de diagonaliser une seule matrice pour obtenir tous les résultats. Ainsi, au niveau des calculs, une des deux dimensions du tableau de données n'est pratiquement pas limitée. La plupart des logiciels diagonalisent une matrice de dimension égale au nombre des colonnes, $X'DXM$ par exemple, dont les facteurs $G_s$ sont vecteurs propres. La construction de cette matrice ne nécessite qu'une seule lecture ligne à ligne du tableau de données, propriété précieuse dans le cas d'un très grand nombre d'individus ne permettant pas le stockage des données en mémoire.

En ACP, où les individus et les variables ne sont pas traités de la même façon, ce sont les variables qui constituent les colonnes car elles sont le plus souvent moins nombreuses que les individus. Lorsqu'il n'y a pas de pondération des variables, c'est la matrice des corrélations qui est diagonalisée si les variables sont réduites ; c'est la matrice des covariances, lorsque les variables ne sont pas réduites. Certains logiciels diagonalisent la plus petite des deux matrices $X'DXM$ et $XMX'D$ ce qui permet d'analyser des tableaux dans lesquels un petit nombre d'individus est décrit par un très grand nombre de variables.

### 5.5.1 Simplification en AFC

En AFC, la matrice $X'DXM$ a pour terme général :

$$a_{jj'} = \sum_i \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j}} - f_{.j'}$$

On se contente généralement du premier terme ; la matrice obtenue correspond alors à une analyse des nuages non centrés pour laquelle $X$ se réduit à $f_{ij}/f_{i.}f_{.j}$ tandis que $M$ et $D$ restent inchangées. Après diagonalisation de cette matrice, on supprime des résultats son premier vecteur propre, associé à la valeur propre 1 et dont toutes les coordonnées sont égales (appelé facteur trivial). Un calcul simple permet en effet de vérifier que ce vecteur est aussi vecteur propre de $X'DXM$, associé à une valeur

propre nulle et que les autres vecteurs propres et valeurs propres sont exactement ceux de $X'DXM$.

Indiquons le principe de ce calcul. Les deux matrices définies par les deux éléments du terme général de $X'DXM$ admettent, comme vecteur propre associé à la même valeur propre 1, le vecteur dont toutes les coordonnées sont égales à 1 :

$$\sum_{j'} \sum_{i} \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j}} = \sum_{j'} f_{.j'} = 1$$

La seconde matrice est de rang 1. Elle annule donc tous les vecteurs orthogonaux à ce premier vecteur et notamment tous les autres vecteurs propres de la première matrice.

### 5.5.2 Diagonalisation d'une matrice non symétrique particulière

Lorsque $M$ n'est pas un multiple de l'identité (ce qui est toujours le cas en AFC et qui se produit en ACP lorsque les variables ont des poids différents), la matrice à diagonaliser $X'DXM$ n'est pas symétrique. Or, les algorithmes de diagonalisation de matrices symétriques sont beaucoup plus efficaces que ceux d'une matrice quelconque. Aussi, on construit et on diagonalise plutôt la matrice symétrique $M^{1/2}X'DXM^{1/2}$. Cette matrice a les mêmes valeurs propres que $X'DXM$ et il suffit d'appliquer la matrice $M^{-1/2}$ à ses vecteurs propres pour obtenir ceux de $X'DXM$ puisque, si $u$ est vecteur propre de $X'DXM$, on a :

$$\begin{aligned} X'DXMu &= \lambda u \\ M^{1/2}X'DXM^{1/2}M^{1/2}u &= \lambda M^{1/2}u \end{aligned}$$

### 5.5.3 Enchaînement des calculs (*cf.* Tableau 5.3)

À l'issue de la diagonalisation de $X'DXM$, les facteurs $G_s$, définis sur $J$, s'obtiennent en multipliant les vecteurs propres $M$-normés de $X'DXM$ par $\sqrt{\lambda_s}$. Les facteurs $F_s$ s'obtiennent directement par projection du nuage $N_I$ sur $u_s$ ; cette opération est valable pour tous les éléments, actifs ou supplémentaires. Les projections des colonnes supplémentaires se déduisent ensuite de $F_s$ par la formule de transition de $F_s$ vers $G_s$.

Les facteurs et les inerties constituent les résultats de base d'une analyse factorielle. Ces résultats sont toujours complétés par des ensembles d'indices, appelés aides à l'interprétation, qui peuvent varier d'un logiciel à l'autre mais qui comprennent toujours au moins les qualités de représentation et les contributions à l'inertie de chaque élément, ligne ou colonne.

**Tableau 5.3** Enchaînement des calculs.

| Relation utilisée | Résultat |
|---|---|
| $X'DXMu_s = \lambda_s u_s$ | $u_s$ et $\lambda_s$ |
| $G_s = \sqrt{\lambda_s} u_s$ | Coordonnées des colonnes actives |
| $F_s = XMu_s$ | Coordonnées des lignes actives ou supplémentaires |
| $G_s = \frac{1}{\sqrt{\lambda_s}} X'DF_s$ | Coordonnées des colonnes supplémentaires |

## 5.6 RECONSTITUTION DES DONNÉES ET APPROXIMATION DE *X*

La projection d'un nuage sur ses axes d'inertie correspond à un changement de base orthonormée. En écrivant, par exemple, le vecteur $x_i$ représentant la ligne $i$ dans la base orthonormée des axes $u_s$, on obtient :

$$x'_i = \sum_s F_s(i) u_s$$

D'où, pour sa composante $x_{ij}$ sur la base canonique :

$$\begin{aligned} x_{ij} &= \sum_s F_s(i) u_s(j) \\ &= \sum_s \frac{F_s(i) G_s(j)}{\sqrt{\lambda_s}} \end{aligned}$$

Cette dernière expression, appelée formule de reconstitution des données, permet de calculer les valeurs $x_{ij}$ en fonction des facteurs et des valeurs propres de l'analyse. En limitant la somme à ses premiers termes, on obtient des valeurs approchées. La formule de reconstitution d'ordre $S$ ne retient que les $S$ premiers termes de la somme ; plus $S$ est grand, plus l'approximation se rapproche des données initiales.

### *Interprétation dans l'espace des matrices*

La formule de reconstitution des données s'écrit matriciellement :

$$X = \sum_s \frac{1}{\sqrt{\lambda_s}} F_s G'_s = \sum_s \sqrt{\lambda_s} v_s u'_s$$

La matrice $X$ est ainsi décomposée en une somme de matrices de rang 1 (le rang d'une matrice est la dimension de l'epace vectoriel engendré par ses colonnes ou par ses lignes).

Considérons l'espace des matrices de dimension $IJ$, noté $R^{IJ}$, muni de la métrique diagonale des produits $m_j p_i$. Dans cet espace, les matrices $v_s u'_s$ (de rang 1) forment

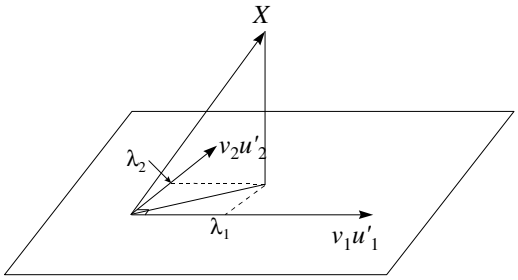



**Figure 5.5** Dans l'espace $R^{IJ}$, la reconstitution d'ordre 2 de $X$ est une projection de $X$ sur un plan.

un système orthonormé et $\sqrt{\lambda_s} v_s u_s'$ est la projection de $X$ sur $v_s u_s'$ :

$$\begin{aligned}
\langle v_s u_s', v_t u_t' \rangle_{m_j p_i} &= \sum_i \sum_j v_s(i) u_s(j) v_t(i) u_t(j) m_j p_i \\
&= \sum_i v_s(i)\ v_t(i)\ p_i \sum_j u_s(j)\ u_t(j)\ m_j \\
&= \begin{cases} 0 \text{ si } s \neq t \\ 1 \text{ si } s = t \end{cases} \\
\langle X, v_s u_s' \rangle_{m_j p_i} &= \sum_i \sum_j x_{ij} v_s(i) u_s(j) m_j p_i = \sum_i p_i v_s(i) F_s(i) = \sqrt{\lambda_s}
\end{aligned}$$

L'analyse factorielle décompose la matrice $X$, en tant que vecteur de l'espace $R^{IJ}$, sur un système orthonormé de matrices de rang 1. La restriction de la formule de reconstitution des données, à ses $S$ premiers termes, correspond à une approximation de $X$ par projection sur le sous-espace engendré par les $S$ premiers vecteurs (*cf.* **Figure 5.5**). Cette approximation est une matrice de rang $S$.

Le carré de la norme de la différence entre $X$ et son approximation d'ordre $S$ est égal à la somme des valeurs propres d'ordre supérieur à $S$.

On peut définir l'analyse factorielle par cette décomposition. L'objectif est alors d'approcher le tableau $X$ avec un tableau de rang fixé $S$ ($S$ étant supérieur ou égal à 1 et inférieur à $I$ et à $J$). On réalise l'ajustement avec le critère des moindres carrés pondérés, la case $(i, j)$ ayant le poids $m_j p_i$. On cherche alors une suite orthogonale de matrices de rang 1, qui s'écrivent donc comme le produit d'un vecteur $A_s$ de $R^I$ et d'un vecteur $B_s$ de $R^J$, qui minimisent l'expression :

$$\sum_i \sum_j \left( x_{ij} - \sum_{s=1}^{S} A_s(i) B_s(j) \right)^2 m_j p_i$$

Quelques calculs, en procédant par itération sur $s$, permettent de vérifier que la solution unique est donnée par les premiers facteurs de l'analyse factorielle.

## 5.7 UNE ÉQUIVALENCE EN ACM

Nous montrons ici que les facteurs sur les modalités en ACM peuvent être acquis en mettant en œuvre une AFC indifféremment sur un tableau disjonctif complet ou sur un tableau de Burt. Par commodité, nous appliquons l'AFC à des nuages non centrés.

Dans une AFC sur TDC, les matrices $X$, $M$ et $D$ ont pour terme général respectivement : $(I y_{ij}/I_j)$, $(I_j/IQ)$ et $1/I$ (*cf.* section 5.3.3). En notant $Y$ le TDC et $E$ la matrice diagonale des effectifs des modalités, on a :

$$X = IYE^{-1} \text{ et } M = E/(IQ)$$

La matrice diagonalisée s'écrit :

$$X'DXM = E^{-1}Y'Y/Q$$

Notons $B$ le tableau de Burt. Dans une AFC sur ce tableau, les matrices $X$ et $M$ ont pour terme général respectivement : $(II_{jk})/(I_jI_k)$ et $(I_j/IQ)$. Le tableau de Burt étant symétrique, les matrices $M$ et $D$ sont identiques. En remarquant que le tableau de Burt est lié au TDC ($B = Y'Y$), ces matrices peuvent s'écrire :

$$\begin{aligned} X &= IE^{-1}Y'YE^{-1} \\ M &= D = E/(IQ) \end{aligned}$$

La matrice diagonalisée s'écrit : $X'DXM = E^{-1}Y'YE^{-1}Y'Y/Q^2$

Si $u_s$ est l'axe de rang $s$ issu de l'AFC sur le TDC, alors il vérifie :

$$(1/Q)E^{-1}Y'Yu_s = \lambda_s u_s$$

Il vérifie aussi l'équation de l'AFC appliquée au tableau de Burt :

$$(1/Q)^2 E^{-1} Y' Y E^{-1} Y' Y u_s = \lambda_s^2 u_s$$

Compte tenu de la relation entre axes et facteurs ($u_s = F_s/\sqrt{\lambda_s}$), les deux analyses conduisent aux mêmes axes dans $R^J$ et aux mêmes facteurs sur les modalités. Toutefois, la valeur propre de rang $s$ issue de l'analyse du tableau de Burt est égale au carré de son homologue de l'analyse du TDC.

Chapitre **6**

# Exemple de traitement de tableau multiple par ACM et AFC

## 6.1 L'ENQUÊTE OUEST-FRANCE

Le quotidien régional *Ouest-France* a réalisé en 1973 une enquête auprès de 340 abonnés de Rennes et de sa région. Le but général de cette enquête était de « mieux connaître ses lecteurs », de savoir ce qui était lu dans le journal et de répondre à la question : « Qui lit quoi ? ».

Pour cela, 26 rubriques, qui paraissent quotidiennement et qui couvrent à peu près entièrement les sujets abordés dans le journal, ont été choisies (*cf.* **tableau 6.1**). À chaque enquêté, on demande d'indiquer, parmi les 26 rubriques citées, celles qu'il lit habituellement. Pour chaque rubrique, il y a deux solutions possibles : soit elle est déclarée *lue habituellement*, soit elle ne l'est pas. Chacune d'elles définit donc, sur l'ensemble des enquêtés, une variable qualitative à deux modalités : *rubrique-lue* et *rubrique-non-lue*. Le comportement de lecture d'un enquêté est donc caractérisé par un ensemble de 26 variables qualitatives à 2 modalités.

Pour préciser le « Qui » dans « qui lit quoi ? », plusieurs questions décrivant les individus complètent le questionnaire. L'ensemble de ces questions constitue ce que l'on appelle le signalétique. Ce signalétique est donné dans le **tableau 6.2**. Pour la plupart des questions (zone d'habitat, sexe, etc.), l'enquêté doit choisir une (et une seule) réponse parmi un ensemble proposé. Ces questions définissent donc aussi des variables qualitatives. Pour deux questions, l'âge et le nombre d'enfants, la réponse est un nombre. Afin de rendre homogène l'ensemble des variables, ces deux questions sont transformées en variables qualitatives à quatre modalités. La première, l'âge, est divisée en 4 tranches d'effectifs à peu près égaux : *16-26 ans*, *27-38 ans*, *39-55 ans*

**Tableau 6.1** Les 26 rubriques du journal Ouest-France et l'effectif de leurs lecteurs.

| Rubrique | Effectif | Rubrique | Effectif |
|---|---|---|---|
| informations locales | 276 | Lariflette | 159 |
| faits divers | 250 | reportage de la page 2 | 150 |
| page télé | 241 | jeunesse éducation | 139 |
| accidents | 224 | courrier des lecteurs | 128 |
| informations départementales | 221 | courrier des consommateurs | 127 |
| informations sociales | 208 | au jour le jour | 120 |
| informations politiques | 206 | pour vous Madame | 117 |
| décès | 204 | petites annonces | 112 |
| informations étrangères | 177 | cérémonies officielles | 109 |
| informations économiques | 167 | annonces sur l'emploi | 91 |
| sports | 164 | informations agricoles | 78 |
| l'article de première page | 161 | feuilleton | 46 |
| spectacles | 161 | informations maritimes | 17 |

et + *de 56 ans*. Les modalités de la seconde, le nombre d'enfants, sont : *pas d'enfant*, *1 enfant*, *2 enfants*, *3 enfants ou plus* (la dernière modalité regroupe les parents de famille nombreuse, en minorité dans l'échantillon).

L'ensemble des données est donc composé de deux groupes de variables qualitatives dont l'objet est différent : le signalétique décrit le *qui*, tandis que les rubriques décrivent le *quoi*.

## 6.2 ANALYSE SIMULTANÉE DE PLUSIEURS GROUPES DE VARIABLES

Le cas de cette enquête illustre une situation très courante : les variables définies sur un ensemble d'individus ne constituent pas un ensemble homogène mais sont structurées en plusieurs groupes. Les enquêtes, comme celle de *Ouest-France*, comprennent presque toujours, en plus du thème particulier de l'enquête, un questionnaire concernant le signalétique. En effet, chaque enquêté n'est pas intéressant en lui-même mais en tant que représentant de certaines catégories de la population. En outre, généralement (bien que ce ne soit pas le cas ici), le thème de l'enquête lui-même peut se subdiviser en plusieurs sous-thèmes qui constituent autant de sous-groupes de variables.

Ce type de structure existe aussi pour des variables numériques. Un exemple de cette nature est étudié dans le chapitre suivant : les variables sont des notes affectées à un ensemble de vins suivant certains critères de dégustation, soit olfactifs, soit visuels, soit gustatifs.

**Tableau 6.2** Le signalétique et les effectifs de ses modalités.

| | | |
|---|---|---|
| Zone d'habitat | Rurale | 132 |
| | Centre ville | 77 |
| | Z.U.P. | 72 |
| | Résidentielle | 43 |
| | non-réponse | 16 |
| Sexe | Homme | 198 |
| | Femme | 137 |
| | non-réponse | 5 |
| Situation de famille | Célibataire | 77 |
| | Marié | 229 |
| | Veuf | 24 |
| | Autre | 4 |
| | non-réponse | 6 |
| Age | 18-26 ans | 75 |
| | 27-38 ans | 91 |
| | 39-55 | 106 |
| | + de 55 ans | 61 |
| | non-réponse | 7 |
| Enfants à charge | Pas d'enfant | 159 |
| | 1 enfant | 46 |
| | 2 enfants | 63 |
| | 3 enfants et + | 72 |

| | | |
|---|---|---|
| CSP | Agriculteur | 35 |
| | Gros Com. Indus. | 14 |
| | Com. Artisan | 43 |
| | Cad. sup-Prof. lib. | 36 |
| | Cad.moyen | 55 |
| | Employé | 31 |
| | Ouvrier | 27 |
| | Etudiant-scolaire | 8 |
| | Retraité-div.-inactif | 25 |
| | Femme foyer | 10 |
| | non-réponse | 66 |
| Niveau d'instruction | Primaire | 117 |
| | Primaire supérieur | 66 |
| | Techniq. commerc. | 23 |
| | Secondaire | 51 |
| | Supérieur | 76 |
| | non-réponse | 7 |
| Mode d'habitat | Maison propriétaire | 113 |
| | Maison locataire | 62 |
| | Appart. propriétaire | 43 |
| | Appart. locataire | 114 |
| | non-réponse | 8 |

Il existe aussi des tableaux « mixtes » qui présentent des groupes de variables numériques et des groupes de variables qualitatives.

Un autre exemple de tableaux comprenant plusieurs groupes de variables est celui de mesures (numériques ou qualitatives) effectuées à plusieurs dates. Les variables mesurées peuvent être les mêmes à chaque date ou varier dans le temps. Contrairement au cas de l'enquête *Ouest-France*, le nombre de tableaux peut être alors très grand.

Pour analyser des données structurées en plusieurs groupes de variables, il est possible d'appliquer les méthodes classiques d'analyse factorielle : ACP pour des variables numériques et ACM pour des variables qualitatives. Une méthodologie s'est dégagée usant très largement de la technique des éléments supplémentaires : un ou plusieurs tableaux servent de base à l'analyse, les autres tableaux sont mis en supplémentaire.

Dans l'enquête *Ouest-France*, qui comporte deux groupes de variables qualitatives, deux solutions de ce type sont possibles :

1. une ACM de l'ensemble des rubriques (en principal) et du signalétique (en supplémentaire) ;
2. une ACM de l'ensemble du signalétique (en principal) et des rubriques (en supplémentaires) ;
   À ces deux analyses, dans lesquelles l'un des deux groupes est privilégié et sert de base de référence, on peut ajouter :
3. une ACM avec l'ensemble des rubriques et du signalétique en principal.
   Enfin, comme les deux groupes sont qualitatifs, s'ajoute une possibilité inexistante pour des groupes numériques :
4. une AFC du tableau croisant les deux groupes de variables.

L'objet de ce chapitre est double. D'une part, nous étudions et comparons les objectifs de ces différentes approches en les commentant sur l'enquête *Ouest-France*. D'autre part, nous montrons leurs limites et donc la nécessité d'ajouter, à la panoplie des méthodes factorielles, une technique qui inclut la notion de groupes de variables et qui puisse donner des solutions aux questions laissées sans réponse par les méthodes classiques.

Mais, avant d'aborder ces analyses, nous consacrons une section à la résolution du problème des réponses manquantes dans les questionnaires.

## 6.3 LE PROBLÈME DES RÉPONSES MANQUANTES

Dans l'enquête *Ouest-France*, ce problème se pose pour plusieurs variables du signalétique. Pour la CSP notamment, 66 individus n'ont pas indiqué de catégorie ; pour chacune des autres variables, le nombre de réponses manquantes ne dépasse pas 8. Pour la lecture des rubriques, le problème ne se pose pas : une rubrique non citée dans les lectures habituelles est considérée comme *non-lue*.

La manière de traiter les non-réponses à une question dépend de plusieurs éléments : le pourcentage d'individus concernés, la signification de cette non-réponse et surtout la manière dont la question intervient dans l'analyse (en élément actif ou supplémentaire).

### 6.3.1 Les non-réponses dans les variables supplémentaires

Dans ce cas, les non-réponses n'ont aucune influence sur l'ensemble des résultats et le problème n'est pas crucial. Une première solution consiste à créer pour chaque question concernée une modalité *non-réponse*. Une deuxième solution est envisageable si l'on applique un programme classique d'AFC au TDC : elle consiste à mettre en supplémentaire un tableau disjonctif incomplet (sans les modalités *non-réponse*). La seule différence entre ces deux solutions est que les modalités *non-réponse* n'apparaissent pas dans les résultats de la deuxième.

### 6.3.2 Les non-réponses dans les variables actives

Dans ce cas, l'ensemble des résultats dépend de la manière dont les non-réponses sont traitées et le problème doit être étudié avec soin.

La solution qui consiste à **créer une modalité supplémentaire** est encore possible. Mais il faut prendre garde au fait que cette modalité aura autant d'importance dans les typologies des individus et des variables que les autres modalités. Or, cela ne se justifie qu'à deux conditions. La première condition est que cette non-réponse traduise une attitude particulière (soit le refus de répondre, soit le fait de ne pas savoir, ou toute autre modalité de réponse non prévue dans le questionnaire) : s'il s'agit seulement d'une omission involontaire de l'enquêteur ou de l'enquêté, son influence doit être minimisée. La deuxième condition est que le pourcentage des réponses manquantes ne soit pas trop faible. On rejoint ici le problème des modalités rares dont l'influence risque d'être trop grande par rapport à la part très marginale de la population qu'elles concernent. Dans l'enquête, seule la CSP répond à ces deux conditions. Nous créons donc pour elle une modalité *non-réponse*. Pour les autres questions, le problème est le suivant : il faut obtenir les résultats d'une ACM en **minimisant l'influence des données manquantes** ou, plus généralement, celle de modalités qui traduisent une information dont on ne veut pas tenir compte (par exemple des modalités trop rares).

Une seconde solution consiste à **supprimer les individus dont les réponses au questionnaire sont incomplètes**. C'est une perte d'information qui n'est pas très regrettable si le nombre d'individus est très grand et les non-réponses rares. Dans l'enquête, cette solution n'est pas envisageable : sans tenir compte de la CSP, 37 individus présentent une seule donnée manquante et 6 en présentent deux ; ainsi, on serait conduit à se priver de 43 individus sur les 340.

Une troisième solution consiste à **ventiler aléatoirement les réponses inconnues sur l'ensemble des autres modalités de la même variable**. Cette technique a l'inconvénient de fausser les données ce qui pose un problème lorsque l'effectif à ventiler est assez important.

Une quatrième solution consiste à appliquer une **variante de l'ACM adaptée aux données manquantes**. Formellement, on peut la définir comme une variante de l'AFC appliquée à un tableau disjonctif incomplet[1]. Comme la plupart des propriétés de l'AFC d'un TDC qui définissent l'ACM tiennent au fait que la marge sur les individus est constante et que cette propriété n'est pas vérifiée pour les tableaux disjonctifs incomplets, le principe de la variante est de remplacer la marge réelle de ces tableaux par une marge constante partout où elle intervient (profil et poids des lignes, métrique et origine des axes du nuage des colonnes). Toutes les propriétés fondamentales de l'ACM sont vérifiées pour cette variante : dualité entre le nuages des individus et celui

1. Traitement des questionnaires avec non-réponse, Analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte. Publication ISUP XXXII fasc.3 1987 B.Escofier.

des modalités ; coïncidence entre une modalité et le barycentre de la population qu'elle caractérise ; maximisation de la somme des rapports de corrélation par les facteurs obtenus (le rapport de corrélation d'une variable ayant des réponses manquantes est calculé en plaçant les individus aux réponses inconnues au barycentre).

Dans ce chapitre, après avoir constaté que ces quatre méthodologies conduisent à des résultats assez proches, nous avons décidé de conserver la première solution, qui conservent les non-réponses telles quelles.

### 6.3.3 Les non-réponses dans les tableaux croisés

Pour le tableau croisé, la solution est très simple : les effectifs sont calculés avec les réponses effectivement connues, sauf pour la CSP dont la modalité *non-réponse* est introduite.

## 6.4 PREMIÈRE ANALYSE : ACM DES RUBRIQUES

Dans cette première analyse, celle commentée le plus largement, les éléments principaux sont les 52 modalités des 26 variables concernant la lecture des rubriques. Les 38 modalités des 8 variables du signalétique interviennent en supplémentaire. Avant d'étudier les résultats de cette analyse, indiquons brièvement ce que l'on peut en attendre.

1. Une typologie des individus suivant leur profil de lecture : deux individus sont proches s'ils lisent les mêmes rubriques du journal.
2. Une étude des liaisons entre la lecture (ou la non-lecture) des différentes rubriques : si plusieurs rubriques sont souvent lues par les mêmes lecteurs, elles constituent un groupe qui sera mis en évidence. Si, à l'inverse, il existe des phénomènes d'exclusion (les lecteurs de la rubrique A ne lisant jamais la rubrique B), ils seront détectés.
3. Avec les éléments supplémentaires, une étude de la liaison entre chaque variable du signalétique, **prise séparément**, et les principaux facteurs de variabilité des profils de lecture.

### 6.4.1 Plan des deux premiers facteurs

#### *a) Les variables actives : rubrique-lue et rubrique-non-lue*

Après s'être assuré que la répartition des individus est à peu près régulière sur le premier plan factoriel, on examine la projection des rubriques-non-lues et des rubriques-lues (*cf.* **Figure 6.1**). Notons d'abord que les deux modalités d'une même rubrique (*lue* et *non-lue*) sont toujours alignées avec l'origine des axes. En effet, en ACM, l'origine des axes est au barycentre des modalités d'une même variable (*cf.* section 4.3.5) ;

lorsqu'il n'y a que deux modalités, comme c'est le cas ici, l'origine est située sur le segment qui les joint. Certaines modalités, comme les informations économiques par exemple, ont des positions à peu près symétriques car les effectifs des lecteurs et des non-lecteurs de cette rubrique sont presque égaux (167 et 173). Lorsque ces effectifs ne sont pas du tout équilibrés (cas des informations maritimes qui n'intéressent que 17 personnes sur 340), la modalité lourde (non-lue) est près de l'origine tandis que la modalité légère (lue) est excentrée. En termes de mécanique, on retrouve le principe du bras de levier.

➤ Séparation des modalités lue et des modalités non-lue

Un simple coup d'œil à ce graphique révèle une structure particulière ; le phénomène serait beaucoup plus frappant encore si nous avions pu disposer de couleurs, faisant apparaître les rubriques lues en rouge et les non-lues en vert ! En effet, il existe une séparation très nette entre les deux types de modalités (par la deuxième bissectrice) : toutes les modalités *lue* sont au-dessus et toutes les modalités *non-lue* sont en dessous. Or rien ni dans le codage ni dans la méthode ne les différencie *a priori*.

Cette séparation, si nette sur le graphique, provient des données c'est-à-dire du comportement de lecture des enquêtés : globalement, il existe une certaine ressemblance entre l'ensemble des modalités *lue* d'une part et entre l'ensemble des *non-lue* d'autre part. Ceci implique que les tendances les plus marquantes dans la lecture du journal ne sont pas des exclusions systématiques (quand on lit le *sport*, on ne lit pas *pour vous Madame* et réciproquement) mais plutôt un effet « boule de neige » : quand on lit des rubriques, on a tendance à en lire d'autres, quelles qu'elles soient. Ce qui ne veut pas dire que l'attitude exclusive évoquée ci-dessus n'existe pas (on la découvre dans la suite) mais qu'elle est moins importante que le phénomène illustré sur ce premier plan.

La ligne de partage ne correspond pas à l'un des deux premiers facteurs mais, comme les inerties de ces facteurs sont très proches (0.155 et 0.125), le plan est une structure plus stable que chacun des facteurs pris séparément et on a tendance à l'étudier globalement. Nous pouvons cependant interpréter séparément chacun de ces facteurs.

➤ Groupes de rubriques

Sur le plan des deux premiers facteurs, certains regroupements sont visibles, notamment celui des informations étrangères, économiques, politiques et sociales avec les articles de fond de la page 1 et de la page 2 ainsi que la rubrique *jeunesse et éducation*. Les modalités *lue* de ces rubriques sont toutes situées en haut du graphique (coordonnée positive sur le deuxième facteur) et les modalités *non-lue* sont toutes situées en bas du graphique (coordonnée négative sur le deuxième facteur). Le point commun entre ces différentes rubriques est leur aspect relativement intellectuel. Cela explique sans doute qu'elles intéressent (ou n'intéressent pas) les mêmes sous-populations. Notons

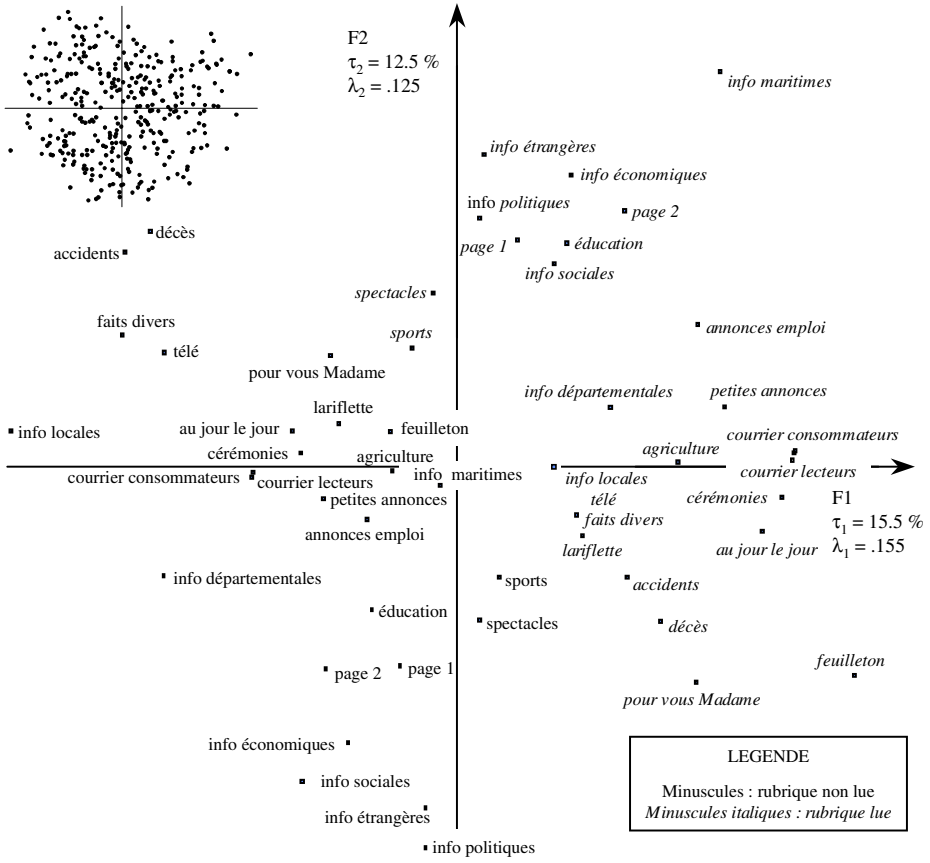



**Figure 6.1** Les deux premiers facteurs de l'analyse des rubriques : les rubriques. En haut à gauche, l'allure du nuage des individus.

que les proximités entre ces rubriques prises deux à deux ne sont pas forcément très fortes ; par contre, elles sont globalement assez proches entre elles (la lecture de l'une de ces rubriques est associée fréquemment à la lecture de deux ou trois autres rubriques du groupe). L'intérêt d'une analyse multidimensionnelle est de mettre en évidence de tels phénomènes.

### ➤ Le premier facteur

Les rubriques dont la contribution à l'inertie du premier facteur est la plus importante (en cumulant celle des deux modalités) sont : le courrier des lecteurs, le courrier des consommateurs, les décès et les autres rubriques d'information plutôt anecdotiques. Les modalités *lue* sont toutes situées à l'extrême-droite et les *non-lue* sont situées à l'extrême-gauche. Ce facteur oppose donc les lecteurs et les non-lecteurs de plusieurs

de ces rubriques. On peut aussi suivre l'ordre des rubriques lues et non-lues qui apparaît sur le premier axe et qui donne un classement des enquêtés : depuis les lecteurs assidus des rubriques très anecdotiques (courrier des lecteurs, feuilleton, cérémonies officielles, etc.) jusqu'à ceux qui non seulement ne s'intéressent pas à ces rubriques mais ne lisent même pas les informations locales qui ont pourtant un large public (276 sur les 340 enquêtés).

Schématiquement, ce facteur classe les lecteurs suivant l'intérêt qu'ils portent aux rubriques anecdotiques.

➤ Le deuxième facteur

Le deuxième facteur oppose les lecteurs des rubriques « intellectuelles » aux non-lecteurs de ces mêmes rubriques. Les premiers n'éprouvent guère d'intérêt pour les rubriques décès et accidents dont les modalités *non-lue* ont des coordonnées positives sur le deuxième axe.

➤ Les rubriques mal représentées sur le premier plan

La modalité *lue* de la rubrique *sports* est assez proche de celles des rubriques « intellectuelles ». Mais la qualité de représentation des deux modalités des sports (identiques car elles sont alignées avec l'origine) est faible sur les deux premiers axes (0.010 et 0.068) ; cela indique que sa position dans l'ensemble des rubriques est mal exprimée sur ce plan. Le rapprochement avec les rubriques intellectuelles existe, mais ce n'est pas ce qui caractérise le plus la lecture des sports : on en conclut aussi que l'attitude vis-à-vis de cette rubrique est assez indépendante de la dispersion générale des profils de lecture reflétée par le premier plan. Pour préciser la situation des sports dans la lecture du journal, il faut étudier plutôt le troisième facteur où sa qualité de représentation est la plus forte (0.279).

Nous n'attachons guère d'importance à la rubrique *informations maritimes* dont la position très excentrée est due à son faible effectif. Elle est en réalité peu liée aux deux premiers facteurs, comme l'indiquent ses qualités de représentation et ses contributions à l'inertie (sa qualité de représentation sur ces facteurs vaut 0.002 et 0.041 et les contributions cumulées de ses deux modalités valent 0.005 et 0.013).

### *b) Les individus*

Les observations précédentes ne répondent pas à la question « qui lit quoi ? ». L'analyse factorielle permet aussi d'y répondre. En effet, à la représentation des rubriques, on peut superposer une représentation des individus, ici les 340 enquêtés (pour des raisons de lisibilité, nous n'avons pas reproduit cette superposition). Dans ce nouveau nuage de points, deux aspects sont à retenir : d'une part, deux individus sont proches s'ils lisent (et ne lisent pas) les mêmes rubriques ; d'autre part, un individu est situé, à une homothétie près, au centre de gravité des modalités *lue* ou *non-lue* des rubriques qu'il lit ou ne lit pas. Concrètement, cela signifie qu'un individu situé au bas à droite du graphique

est un lecteur assidu d'un ensemble de rubriques assez peu intellectuelles (courrier des lecteurs, feuilleton, cérémonies officielles, etc.) dont les coordonnées sur l'axe horizontal sont fortement positives comme la sienne. Mais les informations politiques, étrangères, sociales ou économiques ne l'attirent guère car, pour ces rubriques, ce sont les modalités *non-lue* qui ont, comme lui, une coordonnée négative sur l'axe vertical.

Mais la position de tel ou tel point ne nous intéresse guère : le seul intérêt de ce graphique est de voir que les enquêtés se répartissent assez uniformément sur le plan et qu'il n'y a donc pas de classes de profils de lecture très marquées. Par contre, la position des enquêtés nous intéresse pour représenter les tendances du « qui ? » dans la question « qui lit quoi ? ». C'est là qu'intervient le signalétique des enquêtés puisque l'on connaît pour chaque individu son sexe, son niveau d'instruction (codé en cinq niveaux), sa CSP, etc. Pour mieux voir comment ces catégories sont liées aux modes de lecture, il est possible de représenter les deux barycentres des hommes et des femmes, les cinq barycentres du niveau d'instruction, les huit barycentres des CSP, etc. Ce point de vue sur l'analyse des individus se confond avec l'étude des modalités supplémentaires.

### *c) Les variables supplémentaires : le signalétique*

La projection des modalités de ces variables sur le plan 1-2 (seulement les plus éloignés du barycentre) est donnée **figure 6.2** où sont rappelées quelques-unes des rubriques. L'une des variables est étroitement liée au premier plan : c'est le **niveau d'instruction**. Les cinq niveaux d'instruction vont du plus faible au supérieur en passant par les niveaux intermédiaires. Il est remarquable de voir ces 5 niveaux ordonnés et alignés ; ils sont de plus très éloignés de l'origine qui représente le barycentre des 340 enquêtés.

Le fait de retrouver l'ordre naturel des cinq modalités du niveau d'instruction est un argument qualitatif mais essentiel pour conclure que cette variable est liée à la structure des profils de lecture schématisée sur le premier plan. L'éloignement, par rapport à l'origine, des cinq points est un autre argument qui peut être quantifié par le calcul du rapport de corrélation. Le carré de ce rapport, pour un facteur donné, est proportionnel à la somme des contributions des modalités de cette variable au facteur (*cf.* section 4.3.6 page 96). Parmi les variables supplémentaires, c'est le niveau d'instruction qui a le plus fort rapport de corrélation avec le second facteur ($\eta^2 = 0.247$) : en remplaçant les 340 individus par les cinq barycentres des classes de niveau d'instruction, on conserve presque le quart de l'inertie ! Le niveau d'instruction est donc très lié au profil de lecture ($\eta^2$(F1, niveau d'instruction) $= .157$) ; calculé pour le plan, l'inertie des barycentres des modalités du niveau d'instruction rapportée à l'inertie totale vaut : .207). Les lecteurs dont le niveau d'instruction est faible sont en moyenne en bas à droite du graphique : ils ont déjà été décrits. Ils s'opposent aux enquêtés de niveau d'instruction élevé, situés en haut à gauche. Ces derniers lisent les informations « intellectuelles » et passent sans s'arrêter sur les pages des décès, des accidents et des

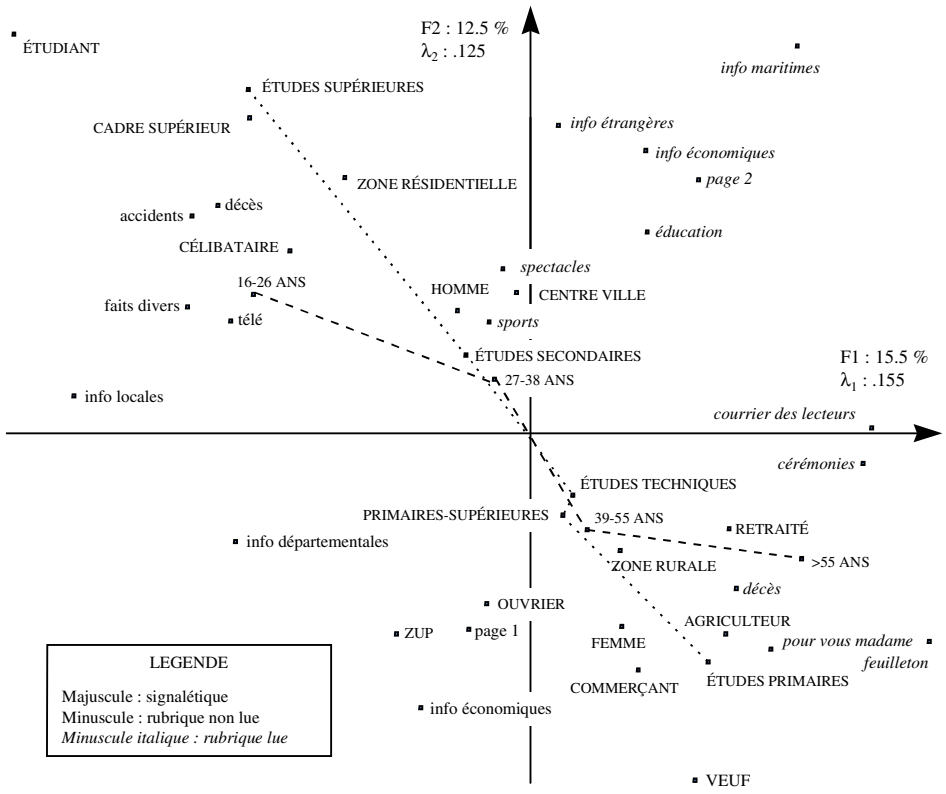



**Figure 6.2** Les variables du signalétique dans l'analyse des rubriques.

autres informations « anecdotiques ». C'est tout au moins ce qu'ils déclarent, car il ne faut pas oublier que nous disposons de réponses et non de comportements de lecture ; mais c'est un autre problème. Le respect de l'ordre des niveaux d'instruction montre que plus on est instruit, plus on est « intellectuel » dans la lecture de « *Ouest-France* ».

Que les cinq points représentant les niveaux d'instruction soient tous sur une même ligne montre que, dans la dispersion du nuage des profils de lecture, il existe une autre dimension indépendante du niveau d'instruction. Cette dimension, figurée sur le plan par la première bissectrice des deux axes, oppose les individus qui lisent un grand nombre de rubriques à ceux qui en lisent peu.

Nous ne commentons pas complètement l'ensemble des résultats concernant les variables. Nous laissons au lecteur le plaisir d'approfondir les interprétations. Notons seulement que la graduation obtenue pour le niveau d'instruction se retrouve dans la variable *âge* dont les modalités s'échelonnent aussi le long de la deuxième bissectrice : les enquêtés les plus âgés sont situés vers le bas et la droite du graphique.

Cette structure se retrouve aussi dans l'étude de la CSP : les étudiants et les cadres supérieurs sont des lecteurs « intellectuels » ; les retraités, agriculteurs, commerçants et ouvriers sont des lecteurs plus « anecdotiques ». Une seule variable, la zone d'habitat, traduit une dispersion en partie orthogonale à la première bissectrice : les habitants de la ZUP, situés en moyenne en bas à gauche du côté des modalités *non-lue*, lisent peu de rubriques dans le journal, beaucoup moins en tout cas que ceux du centre ville.

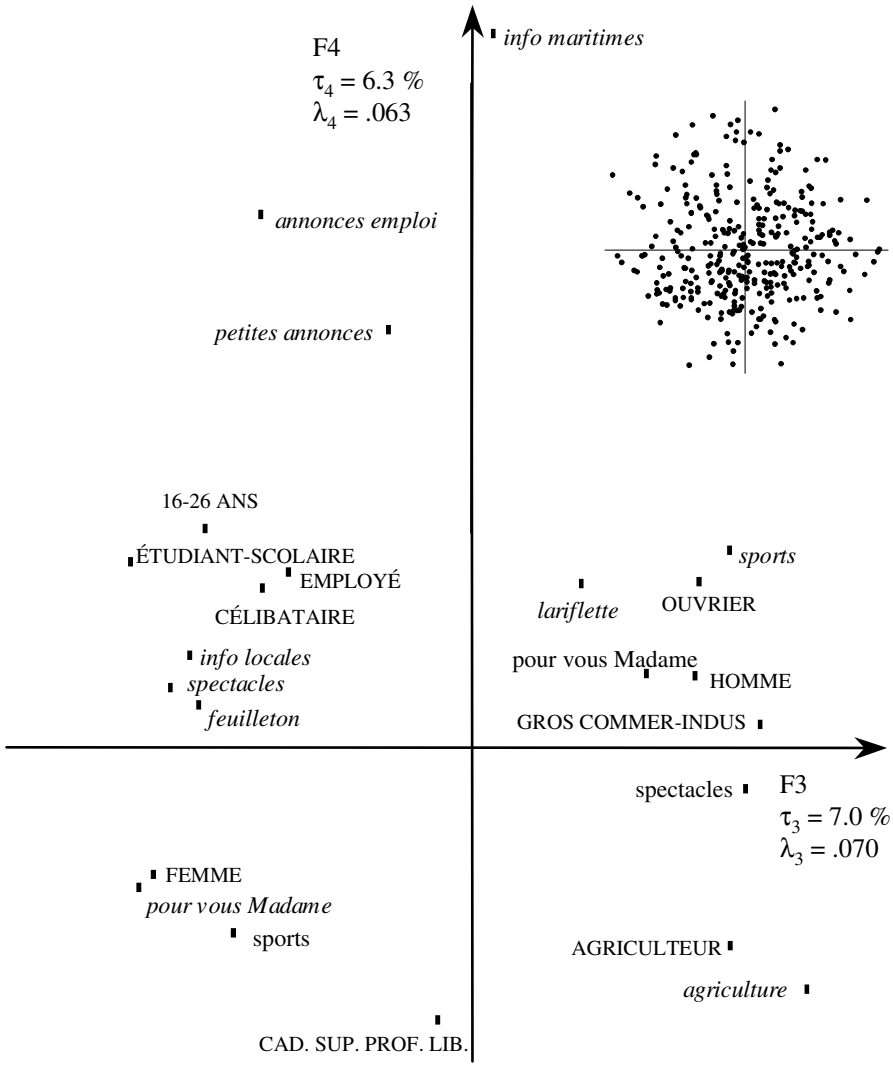



**Figure 6.3** Le plan 3-4 de l'analyse des rubriques. En italiques : rubriques lues ; en haut à droite, l'allure du nuage des individus.

### 6.4.2 Facteurs 3 et 4

La **figure 6.3** donne la projection des points les plus caractéristiques du plan 3-4 ainsi que l'allure du nuage des individus.

Le troisième facteur met en évidence une différence entre les profils de lecture, indépendante de celle traduite sur le premier plan. Trois rubriques contribuent à elles seules à la moitié de l'inertie du facteur : *sports*, *pour vous Madame* et *spectacles*. Ce facteur montre une opposition entre la première de ces rubriques et les deux autres : les lecteurs des *sports* ne sont pas lecteurs de *pour vous Madame* et des *spectacles* et réciproquement.

Les premiers sont plutôt des hommes et les seconds plutôt des femmes (le sexe est la variable du signalétique la plus liée à ce facteur : son rapport de corrélation atteint 0.318 !).

Le quatrième facteur particularise les lecteurs des *petites annonces* et *annonces pour l'emploi*, ces deux variables contribuant à 48 % de l'inertie de cet axe. Ce facteur est peu lié aux variables du signalétique.

## 6.5 DEUXIÈME ANALYSE : ACM DU SIGNALÉTIQUE

La solution inverse de la précédente consiste à mettre en éléments principaux les modalités des 8 variables du signalétique en laissant les modalités de lecture en éléments supplémentaires.

De l'analyse des variables du signalétique on peut attendre :

1. une typologie des individus suivant leur signalétique : deux individus sont proches si leurs signalétiques se ressemblent (et ce, indépendamment de leurs lectures) ;
2. une étude des liaisons entre les différentes variables du signalétique ;
3. avec les éléments supplémentaires, une étude de la liaison entre les principaux facteurs de variabilité du signalétique et la lecture de chaque rubrique **considérée séparément**.

La **figure 6.4** donne les projections des modalités actives et des modalités supplémentaires les plus éloignées du barycentre sur le plan 1-2. En outre, l'allure du nuage des individus est figurée.

### 6.5.1 Les modalités actives : le signalétique

➤ Le premier facteur

Sa valeur propre, moyenne des rapports de corrélation entre le facteur et chacune des variables actives, vaut 0.406. Cette valeur élevée indique une forte liaison globale

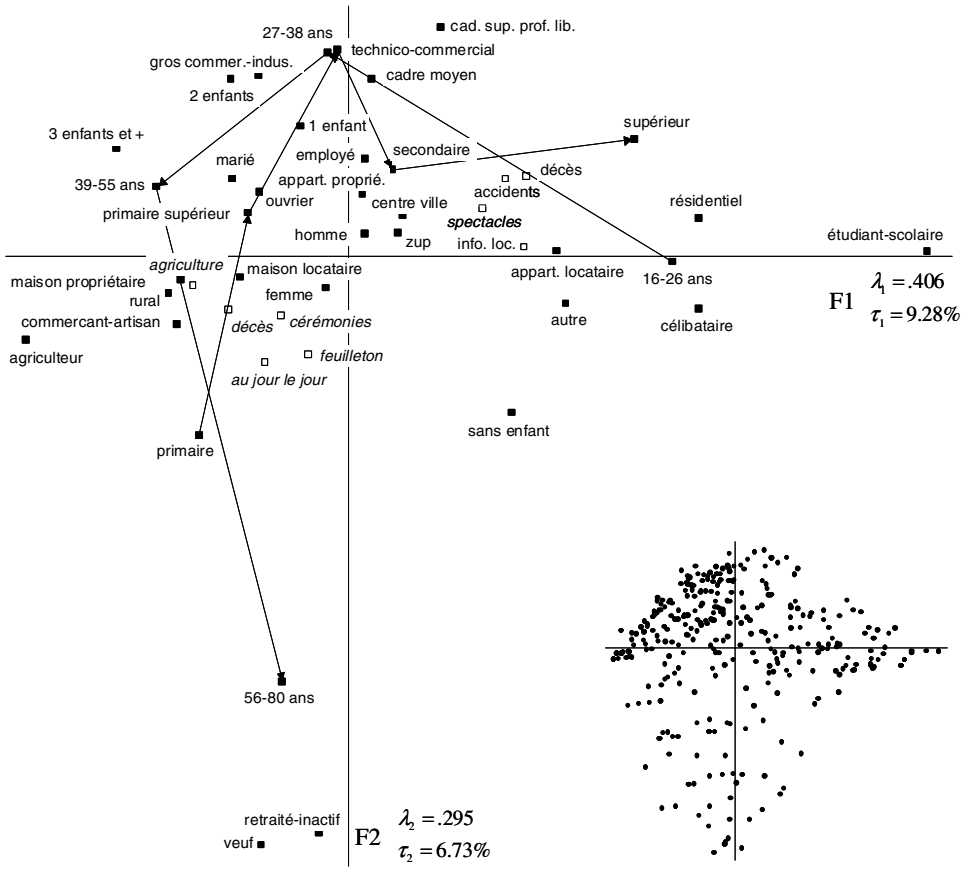



**Figure 6.4** Le plan 1-2 de l'analyse du signalétique. En italiques : rubriques lues ; en bas à droite, l'allure du nuage des individus.

entre les variables. L'examen de chacune des variables du signalétique montre que toutes, sauf une, le sexe, sont très liées à ce facteur et donc très liées entre elles. L'interprétation générale de ce facteur est simple : il met en évidence les plus jeunes, qui cumulent beaucoup de modalités qui les caractérisent bien : dans l'ensemble, ils sont étudiants, de niveau d'instruction élevé, célibataires, sans enfant et habitent une zone résidentielle ; toutes ces modalités, comme celles des CSP assez élevées dans la hiérarchie sociale, sont situées du même côté.

### ➤ Le deuxième facteur

Il détache très nettement des autres un groupe de trois modalités souvent partagées : âgé, veuf, et retraité. Ces trois modalités contribuent à un peu plus de la moitié de l'inertie de ce deuxième facteur. Ce deuxième axe met donc en évidence les plus âgés.

Finalement, on peut dire que ce plan est très structuré autour le la variable âge avec laquelle varient la plupart des autres variables du signalétique. Notons au passage la mise en évidence d'une liaison non linéaire entre l'âge et le nombre d'enfants : la modalités pas d'enfants (à charge) est associée à la fois aux plus jeunes et aux plus âgés.

➤ Description de l'échantillon

Cette analyse permet de décrire l'échantillon, ou plus exactement les liaisons entre les variables dans l'échantillon. Les résultats ne sont pas très originaux (que les jeunes soient plutôt célibataires et les veufs plutôt âgés n'est pas une découverte !). Mais cette banalité est un bon signe : une incohérence avec la réalité (bien connue) de la population générale signalerait un biais de l'échantillon par rapport à la population qu'il est censé représenter. Cette analyse attire aussi l'attention sur la liaison entre certaines variables : par exemple ici, le niveau d'instruction est très lié à l'âge de l'enquêté (il varie en sens inverse). Il faut prendre garde dans l'interprétation des résultats à ce lien sous-jacent : l'infléchissement avec l'âge des profils de lecture vers des rubriques de moins en moins intellectuelles (mis en évidence dans la première analyse) traduit peut-être essentiellement le moindre niveau d'instruction des enquêtés âgés.

### 6.5.2 Les modalités supplémentaires : les rubriques

Les modalités de lecture et de non-lecture des 26 rubriques sont presque toutes situées très près de l'origine des axes. Notons quand même quelques exceptions : les spectacles (qui intéressent les étudiants), les décès (lus beaucoup plus par les personnes âgées que par les jeunes), les informations agricoles (lues par les agriculteurs), etc.

### 6.5.3 Contradiction apparente entre les deux analyses

La plupart des rubriques étant très proches de l'origine, on en conclut que le signalétique est très peu lié à la lecture d'une rubrique particulière. Et pourtant nous avons vu dans la première analyse que profil de lecture et signalétique sont très liés !

Ces résultats ne sont pas contradictoires. Chacune des deux analyses focalise l'attention sur un aspect des données et ces points de vue se complètent mutuellement. Dans la première analyse, la typologie des individus traduit le comportement de lecture devant l'ensemble des rubriques. Ce comportement général est très lié à plusieurs variables du signalétique, notamment le niveau d'instruction. Dans cette deuxième analyse où les rubriques sont en éléments supplémentaires, elles apparaissent séparément. On en conclut que le signalétique est très lié au profil général de lecture mais peu à la lecture d'une rubrique particulière (un niveau social élevé implique très fréquemment une lecture « intellectuelle » du journal qui privilégie ce type de rubriques, mais non systématiquement l'une d'entre elles).

Cet exemple illustre une fois de plus la richesse de l'analyse multidimensionnelle qui permet de déceler des liaisons que l'étude séparée de chaque dimension (ici chaque rubrique) ne peut révéler.

### 6.5.4 Perte des dispersions spécifiques des profils de lecture

L'opposition entre lecture et non-lecture, valable pour les 26 rubriques, est une des caractéristiques les plus marquantes de la première analyse. Elle traduit une dispersion des profils de lecture liée au nombre total de rubriques lues. Cela n'apparaît plus du tout dans cette analyse où la typologie des individus est faite sur leur signalétique uniquement : toute dispersion qui ne lui est pas liée est forcément invisible.

Inversement, la séparation du groupe veuf-retraité-agé apparaît beaucoup moins nettement dans la première analyse que dans la deuxième.

## 6.6 UNE ANALYSE NON SATISFAISANTE : ACM DES RUBRIQUES ET DU SIGNALÉTIQUE

Il est possible d'envisager une analyse dans laquelle les deux groupes de variables sont en actifs. Mais cette analyse est délicate car les variables des deux groupes sont hétérogènes. Étudions les problèmes spécifiques qu'elle pose.

### 6.6.1 Typologie des individus

Dans la typologie des individus obtenue par cette ACM, le signalétique et le profil de lecture interviennent simultanément, deux individus étant proches s'ils se ressemblent socialement **et** lisent les mêmes rubriques. Au premier abord, les deux groupes de variables semblent intervenir également. Mais cette égalité apparente peut cacher un déséquilibre important : rien n'empêche que l'un des deux groupes prédomine sur l'autre : la typologie « mixte » serait alors pratiquement la typologie induite par ce groupe. Ce n'est pas le but cherché : lorsqu'on considère simultanément deux groupes aussi hétérogènes, on souhaite (implicitement au moins) qu'ils interviennent réellement tous deux dans la typologie.

### 6.6.2 Typologie des modalités

L'intérêt d'une typologie conjointe des modalités des variables des deux groupes est d'étudier l'ensemble des liaisons : à la fois à l'intérieur de chaque groupe et entre les deux groupes. S'il y a déséquilibre entre les deux groupes, les liaisons internes du groupe dominant seront mises en évidence aux dépens des liaisons internes de l'autre groupe et des liaisons inter-groupes.

En conclusion, que ce soit pour l'étude des individus ou pour celle des variables, dans une analyse où plusieurs groupes de variables hétérogènes interviennent simultanément en actifs, il est nécessaire d'équilibrer leur influence.

### 6.6.3 Indices concernant les groupes

Dans ces données, en plus des trois types d'objets classiques de l'ACM (les individus, les modalités et les variables), un quatrième type apparaît : les groupes de variables. L'interprétation des résultats doit s'enrichir de chacun de ces niveaux. Pour les trois premiers types d'objets, on dispose de deux indices d'aides à l'interprétation : la contribution à l'inertie d'un facteur et un indice de liaison avec le facteur (qui est la qualité de représentation pour les individus et les modalités, et le rapport de corrélation pour les variables qualitatives). Tout naturellement, on souhaite disposer d'indices analogues pour les groupes de variables. La contribution à l'inertie mesure l'importance d'un groupe dans la typologie traduite par un facteur ; elle se définit en cumulant les contributions des variables du groupe. La liaison entre un groupe de variables et un facteur est une notion complexe dont nous proposons un indice de mesure dans la section 8.5 page 194.

En pratique, pour guider l'interprétation des résultats de données hétérogènes, des indices concernant les groupes sont nécessaires : par exemple, avant d'étudier en détail une typologie d'individus, il faut savoir si cette typologie correspond surtout à leur signalétique, à leur profil de lecture ou à ces deux aspects.

### 6.6.4 L'Analyse Factorielle Multiple, alternative de cette ACM

L'Analyse Factorielle Multiple (AFM), méthode d'analyse de tableaux comprenant plusieurs groupes de variables, apporte une solution très satisfaisante au problème de l'équilibre des groupes. Ses résultats sont plus complets que ceux de l'ACM. Ils comprennent des indices d'aides à l'interprétation concernant les groupes et bien d'autres indices permettant en plus de comparer les groupes entre eux.

Nous donnons donc les résultats de l'AFM plutôt que ceux de l'ACM. On les trouve dans la dernière section du chapitre suivant, après une présentation générale de cette méthode qui s'appuie sur un exemple plus simple concernant des variables numériques et non des variables qualitatives.

## 6.7 TROISIÈME ANALYSE : AFC DU TABLEAU CROISANT SIGNALÉTIQUE ET RUBRIQUES

Nous commentons ci-après l'AFC du tableau croisant l'ensemble des 52 modalités de lecture et les 38 modalités du signalétique. Ce tableau est formé d'une juxtaposition de tableaux de contingence ; c'est un sous-tableau du tableau de Burt défini par les

variables des deux groupes (*cf.* **Figure 6.5**). Cette quatrième analyse est assez différente des autres en ce sens qu'elle est focalisée sur la liaison entre les variables du signalétique et la lecture des rubriques.

Ce tableau est structuré, en ligne et en colonne, par les variables. La marge de chaque sous-tableau défini par l'ensemble des modalités d'une variable du signalétique (ou d'une rubrique) est proportionnelle à celle du tableau entier. Ceci implique que le barycentre des modalités d'une même variable est, comme en ACM, situé à l'origine des axes.

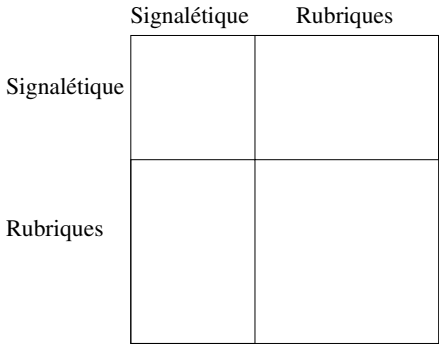



**Figure 6.5** Le tableau croisant signalétique et rubriques est un sous-tableau du tableau de Burt.

### 6.7.1 Plan des deux premiers facteurs (*cf.* Figure 6.6)

Le premier facteur extrait un fort pourcentage d'inertie (58 %).

#### *a) Le signalétique sur le premier facteur*

La variable du signalétique qui contribue le plus à l'inertie de ce facteur est le niveau d'instruction dont les 5 modalités ordonnées s'étagent de gauche (niveau supérieur) à droite (niveau le plus faible). La contribution cumulée de ces 5 modalités dépasse 25 %. L'âge, qui a aussi une contribution très importante, est gradué des plus jeunes (à gauche) aux plus âgés. L'opposition entre les CSP *agriculteur* d'une part et *cadre supérieur* et *étudiant* d'autre part, montre que ce facteur est lié au niveau social.

#### *b) Les rubriques sur le premier facteur*

Les modalités les plus extrêmes sont des modalités de non-lecture : celles des décès, des accidents et des informations locales. Ces 3 modalités, concernant des rubriques très anecdotiques, sont situées du même côté que les modalités de lecture des rubriques intellectuelles : informations étrangères, politiques, économiques et les spectacles. C'est un axe de niveau intellectuel.

### *c) La liaison signalétique-rubriques sur le premier facteur*

Du point de vue du signalétique, c'est un facteur de niveau social ; du point de vue des rubriques, c'est un facteur de niveau intellectuel des rubriques. Les modalités caractérisant un niveau social élevé sont liées aux modalités de lecture des rubriques les plus intellectuelles et aux modalités de non-lecture des rubriques les plus anecdotiques (et inversement). Sur ce facteur, on retrouve une structure assez proche de celle remarquée dans la première analyse (non pas le long d'un facteur mais le long de la deuxième bissectrice du plan 1-2 sur laquelle s'étageaient notamment les 5 modalités ordonnées du niveau d'instruction).

Le lien entre les profils de signalétique et de lecture est donc en très grande partie exprimé par la liaison entre le niveau social des lecteurs et le niveau intellectuel des rubriques lues.

### *d) Le deuxième facteur*

Le deuxième facteur extrait 15 % de l'inertie. Beaucoup moins important que le premier, il peut encore s'interpréter clairement.

Pour le signalétique, plus de la moitié de l'inertie de ce facteur est fournie par les deux modalités de la variable *sexe*. Pour les rubriques, la lecture du feuilleton et de *pour vous Madame* s'oppose à celle des informations agricoles et des sports.

Ce deuxième facteur montre qu'un deuxième élément important dans le lien entre le signalétique et la lecture des rubriques est l'opposition entre :

1. les hommes, lecteurs des *informations agricoles* et des *sports* ;
2. les femmes, lectrices de *pour vous Madame* et du *feuilleton*.

Cette liaison a déjà été en partie décelée sur le troisième facteur de la première analyse, l'ACM des rubriques.

## 6.7.2 Intérêt et limites de cette analyse

Cette analyse fait jouer un rôle symétrique aux deux groupes et met très bien en évidence différents aspects de leurs relations. Mais elle a trois limites principales :

1. elle ne permet d'étudier que des groupes de variables **qualitatives** ;
2. elle ne permet de comparer que **deux** groupes de variables ;
3. elle ne met en évidence que les points communs entre ces deux groupes : toutes les dimensions spécifiques du signalétique (groupe veuf-âgé-retraité) et des profils de lecture (nombre de rubriques lues) non liées à l'autre profil sont supprimées. L'importance relative des structures communes et des structures spécifiques ne peut absolument pas être mesurée.

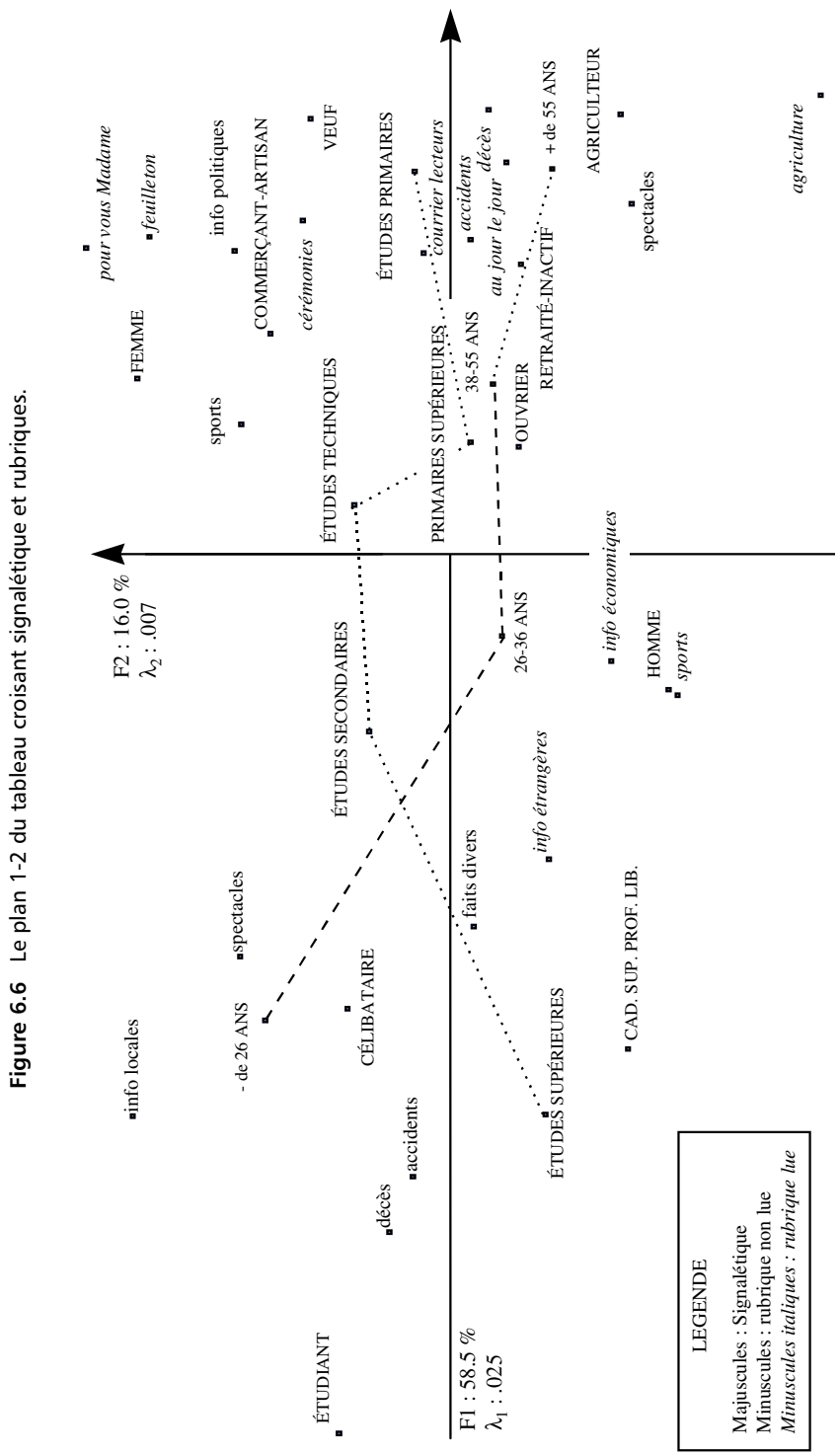



**Figure 6.6** Le plan 1-2 du tableau croisant signalétique et rubriques.

## 6.8 CONCLUSION

Ces trois analyses ont permis d'étudier plusieurs aspects de ces données.

Mais nous avons mis en évidence les limites des techniques classiques qui ne permettent ni la comparaison globale de plusieurs groupes de variables (structures communes et spécifiques) ni la construction des typologies des lignes et des colonnes dans laquelle les groupes sont équilibrés.

L'AFM, présentée dans le chapitre suivant, est conçue spécialement pour l'analyse de tableaux comportant plusieurs groupes de variables. Elle ne présente pas ces limites et permet une comparaison systématique des groupes à travers l'ensemble de points de vue très divers que la richesse de la structure de groupes de variables suggère.

Chapitre 7

# L'Analyse Factorielle Multiple à partir de deux applications

Ce chapitre présente une méthode factorielle adaptée au traitement de tableaux dans lesquels un ensemble d'individus est décrit par plusieurs groupes de variables : l'Analyse Factorielle Multiple (AFM). Cette présentation s'articule autour de deux exemples.

- Le premier rassemble des appréciations sensorielles fournies par un ensemble de dégustateurs sur un ensemble de vins. Les différents résultats de l'AFM sont commentés de façon à mettre en évidence les problèmes auxquels ils apportent des solutions. Les premiers résultats présentés sont très proches de ceux des méthodes classiques ; les résultats sans équivalents dans l'ACP ou l'ACM sont introduits progressivement.
- Le second est l'enquête *Ouest-France* étudiée par les méthodes classiques dans le chapitre précédent.

## 7.1 L'EXEMPLE DES VINS

### 7.1.1 Présentation des données

Cet exemple est issu de recherches réalisées pendant plusieurs années par l'équipe constituée autour de C. Asselin et R. Morlat au Centre de Recherches INRA d'Angers.

Les données se présentent initialement sous la forme suivante : 36 dégustateurs ont jugé chacun 21 vins à l'aide d'une fiche comprenant 29 variables. Les variables sont des caractéristiques du vin dont le dégustateur doit apprécier l'intensité à l'aide d'une échelle comportant cinq modalités ordonnées (très faible ou nul, faible, moyen, fort, très fort) codées de 1 à 5.

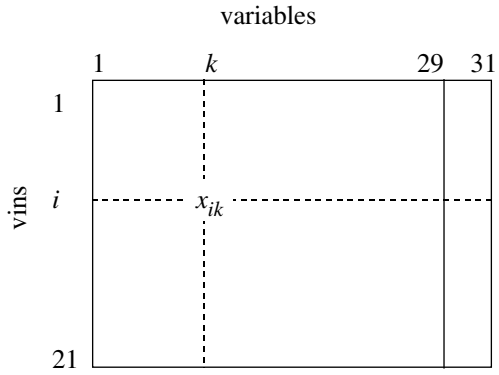



**Figure 7.1** Tableau analysé. Pour les 29 premières variables, $x_{ik}$ est la moyenne des appréciations des 36 juges sur le vin $i$ à propos de la variable $k$. Pour les deux dernières variables, $x_{ik}$ est le numéro de la modalité du vin $i$ pour la variable $k$.

À partir de ces données, un fichier plus petit a été construit (*cf.* **Figure 7.1**) : pour chaque vin et chaque variable de la fiche, on a calculé la moyenne des appréciations de l'ensemble des juges. Lorsqu'une donnée est manquante, elle n'intervient pas dans la moyenne.

À ce fichier de 29 variables numériques, on ajoute deux variables qualitatives qui caractérisent l'origine des vins : l'aire d'appellation (Saumur, Bourgueil, Chinon) et le type de sol (séquence de référence, milieu 2, milieu 3 et milieu 4 ; la séquence de référence est, selon l'hypothèse des chercheurs, le type de sol qui possède les meilleures potentialités viticoles).

Le tableau obtenu n'est pas homogène puisqu'il présente à la fois des variables quantitatives et qualitatives. Dans la suite, les variables qualitatives apparaissent au travers de leurs modalités : lors de l'interprétation, on parle peu de la variable *appellation* mais surtout de la modalité *Saumur*, de la modalité *Bourgueil*, etc.

### 7.1.2 Description de la problématique

L'objectif général de l'étude est la caractérisation de ces vins rouges. On cherche d'abord une typologie des vins permettant de répondre à des questions du type suivant : quels sont les vins qui globalement, c'est-à-dire du point de vue de l'ensemble des variables, se ressemblent ?

Pour cela, nous utilisons la méthodologie factorielle qui met en évidence les principales dimensions de variabilité et décrit les individus (ici les vins) à l'aide de ces

dimensions. Dans cette optique, une ACP semble bien adaptée au tableau. Dans cette ACP, comme dans l'AFM par la suite, les variables sont normées pour qu'elles aient la même influence *a priori*.

Toutefois, l'examen de la fiche de dégustation montre que les variables sont structurées en groupes. Tout d'abord, les variables qui caractérisent l'origine des vins jouent un rôle bien à part : elles ne doivent pas participer à la construction des principaux facteurs de variabilité mais simplement intervenir à titre illustratif. Cela étant, même parmi les variables sensorielles, on distingue :

- 5 variables relatives à l'**olfaction au repos** ; intensité olfactive, qualité aromatique, note fruitée, note florale, note épicée ;
- 3 variables relatives à la **vision** ; intensité visuelle, nuance (orangé/violet), impression de surface (larmes) ;
- 10 variables relatives à l'**olfaction après agitation** : intensité olfactive, qualité olfactive, note fruitée, note florale, note épicée, note végétale, note phénolique, intensité aromatique de bouche, persistance aromatique de bouche, qualité aromatique de bouche ;
- 9 variables relatives à la **gustation** : intensité d'attaque, acidité, astringence, alcool, équilibre acidité-astringence-alcool, velouté, amertume, intensité de fin de bouche, harmonie ;
- 2 variables relatives à un **jugement d'ensemble** ; qualité d'ensemble, typicité.

Le dernier groupe comporte deux variables synthétiques : nous décidons de leur faire jouer un rôle illustratif.

Les variables sur lesquelles nous appuyons principalement l'analyse sont donc structurées en quatre groupes : olfaction au repos, vision, olfaction après agitation, gustation. L'existence de cette structure pose d'abord un problème technique : une ACP globale ne risque-t-elle pas d'être influencée de façon prépondérante par un seul groupe ? Auquel cas, la prise en compte simultanée des quatre groupes serait illusoire.

Ainsi, le premier problème posé par le traitement simultané de plusieurs groupes de variables est la pondération de ces groupes. Dans un premier temps, l'AFM peut être vue comme une analyse factorielle (ici une ACP) dans laquelle l'influence des groupes de variables est équilibrée. C'est dans cet esprit que nous effectuons une première présentation des résultats de l'exemple dans la section 1.4. L'aspect technique de la pondération est présenté dans la section 1.3.

En outre, la prise en compte de la structure en groupes d'un ensemble de variables enrichit la problématique de l'étude. De même que l'on cherche à comparer des vins (en termes de ressemblances) ou des variables (en termes de liaisons), on peut chercher à comparer globalement les groupes de variables. On dira que deux groupes de variables se ressemblent si deux vins proches pour l'un des deux groupes (par exemple, l'aspect visuel) sont aussi proches pour l'autre (par exemple, le goût). On

tente donc de mettre en évidence une typologie des groupes, c'est-à-dire, dans notre exemple, des aspects sensoriels mis en jeu dans la dégustation. L'AFM fournit une telle typologie : son application à l'exemple est décrite aux sections 1.6 et 1.7.

L'existence de groupes de variables conduit à vouloir décrire les vins, non seulement au travers de l'ensemble des variables mais aussi au travers de chacun des groupes. Pour cela, il est toujours possible de réaliser des analyses séparées des groupes. Toutefois leurs résultats, étant obtenus indépendamment, sont difficilement comparables entre eux : par exemple, une ressemblance, même forte, entre sous-espaces factoriels peut être masquée par des rotations. Pour comparer les représentations des vins vus par chacun des groupes, il est nécessaire de les situer dans un référentiel commun. L'AFM répond à ce problème en fournissant une représentation factorielle dans laquelle figurent les représentations des vins décrits par chacun des groupes de variables considéré séparément. Son application à l'exemple est décrite section 1.5.

**En résumé**, la prise en compte d'une structure en groupes d'un ensemble de variables pose un problème technique (la pondération des groupes) et enrichit la problématique d'une étude (comparaison des groupes ; comparaison des typologies des vins définies par chaque groupe). L'AFM propose une solution technique (la pondération décrite dans la section suivante) au problème technique (équilibrer l'influence des groupes) et fournit des représentations adaptées aux différents aspects de l'objectif.

### 7.1.3 Pondération des groupes de variables

Deux éléments interviennent dans le rôle que peut jouer un groupe de variables dans une analyse d'ensemble :

- son inertie totale (égale au nombre de variables lorsqu'elles sont normées) ; plus cette inertie est importante, plus grande est l'influence du groupe ;
- la structure du groupe ; plus le groupe possède une structure forte, c'est-à-dire plus ses variables sont liées, et plus son influence sera déterminante dans la construction des principales dimensions de variabilité.

Dans la construction du premier axe d'une analyse globale, la direction principale d'inertie de chaque groupe joue un rôle majeur. Or, les inerties associées à ces directions (i.e. la première valeur propre des ACP séparées) peuvent être très variables. Dans l'exemple (*cf.* **Tableau 7.1**), les premières valeurs propres des groupes 3 et 4 sont beaucoup plus élevées que celles des groupes 1 et 2 : c'est là une conséquence du plus grand nombre de variables (et donc d'une inertie totale plus grande) pour ces groupes 3 et 4. Mais, dans le détail, la première valeur propre du groupe 2 est plus élevée que celle du groupe 1 et ce malgré un nombre de variables plus petit pour le groupe 2 : l'inertie totale du groupe 2 (3) est plus petite que celle du groupe 1 (5) mais est concentrée dans une direction (% d'inertie de l'axe $1 = 94,49$ %). Cet exemple illustre le fait qu'il n'est pas suffisant d'effectuer une normalisation des inerties totales.

**Tableau 7.1** Inerties des ACP séparées des quatre groupes actifs.

| Groupe | Inerties | | | | Pourcentages | | |
|---|---|---|---|---|---|---|---|
| | totale | axe 1 | axe 2 | axe 3 | axe 1 | axe 2 | axe 3 |
| 1 olfaction au repos | 5 | 2,24 | 1,52 | 0,82 | 44,84 | 30,33 | 16,31 |
| 2 vision | 3 | 2,83 | 0,15 | 0,01 | 94,49 | 5,03 | 0,48 |
| 3 olfaction après agitation | 10 | 4,70 | 2,48 | 1,05 | 47,01 | 24,83 | 10,46 |
| 4 gustation | 9 | 5,64 | 1,79 | 0,67 | 62,69 | 19,90 | 7,49 |

L'AFM équilibre l'influence des groupes de variables en donnant à chaque variable un poids. Ce poids doit être le même pour toutes les variables d'un même groupe afin de conserver la structure interne de chaque groupe.

Le poids donné par l'AFM à chacune des variables d'un groupe est égal à l'inverse de l'inertie de la première composante principale de ce groupe. Ainsi, dans l'exemple, les poids des variables de chacun de ces quatre groupes dans l'AFM sont respectivement 0.45, 0.35, 0.21 et 0.18.

Lorsque l'on affecte un même poids à toutes les variables d'un groupe, l'inertie du nuage associé est multipliée par ce poids dans chaque direction de l'espace. Avec le poids indiqué, l'inertie de la première composante principale de chaque groupe de variables est égale à 1 ; par suite, la somme des inerties des variables d'un même groupe sur un axe quelconque de l'espace est inférieure ou égale à 1. De cette façon, le rôle de chacun des groupes est équilibré en ce sens qu'aucun groupe ne peut influencer à lui seul la première composante principale de l'ensemble (qui maximise l'inertie projetée de toutes les variables). Cette pondération est une caractéristique majeure de l'AFM ; elle lui confère plusieurs propriétés qui apparaîtront par la suite.

L'AFM consiste d'abord en une analyse factorielle (ici une ACP normée) des variables ainsi pondérées.

Le **tableau 7.2** présente la décomposition, selon les quatre groupes, de l'inertie des trois premières composantes principales de l'AFM. Son interprétation, qui fait référence à la pondération des variables, en illustre les conséquences.

Pour la première composante principale, les inerties des variables de chacun des quatre groupes sont toutes assez proches entre elles : les quatre groupes contribuent de façons équilibrées à cette direction, ce qui était l'objectif de la pondération. En outre, ces inerties sont proches de la valeur maximum : 1. Du fait de la pondération des variables, la valeur 1 ne peut être atteinte que dans le cas extrême où cette composante est confondue avec la première composante principale de l'ACP séparée du groupe. Cette première composante principale est donc très liée à chacun des groupes en ce sens qu'elle constitue une direction d'inertie importante pour chacun.

**Tableau 7.2** Décomposition de l'inertie des trois premières composantes principales de l'AFM selon les quatre groupes.

| | 1e composante | 2e composante | 3e composante |
|---|---|---|---|
| inertie totale | 3.46 | 1.37 | .62 |
| 1 olfaction au repos | .78 | .62 | .37 |
| 2 vision | .85 | .04 | .01 |
| 3 olfaction après agitation | .92 | .47 | .18 |
| 4 gustation | .90 | .24 | .05 |

La situation est différente pour la deuxième composante principale. Le groupe 2 ne contribue pas à cette direction. Ce fait est à rapprocher du résultat suivant : l'ACP du seul groupe 2 conduit à une première composante principale associée à un pourcentage d'inertie de 0.94. Ainsi ce groupe est presque unidimensionnel alors que les autres groupes sont plus « riches », c'est-à-dire comportent au moins quelques variables peu corrélées entre elles. La pondération ne modifie pas cette structure inhérente aux données : les groupes riches influencent plus d'axes que les groupes pauvres.

### 7.1.4 Typologie des vins et principales dimensions

Les règles d'interprétation de ces premiers résultats sont identiques à celles d'une ACP. Compte tenu des pourcentages d'inertie (49,4 % pour l'axe 1 ; 19,5 %pour l'axe 2 et 8,8 % pour l'axe 3) nous limitons, dans cette présentation méthodologique, l'essentiel de l'interprétation au premier plan factoriel.

#### *a) Représentation des variables*

La **figure 7.2** montre que les variables les plus corrélées au premier facteur sont, pour chacun des groupes :

- **Olfaction au repos :** qualité globale des arômes, fruité.
- **Vision :** impression de surface, intensité, nuance (violacée).
- **Olfaction après agitation :** persistance aromatique, intensité olfactive rétronasale, qualité globale des arômes.
- **Gustation :** intensité fin de bouche, harmonie, intensité d'attaque, velouté.
- **Jugement d'ensemble** : qualité globale, typicité.

Ce premier axe recouvre des notions classiquement (dans le monde du vin) regroupées dans les mots « puissance » et « harmonie » qui possèdent des connotations nettement positives. Ces deux termes ne sont absolument pas synonymes en général, mais sont, pour la population de vins étudiés ici, très liés.

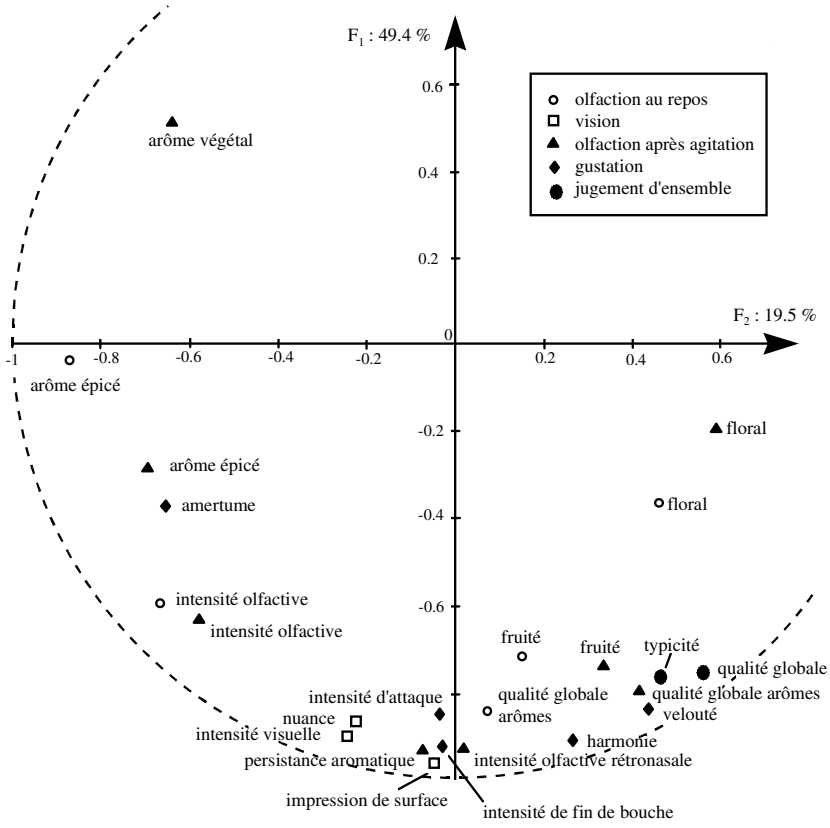



**Figure 7.2** Représentation des variables (par leur corrélation avec les axes) sur le premier plan factoriel.

Les variables les plus corrélées au second facteur sont :

- **Olfaction au repos :** épicé, intensité olfactive.
- **Olfaction après agitation :** épicé, végétal, intensité olfactive.
- **Gustation :** amertume.

Ce deuxième axe est lié à la reconnaissance d'une particularité perçue de façon intense par de nombreux juges comme la caractéristique *épicé* ou *végétal*. Il correspond à une particularité essentiellement olfactive (elle n'est associée qu'à une seule caractéristique provenant d'un autre sens : l'amertume).

*b) Représentation des vins (cf. Figure 7.3)*

Le premier axe étant interprété comme un axe de « puissance et harmonie », la coordonnée d'un individu s'interprète en ces termes. Ainsi, le vin 1DAM a été considéré comme le plus « puissant et harmonieux ». À l'opposé, les vins 1VAU et 2ING, possédant les coordonnées les plus fortes, ont été perçus comme particulièrement peu « puissants et harmonieux ». Ces deux derniers vins se démarquent franchement des autres le long de cette dimension à laquelle ils contribuent pour $32.8 + 26.4 = 59.2$ %.

Le deuxième axe est essentiellement dû aux deux vins notés Smi4 (contribution de ces deux vins à cet axe : 29.7 % + 39.3 % = 69 %). Il s'agit en fait du même vin présenté deux fois aux dégustateurs. On interprète donc cet axe comme le « cas particulier du vin Smi4 ».

En outre, dans une question ouverte relative à l'olfaction, ce vin a été très souvent (8 fois pour l'un, 9 fois pour l'autre) associé à *sous-bois* et/ou *champignon*, termes très peu cités pour les autres vins. Ces données renforcent l'interprétation de cet axe en tant que particularité olfactive du vin Smi4. Remarquons au passage que le fait que les dégustateurs aient jugé de la même façon les deux échantillons provenant du même vin est un bon argument en faveur de la valeur des données.

➤ Relation entre les deux premiers facteurs et l'origine du vin

Chaque modalité d'une variable qualitative est représentée au centre de gravité des individus qui la possèdent. À chaque coordonnée d'une modalité sur un axe, on associe une valeur-test (*cf.* § 2.4.4 page 54).

Les modalités *Saumur*, *Chinon* et *Bourgueil* sont très proches de l'origine des axes (valeurs-test $< 1.4$) : l'origine du vin, au sens de l'appellation, est sans rapport avec les principales dimensions de variabilité de ces vins. La modalité *milieu 4* est très éloignée le long de l'axe 2 (valeurs-test $= 3.9$) mais elle ne concerne que deux vins. La modalité *séquence de référence* est fortement éloignée le long de l'axe 1 (valeur-test $= 2.4$) ; rappelons qu'elle correspond à un type de sol qui, d'un point de vue agronomique, possède une excellente potentialité viticole ; cet *a priori* est confirmé par la place de cette modalité sur le plan.

### 7.1.5 Représentation superposée des vins décrits par chaque groupe de variables séparément

La représentation des vins décrite précédemment s'appuie sur l'ensemble des variables des quatre groupes. Un vin comme 1DAM, qualifié de « puissant et harmonieux », occupe une position extrême sur la première composante car les notes qui lui ont été attribuées pour les variables corrélées à cette composante sont en général très élevées. Un retour au tableau de données permet de préciser ce qu'il en est pour chacune d'entre elles, à savoir si la grande « puissance et harmonie » du vin 1DAM s'exprime de façon homogène sur l'ensemble des variables ou si, au contraire, 1DAM présente

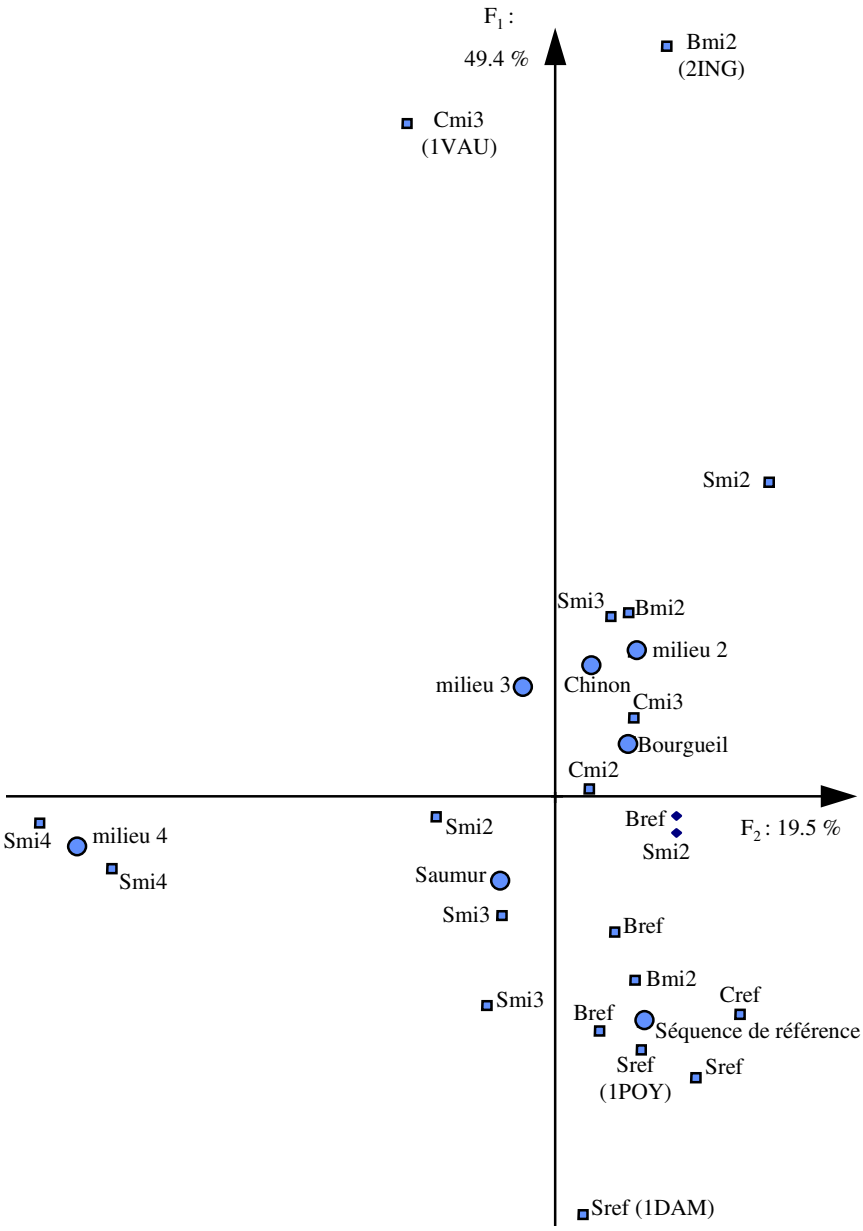



**Figure 7.3** Représentation du nuage des vins vus par l'ensemble des variables. L'emplacement d'un vin est repéré par l'initiale de l'appellation [S = Saumur ; B = Bourgueil ; C = Chinon] suivi de sa modalité de milieu [ref = séquence de référence, mi2 = milieu 2, etc.]. Quatre vins, commentés en détail par la suite, ont un nom particulier [1DAM, 1POY, 1VAU et 2ING]. En outre, le point moyen de chaque modalité est représenté.

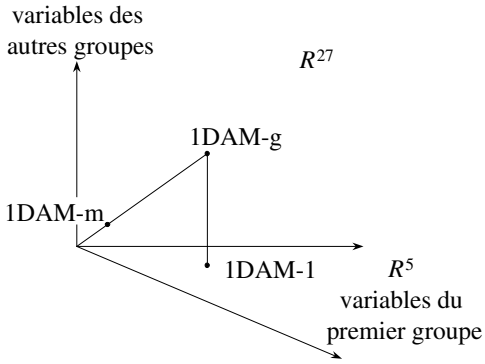



**Figure 7.4** Individu global, partiel et moyen. La représentation du vin 1DAM du point de vue du seul groupe 1 (individu partiel 1DAM-1) est obtenue en projetant le point global (1DAM-g $\in R^{27}$ engendré par toutes les variables) sur le sous-espace (noté $R^5$) engendré par les 5 variables du groupe 1. Le point moyen 1DAM-m est obtenu à partir de 1DAM-g par une homothétie de rapport $1/J = 1/4$.

des points plus ou moins forts. Cette question peut être posée, non plus en termes de variables mais de groupes de variables : la puissance du vin s'exprime-t-elle de façon égale ou inégale dans son aspect visuel, ses parfums, son goût ?

La perception de chaque aspect n'est pas lisible facilement sur les données puisque chacun est mesuré par un groupe de variables. Il est donc utile de disposer d'un outil qui synthétise la perception des vins non plus du point de vue de l'ensemble des variables mais du point de vue de chacun des groupes de variables. Pour cela, en AFM, on s'appuie sur la représentation géométrique suivante.

Remarquons tout d'abord que, dans l'ACP d'un seul groupe de variables, on définit un nuage qui représente l'ensemble des vins perçu à l'aide de ce seul groupe. On dispose ainsi de quatre nuages des vins, dits **nuages partiels**, correspondant chacun à un groupe actif (l'exposé théorique de l'AFM, chapitre 8, montre pourquoi on se limite ici aux groupes actifs).

Ces quatre nuages partiels peuvent être construits dans l'espace de dimension 27 (noté $R^{27}$) engendré par toutes les variables actives. Le nuage *olfaction au repos* est obtenu en projetant le nuage « global » des vins sur le sous-espace de dimension 5 (noté $R^5$) engendré par les cinq premières variables, c'est-à-dire, puisque les variables sont centrées, en annulant toutes les coordonnées qui ne concernent pas cet aspect (*cf.* **Figure 7.4**).

En AFM, on traite comme des lignes supplémentaires les quatre tableaux obtenus en annulant les valeurs des variables (centrées) de trois groupes sur 4. Ainsi, les quatre nuages des vins définis séparément par chacun des quatre aspects mesurés, sont projetés sur les axes factoriels du nuage regroupant ces différents aspects. On

**Tableau 7.3** Quelques valeurs pour les trois variables du groupe olfaction au repos les plus corrélées au premier facteur.

| | qualité globale des arômes | note fruitée | intensité olfactive |
|---|---|---|---|
| maximum | 3.429 | 3.154 | 3.708 |
| 1DAM | 3.429 | 3.154 | 3.607 |
| 1POY | 3.107 | 2.731 | 3.071 |
| moyenne | 3.046 | 2.714 | 3.111 |

obtient ainsi une représentation superposant ces quatre nuages « partiels » au nuage global. En AFM, une homothétie (qui ne modifie strictement pas la forme du nuage) est appliquée au nuage global pour obtenir un nuage « moyen » ; elle met chaque point de ce nuage (par exemple 1DAM) au barycentre des 4 points (1DAM1, 1DAM2, 1DAM3 et 1DAM4) décrivant ce même vin dans ses différents aspects. La lecture des graphiques en est grandement facilitée : il est beaucoup plus rapide de comparer chaque point au barycentre que de comparer les quatre points deux à deux. Le chapitre suivant montre plusieurs propriétés de ces graphiques.

La **figure 7.5** est un extrait de cette représentation superposée appliquée à l'exemple. Elle est limitée, pour des raisons de clarté, à 6 vins extrêmes (*cf.* **Figure 7.3**) :

- 1DAM et 1POY, jugés globalement les plus puissants et harmonieux ;
- 1VAU et 2ING, jugés globalement les moins puissants et harmonieux ;
- le vin Smi4, présenté en double, bien individualisé sur le deuxième axe.

D'après la **figure 7.5**, le vin 1DAM a été perçu comme le plus « puissant et harmonieux » du point de vue de l'olfaction au repos (*cf.* la position extrême de 1DAM1). Par contre, pour ce même groupe de variables, le vin 1POY a été perçu seulement comme un peu plus que moyen (*cf.* la position relativement centrale de 1POY1). Ces informations se retrouvent facilement dans les données (*cf.* **Tableau 7.3**).

La situation est différente du point de vue de la gustation pour laquelle c'est 1POY qui a été perçu comme le plus « puissant et harmonieux ». L'écart est, selon ce sens, moins important que pour l'olfaction. Ces informations se lisent aussi dans les données (*cf.* **Tableau 7.4**).

### 7.1.6 Facteurs communs

Pour mesurer la similitude entre les projections des quatre nuages partiels sur un même axe, on calcule le coefficient de corrélation entre chacune de ces projections et celle du nuage global. Le **tableau 7.5** contient ces valeurs pour les premiers axes.

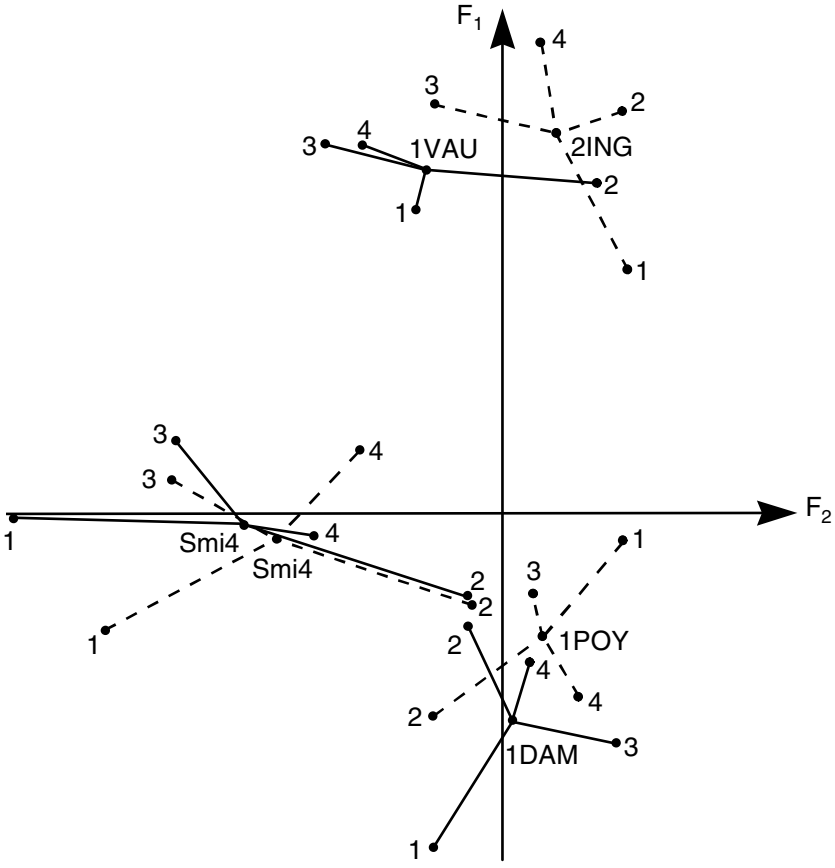



**Figure 7.5** Représentation superposée limitée à 6 vins extrêmes. Chaque vin est caractérisé par chacun des quatre groupes de variables et par l'ensemble des groupes.

**Tableau 7.4** Quelques valeurs pour les quatre variables du groupe gustation les plus corrélées au premier facteur.

| | velouté | intensité fin de bouche | harmonie | intensité d'attaque |
|---|---|---|---|---|
| maximum | 3.286 | 3.676 | 3.786 | 3.519 |
| 1POY | 3.231 | 3.667 | 3.786 | 3.519 |
| 1DAM | 3.036 | 3.643 | 3.643 | 3.464 |
| moyenne | 2.674 | 3.166 | 3.148 | 3.156 |

**Tableau 7.5** Corrélations, pour les sept premiers axes de l'AFM, entre la projection du nuage global et celle de chacun des quatre nuages partiels (i.e. associés à un seul groupe).

| | axe 1 | axe 2 | axe 3 | axe 4 | axe 5 | axe 6 | axe 7 |
|---|---|---|---|---|---|---|---|
| olfaction au repos | .89 | .96 | .89 | .48 | .42 | .27 | .42 |
| vision | .93 | .22 | .16 | .22 | .17 | .08 | .21 |
| olfaction après agitation | .97 | .89 | .90 | .57 | .66 | .49 | .46 |
| gustation | .95 | .87 | .30 | .25 | .52 | .56 | .42 |

Pour synthétiser le **tableau 7.5**, on dira que la première composante de l'AFM, *puissance et harmonie*, est un « facteur commun » aux quatre groupes de variables. En effet, les valeurs assez élevées des quatre coefficients de corrélation de la première colonne indiquent que les projections des quatre nuages partiels sont presque des homothéties de la projection du nuage global (qui est la moyenne entre ces projections). Cette première colonne permet d'affirmer qu'il existe une direction de dispersion presque analogue dans les quatre nuages.

La composante, *cas particulier du vin Smi4*, est une dimension qui se traduit uniquement dans les parfums et le goût puisque seule la corrélation entre la projection du nuage global et celle du nuage défini par les variables visuelles est très faible (0.22). Le second facteur est commun à trois groupes seulement.

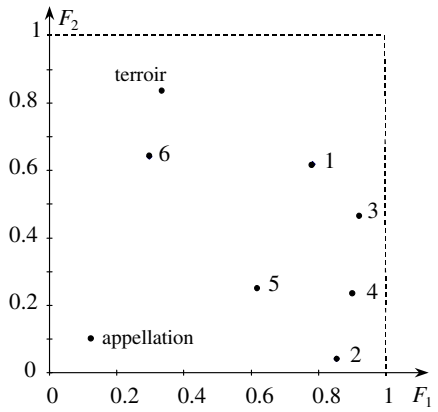
La troisième composante de l'AFM est un facteur commun aux deux olfactions seulement. Ceci incite à examiner cette composante dont nous dirons seulement qu'elle oppose, pour chaque olfaction, les caractères *fruité* et *floral* (ces 4 variables contribuent à 73 % de l'inertie de ce facteur).

Il n'existe pas (actuellement) de seuil à partir duquel on pourrait dire que tel coefficient de corrélation est grand. Ce seuil dépendrait du nombre d'individus, de la dimensionalité des groupes, etc. En l'absence, ces coefficients se servent mutuellement de référence, raison pour laquelle ils sont donnés pour les sept premiers axes.

Ces corrélations permettent donc de juger de l'existence d'un facteur commun à tous les groupes ou à certains d'entre eux. Lorsque cette direction de dispersion commune existe, il est intéressant de mesurer et de comparer son importance dans les différents groupes. L'importance d'un facteur dans un groupe est mesurée par l'inertie cumulée des variables du groupe sur ce facteur (*cf.* **Tableau 7.2**).

### 7.1.7 Représentation des groupes de variables

Les quatre groupes actifs de l'analyse ainsi que les deux groupes supplémentaires sont représentés sur un graphique qui correspond axe par axe aux graphiques des variables et des vins (*cf.* **Figure 7.6**). La coordonnée d'un groupe sur un axe est l'inertie cumulée des variables du groupe sur l'axe correspondant de l'AFM (*cf.* **Tableau 7.2**).




Qualité de représentation des groupes (inertie projetée / inertie totale)
1 olfaction au repos : 0.617
2 vision : 0.729
3 olfaction après agitation : 0.784
4 gustation : 0.772
5 appréciation d'ensemble : 0.444
6 origine : 0.189

**Figure 7.6** Représentation des groupes de variables : carré des liaisons.

La pondération (appliquée aussi aux variables des groupes supplémentaires) implique que les coordonnées d'un point-groupe sont comprises entre 0 et 1.

Ce graphique admet plusieurs interprétations listées ci-après.

### *a) Contributions des groupes à la construction des axes*

Les quatre groupes actifs ont des coordonnées voisines le long du premier axe : ils contibuent également au premier facteur (puissance et harmonie). Les coordonnées des groupes le long du deuxième axe montrent que le deuxième facteur est dû principalement à l'olfaction (groupes 1 et 3) et légèrement à la gustation (groupe 4).

### *b) Liaison entre les composantes principales de l'AFM et les groupes*

La coordonnée d'un groupe sur un axe peut être considérée comme une mesure de la liaison entre le groupe et le facteur correspondant : si cette coordonnée est proche de 0, les variables du groupe ne sont pas corrélées au facteur ; si elle est proche de 1, le facteur correspond à une direction d'inertie importante (voisine du maximum) pour le groupe de variables. Ainsi, le premier facteur est une direction d'inertie très importante pour les quatre groupes actifs et, en ce sens, leur est très lié. Le deuxième facteur a une importance presque aussi grande que le premier pour un seul groupe actif : l'olfaction au repos. Le sixième groupe (origine des vins), traité en illustratif, est beaucoup plus lié au deuxième facteur qu'au premier.

L'interprétation de ce graphique en terme de liaisons lui vaut le nom de carré des liaisons. Il a déjà été vu en ACM (*cf.* 4.3.7) ce qui suggère de représenter aussi les deux variables du groupes 6 séparément (par le carré de leurs rapports de corrélation) : on visualise ainsi l'indépendance entre l'appellation et les deux premières composantes

de l'AFM, la liaison forte entre le terroir et la deuxième composante (qui individualise presque parfaitement la modalité *milieu 4*) et la liaison faible (mais significative, *cf.* § 7.1.4) entre le terroir et la première composante (qui distingue les vins de la séquence de référence).

#### *c) Représentation optimale du nuage des groupes*

Ce graphique s'interprète aussi comme la projection orthogonale d'un nuage de points représentant chacun un groupe. Dans ce nuage, précisé au chapitre suivant, deux groupes sont d'autant plus proches que les structures qu'ils définissent sur l'ensemble des vins (c'est-à-dire les quatre nuages partiels de la section 7.1.5) se ressemblent.

Sur ce plan, les deux groupes supplémentaires (5 et 6) sont mal représentés. Le groupe 6 est éloigné des autres : l'origine des vins (appellation et terroir) est, dans l'ensemble, moins liée à leurs principales caractéristiques organoleptiques que ces caractéristiques ne sont liées entre elles.

Les groupes les plus proches entre eux sont les deux olfactions : ces deux groupes sont proches (entre eux et des autres) du point de vue de la *puissance et harmonie* ; en outre, ce sont surtout eux qui mettent en évidence le cas particulier du vin Smi4.

### 7.1.8 Projections des composantes principales de chaque groupe

La **figure 7.7** représente la projection des deux premières composantes principales normées de chacun des 5 groupes sensoriels sur le plan des deux premières composantes de l'AFM (concrètement, le programme réalise une ACP de chaque groupe et traite les composantes ainsi obtenues comme des variables supplémentaires).

La première composante de l'AFM (*puissance et harmonie*) est très corrélée à la première composante principale de chaque groupe actif. Nous avons déjà indiqué que les quatre groupes actifs possèdent une direction de dispersion commune d'inertie importante. Il apparaît ici que ce facteur commun est, pour ces quatre groupes, proche de leur principale direction de dispersion.

La deuxième composante principale de l'AFM (cas particulier du vin Smi4) est très liée à la deuxième composante principale des trois groupes : *olfaction au repos*, *olfaction après agitation*, et *gustation*. Nous avons déjà conclu à l'existence d'un second facteur commun à ces trois groupes. Ce nouveau résultat précise l'importance relative de ce facteur dans les groupes concernés.

### 7.1.9 Conclusion sur l'exemple des vins

L'AFM prend en compte la structure en groupes de variables à deux niveaux. Elle pondère les variables de façon à équilibrer l'influence des groupes dans l'analyse globale ; de façon indirecte, cette pondération enrichit la signification d'indicateurs (*e.g.* l'inertie projetée des variables d'un groupe) et donc facilite l'interprétation.

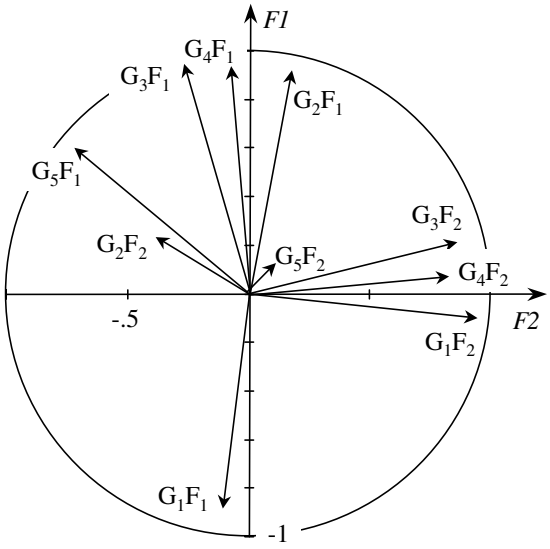



**Figure 7.7** Représentation des deux premières composantes principales de chaque groupe par leur corrélation avec les deux premières composantes de l'AFM. $G_1F_2$ = deuxième composante du groupe 1.

Elle fournit des représentations spécifiques de la structure en groupes (représentation superposée des nuages partiels et représentation des groupes).

## 7.2 AFM APPLIQUÉE AUX DONNÉES DE L'ENQUÊTE *OUEST-FRANCE*

Dans l'exemple des vins, les variables des groupes actifs sont des variables numériques ; dans l'enquête *Ouest-France* commentée dans le chapitre précédent, ce sont des variables qualitatives. Le fait que les variables soient qualitatives et non numériques ne modifie fondamentalement ni la problématique ni les solutions proposées en AFM : il suffit de remplacer la notion d'ACP par celle d'ACM pour les groupes composés de variables qualitatives.

Une AFM appliquée aux données de l'enquête *Ouest-France*, dans laquelle les variables actives sont subdivisées en deux groupes (*lecture* et *signalétique*), permet d'obtenir :

– une représentation optimale des individus dans laquelle leur profil de lecture et leur profil de signalétique interviennent de façon équilibrée ;

- une représentation optimale conjointe des modalités des variables du signalétique et de celles de lecture des rubriques ;
- la réponse à la question : existe-t-il ou non des « facteurs communs » aux deux groupes (*i.e.* des directions de dispersion analogues dans les deux nuages d'individus définis respectivement par le signalétique et les rubriques) ?
- une représentation superposée des nuages d'individus définis par chaque groupe de variables, outil commode notamment pour détecter des individus dont le profil de lecture ne correspond pas à leur signalétique ;
- la réponse à la question : existe-t-il ou non des « facteurs spécifiques » de l'un des groupes (*i.e.* une direction de dispersion d'un des deux nuages qui n'apparaît pas ou peu dans l'autre) ?
- une mesure de l'importance relative, pour chaque groupe, des directions communes ou spécifiques ;
- la comparaison des premiers facteurs de l'analyse séparée du signalétique et de celle des rubriques ;
- une représentation graphique d'un nuage de points représentant chacun un groupe sur des axes correspondant aux axes factoriels des nuages d'individus et de modalités (peu utile dans cet exemple avec deux groupes seulement).

Compte tenu des commentaires des analyses classiques de cette enquête (*cf.* Chapitre 6), la question principale est la suivante : qu'y a-t-il de commun ou de spécifique entre la lecture du journal et le profil signalétique global ? Nous portons donc notre attention, facteur par facteur, sur les indices de comparaison des groupes. Enfin, en harmonie avec les ACM du chapitre 6, les modalités correspondant aux données manquantes sont laissées telles quelles (remarquons au passage que les quatre possibilités de gestion des données manquantes évoquées en ACM valent en AFM).

### 7.2.1 Premier facteur : les modalités

**L'inertie** du premier facteur vaut 1.57. La valeur maximum possible est 2 (cas où le facteur de l'AFM est confondu avec le premier facteur de chaque groupe) et la valeur minimum 1 (inertie maximum de chaque groupe sur un axe). On en conclut que ce facteur n'est pas la principale direction de dispersion des nuages du signalétique et des rubriques (qui ne sont donc pas confondues), mais représente une direction de dispersion qui apparaît dans les deux nuages. Cette remarque est confirmée et précisée par les autres indices.

Les **corrélations** entre ce premier facteur et les projections associées de chacun des deux nuages d'individus définis séparément par le signalétique et les rubriques valent respectivement 0.921 et 0.909 : ces fortes valeurs indiquent qu'il s'agit d'un **facteur commun** aux deux groupes.

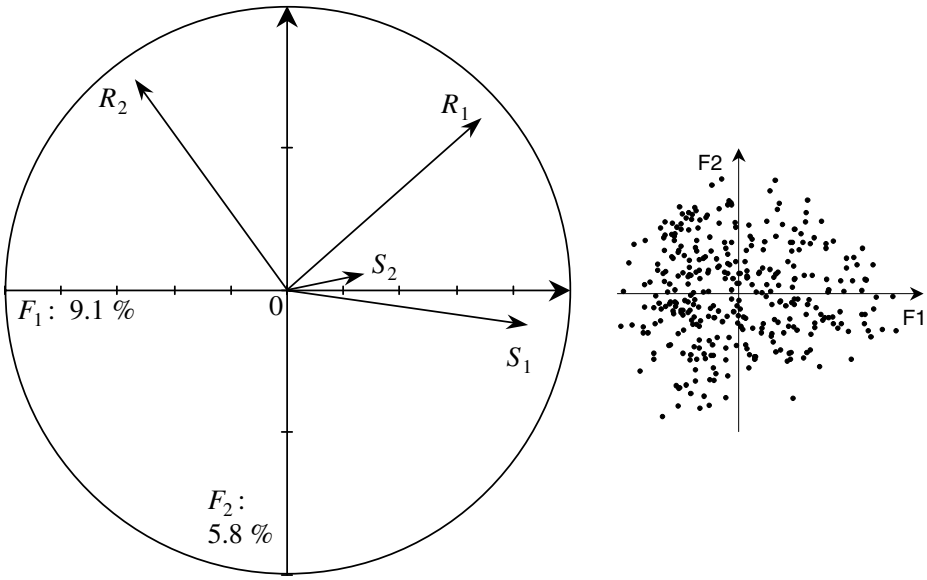



**Figure 7.8** Représentation des deux premiers facteurs des analyses séparées de chaque groupe sur le cercle des corrélations du premier plan de l'AFM. $R_1$ : premier facteur de l'analyse des Rubriques. $S_1$ : premier facteur de l'analyse du Signalétique. À gauche, allure du nuage des individus.

Sur ce premier facteur, l'**inertie de l'ensemble des variables du signalétique** vaut 0.816 tandis que celle de l'ensemble des rubriques vaut 0.752. Ces valeurs sont sensiblement inférieures à la valeur maximum 1 : cette direction de dispersion, commune aux deux groupes, ne se confond pas avec le premier facteur de chaque groupe. Cependant, ces deux valeurs sont suffisamment grandes pour conclure que cette direction de dispersion est assez importante dans les deux groupes.

Les **corrélations entre ce facteur et ceux des deux ACM séparées** ne sont jamais très élevées. Le graphique (*cf.* **Figure 7.8**) représentant les projections des deux premiers facteurs normés des deux ACM (séparées) sur le plan des deux premiers facteurs de l'AFM montre les relations entre les facteurs de ces trois analyses.

Les plans engendrés par les deux premiers facteurs, de l'ACM des rubriques d'une part et de l'AFM d'autre part, sont pratiquement confondus. Le premier facteur de l'AFM n'est autre que la seconde bissectrice du plan 1-2 commenté dans la section 6.4 page 140 (*cf.* Figures 6.1 et 6.2). Cette direction du plan, le long de laquelle s'étagent les cinq niveaux d'instruction et plusieurs autres variables du signalétique, oppose les individus ayant un profil de lecture « intellectuel » (lecture de ce type de rubriques et non-lecture des rubriques anecdotiques) aux individus ayant un profil de lecture inverse. Cette répartition des rubriques et du signalétique est aussi très proche de celle

du premier facteur du tableau croisé qui permet d'analyser leur liaison (*cf.* section 6.7 page 143).

### 7.2.2 Deuxième facteur : les modalités

L'inertie de ce facteur vaut 0.978. Les **coefficients de corrélation** entre ce facteur et la projection associée des deux nuages définis l'un par le signalétique et l'autre par les rubriques valent 0.546 pour le premier et 0.964 pour le second. C'est donc une direction de dispersion des profils de lecture qui n'est pas vraiment liée au signalétique (un cosinus de 0.546 correspond à un angle de 57 degrés). Ce facteur est donc spécifique des profils de lecture.

Sur ce deuxième facteur, l'**inertie cumulée** des variables du signalétique (0.147) et celle des variables de lecture des rubriques (0.825) confirment la prépondérance des variables du deuxième groupe pour lequel, d'ailleurs, cette direction spécifique est un peu plus importante que la direction commune exprimée par le premier facteur.

La **figure 7.8** montre que ce facteur est très proche de la première bissectrice du plan 1-2 de l'ACM des rubriques, direction liée au nombre total de rubriques lues.

### 7.2.3 AFM et ACM des rubriques

Nous ne donnons pas le graphique des projections des modalités des variables sur le plan 1-2 car il se déduit, à très peu de choses près, du plan de l'ACM des rubriques par une rotation $3\pi/4$. Cette coïncidence entre l'analyse d'un groupe et l'AFM est rare lorsque les groupes étudiés sont assez différents entre eux.

Bien que les graphiques soient à peu près identiques, l'interprétation de l'AFM est différente de celle de l'ACM des rubriques puisque l'on se réfère aux deux groupes qui sont tous deux actifs. Le premier facteur étant un facteur commun et le deuxième étant spécifique des profils de lecture, on s'attache à chaque axe séparément plutôt qu'au plan. Rappelons que ces deux axes ne sont pas confondus avec ceux de l'ACM des rubriques mais avec leurs bissectrices.

Les variables les plus liées au premier facteur, et donc liées entre elles, sont d'une part le niveau d'instruction et l'âge et, d'autre part, les rubriques les plus anecdotiques ainsi que les plus intellectuelles. La dispersion des profils de lecture mise en évidence par le deuxième facteur, donc spécifique de la lecture, est l'opposition entre *rubrique-lue* et *rubrique non-lue*.

### 7.2.4 Représentation superposée des individus

La représentation superposée des individus caractérisés d'une part par leur signalétique et d'autre part par leur lecture est un résultat de l'AFM sans équivalent dans les méthodes classiques. Elle n'est intéressante que sur une dimension commune, ici le

premier facteur. La plupart des individus sont représentés par deux points très proches ainsi que l'indique le rapport [*inertieinter*/*inertietotale*] dont la valeur pour ce facteur est 0.837. Ce rapport se réfère au nuage des individus vu par chacun des deux groupes ($680 = 340 \times 2$ points), partitionné en 340 groupes (1 groupe = 1 enquêté) de 2 points chacun (le même enquêté caractérisé par chacun des deux groupes) : il vaut 1 si les 2 images de chaque enquêté coïncident entre elles et donc avec leur centre de gravité. Comme nous ne nous intéressons pas à chaque individu mais plutôt à ce qu'il représente, nous étudions seulement les représentations superposées des barycentres des classes définies par les modalités de toutes les variables. La **figure 7.9** donne un extrait de cette représentation superposée en rappelant l'interprétation générale de ce facteur et la projection des 5 niveaux d'instruction. De cette interprétation on déduit deux cas de figure.

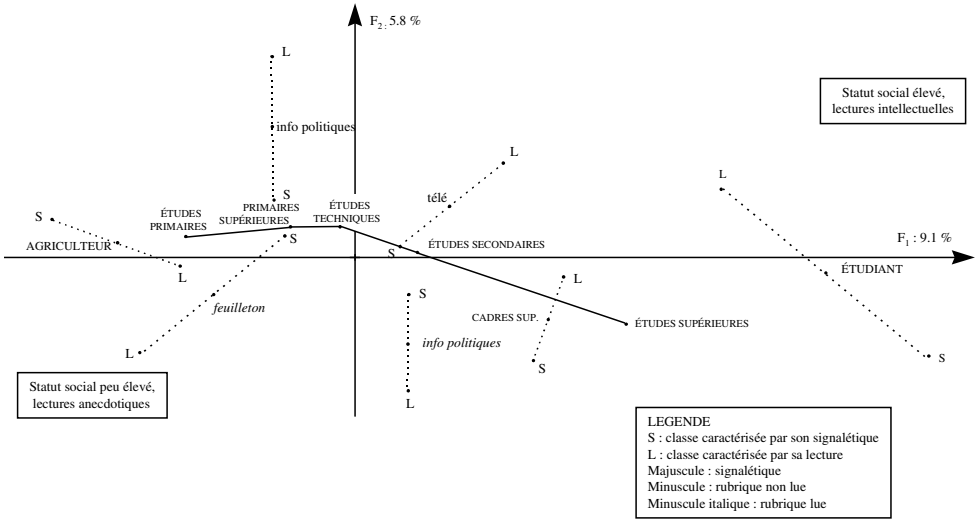



**Figure 7.9** Un extrait de la représentation superposée de l'AFM de l'enquête *Ouest-France*.

– Un individu dont le profil de lecture est situé sensiblement plus à droite que son signalétique a un profil de lecture plus « intellectuel » que ne le laisse présager son signalétique ou, ce qui est équivalent, un statut social moins élevé que ne le laisse présager sa lecture du journal.
– Inversement, un individu dont le profil de lecture est situé sensiblement plus à gauche que son signalétique a un profil de lecture moins « intellectuel » que ne le laisse présager son signalétique ou, ce qui est équivalent, un statut social plus élevé que ne le laisse présager sa lecture du journal.

Pour les barycentres l'interprétation est analogue. De la grande proximité entre les deux points représentant un même individu, découle une grande proximité entre les deux

points représentant leurs barycentres. On peut commeter cependant deux exceptions concernant une modalité du signalétique et une modalité de lecture : les étudiants et les lecteurs du feuilleton.

- Les étudiants sont très extrêmes par leur signalétique ; ils cumulent des modalités caractéristiques des lectures « intellectuelles » : jeune, niveau d'instruction supérieur, habitant la zone résidentielle, etc. Leur profil de lecture, beaucoup plus moyen, est moins intellectuel qu'il ne le devrait !
- Les lecteurs du feuilleton sont beaucoup plus extrêmes par leur lecture (peu intellectuelle) du journal que par leur signalétique relativement moyen. On caractérise donc un lecteur du feuilleton plus à sa lecture des autres rubriques du journal qu'à son signalétique.

Chapitre **8**

# Aspects théoriques et techniques de l'Analyse Factorielle Multiple

À l'aide de deux exemples, le chapitre précédent décrit les grandes lignes de la problématique de l'étude des tableaux multiples ainsi que les principaux résultats de l'AFM. Dans cette première présentation, les considérations théoriques et techniques sont réduites au minimum. Nous reprenons ici l'exposé de l'AFM en faisant toujours référence à l'exemple des vins pour illustrer les objectifs mais en détaillant les calculs ainsi que leurs justifications.

Dans un premier temps, nous adoptons successivement comme cadre les trois espaces dans lesquels l'AFM peut être présentée :

- $R^K$, dans lequel sont situés les nuages des individus ;
- $R^I$, dans lequel est situé le nuage des variables ;
- $R^{I^2}$, dans lequel est situé le nuage des groupes de variables.

Dans un second temps, nous fournissons des compléments qui concernent :

- une autre présentation de la méthode : l'estimation des paramètres du modèle INDSCAL ;
- le cas des variables qualitatives ;
- la mise en œuvre.

## 8.1 DONNÉES ET NOTATIONS

Par souci de clarté, nous restreignons d'abord l'exposé au cas des variables numériques ; la prise en compte des variables qualitatives est étudiée séparément (*cf.* section 8.6 page 197). Comme en ACP, les variables quantitatives sont toujours centrées et, sauf mention explicite du contraire, réduites.

Les données sont constituées par un ensemble d'individus décrits par plusieurs groupes de variables. À 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 en ligne et former ainsi un seul tableau croisant individus et 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 des sous-tableaux ; $K_j$ l'ensemble des variables du groupe $j$ ; ($K = \cup_j K_j$) ; $X_j$ le tableau associé au groupe $j$ (*cf.* **Figure 8.1**).

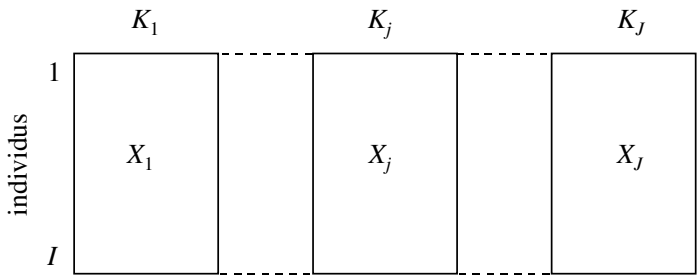



**Figure 8.1** Tableau X des données en AFM.

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 $i$ (avec $\sum_i p_i = 1$) et $m_k$ 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 $K_j$) et $M$ (pour $K$). Il faut distinguer le poids des variables dans les analyses séparées des groupes de variables et dans l'analyse d'ensemble.

**Dans les analyses séparées**, les variables numériques ont presque toujours le poids 1. Il est toutefois possible d'affecter des poids différents d'une variable à l'autre : cette situation se rencontre notamment lorsque les variables sont des facteurs normés issus d'une autre analyse, auquel cas on affecte à une variable-facteur un poids égal à l'inertie à laquelle elle est associée (autre solution : on analyse les facteurs non normés chacun affecté du poids 1).

**Dans l'analyse d'ensemble**, ainsi que cela a été présenté dans l'exemple des vins, les poids initiaux des variables sont modifiés. Le poids initial de chaque variable du groupe $j$ est divisé par $\lambda_1^j$ (en notant $\lambda_1^j$ la première valeur propre de l'analyse factorielle séparée du groupe $j$).

Nous gardons la même notation $m_k$ pour le poids de la variable $k$ quel que soit le stade de l'analyse : pratiquement, du fait du contexte, il n'en résulte aucune ambiguïté. De même, nous gardons une seule notation $\lambda_s^j$ pour la valeur propre de rang $s$ associée à l'ACP de $X_j$ avant ou après la pondération ($\lambda_1^j$ vaut 1 après la pondération).

Cette pondération a pour but d'équilibrer le rôle des groupes dans tous les aspects de l'analyse. Elle est interprétée dans tous les espaces dans lesquels l'AFM est présentée.

## 8.2 L'AFM DANS L'ESPACE DES INDIVIDUS $R^K$

L'espace $R^K$ contient les représentations des individus. Chacune de ses dimensions est associée à une variable. À partir de cet espace, nous cherchons deux types de représentation.

1. Une représentation du nuage des individus caractérisés par l'ensemble des variables. L'exemple des vins a montré que cette représentation était obtenue par une ACP du tableau $X$, les variables étant pondérées.
2. Une représentation superposée des $J$ nuages d'individus caractérisés chacun par un seul groupe de variables. Dans l'exemple des vins, cette représentation faisait figurer sur un même graphique les vins du point de vue de l'olfaction au repos, de la gustation, etc. Ce graphique est obtenu à l'aide de projections de lignes supplémentaires dans l'ACP précédente.

### 8.2.1 Influence de la pondération des groupes sur les $J$ nuages $N_I^j$

À chaque groupe de variables $j$, correspond un nuage représentant les individus. Ce nuage noté $N_I^j$ est situé dans un espace de dimension $K_j$ noté $R^{Kj}$.

Rappelons que la pondération, qui vise à équilibrer le rôle des groupes de variables, revient à diviser par $\lambda_1^j$ le poids initial de chaque variable du groupe $j$. Ce coefficient, étant identique pour toutes les variables du groupe $j$, ne modifie pas la forme des nuages $N_I^j$. En revanche, il normalise ces nuages en ce sens que, avec ces poids, l'inertie maximum de tout nuage $N_I^j$ dans une direction quelconque vaut 1 (de façon équivalente : le premier axe de l'ACP du seul nuage de $N_I^j$ est alors associé à une valeur propre de 1). Enfin, avec cette pondération, deux nuages homothétiques deviennent égaux.

### 8.2.2 Influence de la pondération des groupes sur le nuage $N_I$ associé à toutes les variables

À l'ensemble de toutes les variables, correspond un nuage représentant les individus situé dans l'espace $R^K$. 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 $N_I^j$. Notons $i^j$ le point représentant $i$ dans le nuage $N_I^j$ et $v_k(i)$ la valeur de la variable $k$ pour $i$. Alors :

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

Dans la distance entre deux éléments du nuage $N_I$, l'influence des différents groupes n'est équilibrée que si les distances dans les différents nuages $N_I^j$ sont du même ordre de grandeur. Multiplier les poids initiaux des variables du groupe $j$ par un coefficient $\alpha_j$ est un moyen d'équilibrer l'influence des groupes puisque la distance s'écrit alors :

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

Avec la pondération $\alpha_j = 1/\lambda_1^j$, aucun groupe ne peut être prépondérant dans la première direction d'inertie du nuage moyen. Cependant, le nombre de directions de $N_I$ sur lesquelles le groupe $j$ influe croît avec la dimensionalité de $N_I^j$.

### 8.2.3 Représentation des $J$ nuages $N_I^j$ dans $R^K$ et nuage moyen

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 $J$ sous-espaces orthogonaux deux à deux et isomorphes aux espaces $R^{Kj}$. Soit :

$$R^K = \oplus R^{Kj}$$

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 de dimensions $I$ et $K$, dans lequel $X_j$ est complété par des zéros (*cf.* **Figure 8.2**) ; ce tableau est noté $\tilde{X}_j$.

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 $R^K$.

Soit $N_I^*$ 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 $1/J$. Le nuage $N_I^*$ est un nuage moyen pour les $N_I^j$.

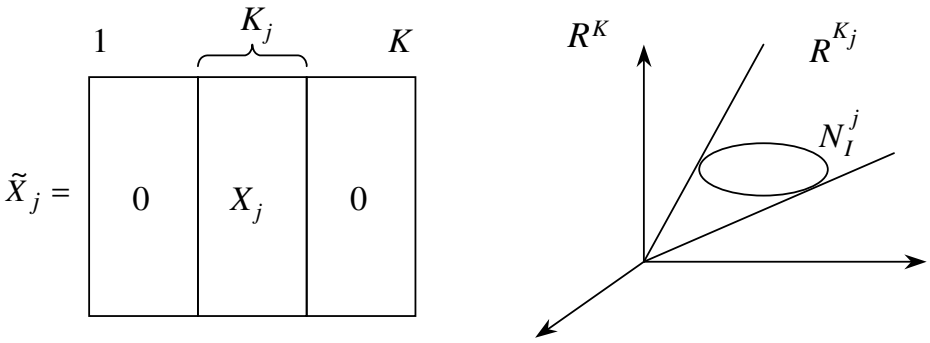



**Figure 8.2** Les nuages $N_I^j$ dans $R^K$. $N_I^j$ appartient au sous-espace $R^{Kj}$ (de $R^K$) engendré par les variables du seul groupe $j$.

### 8.2.4 Représentation du nuage moyen

Cette représentation s'inscrit dans la problématique classique de l'analyse factorielle. On souhaite projeter le nuage $N_I$, ou, ce qui revient au même, son homothétique $N_I^*$, sur un sous-espace de petite dimension tel que la projection obtenue ressemble le plus possible à $N_I$.

Pour cela, on réalise une ACP du tableau $X$. La particularité de cette ACP est que les variables sont pondérées afin d'équilibrer le rôle des groupes.

### 8.2.5 Représentation superposée des *J* nuages définis par chaque groupe de variables

*a) Problématique*

L'intérêt de cette représentation a été mis en évidence dans l'exemple des vins du chapitre précédent. Ainsi, il a été possible de montrer que tel vin, du point de vue de la puissance, était plus caractéristique sur le plan olfactif que sur le plan gustatif.

Toujours dans cet exemple, nous avons indiqué que cette représentation était obtenue par projection, en tant qu'éléments supplémentaires, des nuages $N_I^j$ sur les axes factoriels de $N_I$. Nous justifions ici cette démarche en posant le problème de la recherche directe d'une représentation superposée.

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 cherche à satisfaire deux conditions essentielles pour qu'une telle représentation permette de comparer la position d'un même individu dans les différents nuages.

➤ (C1) Chaque nuage $N_I^j$ doit être « 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. Plus précisément, on cherche à maximiser l'inertie de l'union des $N_I^j$. Soit $N_I^J = \cup_j N_I^j$ .

➤ (C2) Les représentations des nuages $N_I^j$ doivent se « ressembler entre elles »

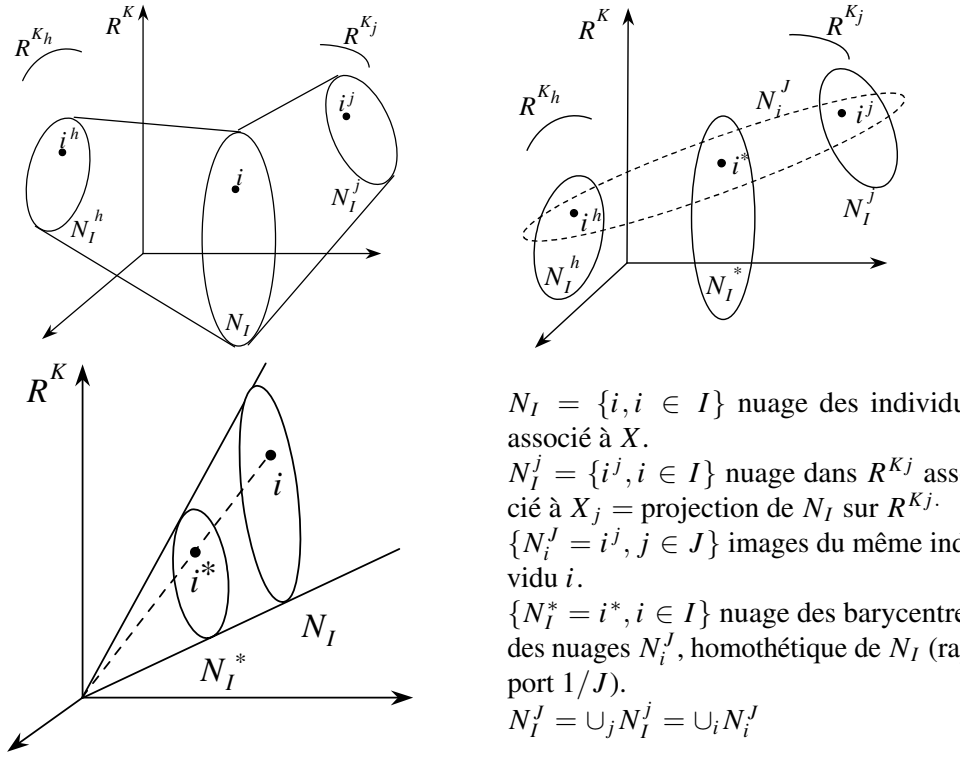
Il n'est pas possible de comparer les positions d'un même point dans les différents nuages si ces représentations sont, dans l'ensemble, très différentes. En particulier, des symétries, rotations ou homothéties, peuvent masquer complètement de fortes ressemblances entre les nuages. Pour assurer cette condition, il faut que les points homologues (représentant le même individu) soient le plus proche possible les uns des autres.

Le nuage $N_I^J$ a été partitionné jusqu'ici en $J$ nuages (contenant chacun $I$ points et notés $N_I^j$) représentant chacun l'ensemble des individus vus au travers d'un groupe de variables. Introduisons maintenant une autre partition de $N_I^J$ : $I$ nuages (contenant chacun $J$ points et notés $N_i^J$) représentant chacun le même individu $i$ vu au travers de chaque groupe de variables (*cf.* **Figure 8.3**).

Le centre de gravité de $N_i^J$ est $i^*$. Selon le théorème de Huygens 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, on cherche à minimiser l'inertie projetée de chaque $N_i^J$ donc l'inertie intra de $N_I^J$.

➤ Compromis entre (C1) et (C2)

Pour satisfaire simultanément les critères associés à (C1) et (C2), le sous-espace cherché devrait 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 : la qualité de représentation des nuages et la 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. À 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.




$N_I = \{i, i \in I\}$ nuage des individus associé à $X$.
$N_I^j = \{i^j, i \in I\}$ nuage dans $R^{Kj}$ associé à $X_j$ = projection de $N_I$ sur $R^{Kj}$.
$\{N_i^J = i^j, j \in J\}$ images du même individu $i$.
$\{N_I^* = i^*, i \in I\}$ nuage des barycentres des nuages $N_i^J$, homothétique de $N_I$ (rapport $1/J$).
$N_I^J = \cup_j N_I^j = \cup_i N_i^J$

**Figure 8.3** Nuages en présence dans $R^K$.

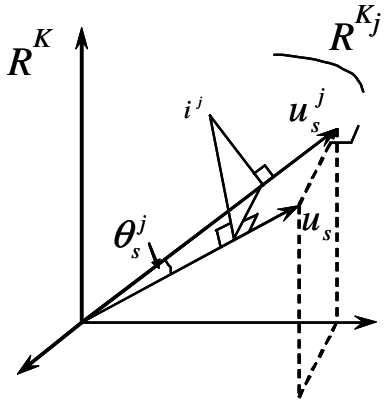
Le théorème de Huygens (inertie inter = inertie totale - inertie intra) suggère un compromis entre une inertie totale maximum et une inertie intra minimum : une inertie inter maximum.

### *b) Solution : 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, notés $u_s$, du nuage $N_I^*$ des centres de gravité. Or ce nuage est homothétique au nuage $N_I$ associé à l'ensemble de toutes les variables. Le sous-espace cherché s'obtient par une ACP du tableau $X$ tout entier.

Les coordonnées des points de $N_I^j$ sont contenues dans le tableau $\tilde{X}_j$ de dimension $(I, K)$ dans lequel $X_j$ est complété par des zéros (*cf.* **Figure 8.2**). En introduisant ces tableaux en supplémentaire dans l'ACP de $X$, on obtient la représentation simultanée des $N_I^j$.

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




$u_s$ : axe principal (de rang $s$) de $N_I$.
$u_s^j$ : composante de $u_s$ dans $R^{Kj}$.
Projeter $i^j$ sur $u_s$ est équivalent à projeter $i$ d'abord sur $u_s^j$ puis sur $u_s$.

**Figure 8.4** Projection de $i^j$.

### *c) Remarques sur la projection des nuages $N_I^j$*

Le nuage $N_I^j$, qui appartient au sous-espace $R^{Kj}$, est projeté sur un vecteur $u_s$ de $R^K$ qui n'appartient pas à $R^{Kj}$. 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 $R^{Kj}$) puis une projection sur $u_s$ qui contracte le nuage en multipliant les coordonnées par $\cos(\theta_s^j)$, en notant $\theta_s^j$ l'angle entre $u_s$ et $u_s^j$ (*cf.* **Figure 8.4**)

Cela 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. À 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 coïncide avec les centres de gravité des $N_i^J$ ne serait plus vérifiée. En outre, les points homologues ($i^j$; $j = 1, J$) ne seraient plus proches entre eux.

### *d) Aides à l'interprétation*

➤ 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 puisque le vecteur $u_s$ de $R^K$, sur lequel $N_I^j$ est projeté, n'appartient pas au sous-espace $R^{Kj}$ dans lequel ce nuage est situé. Ce vecteur $u_s$ fait, avec sa projection $u_s^j$ sur ce sous-espace $R^{Kj}$, un angle déjà noté $\theta_s^j$. D'où :

Qualité de représentation de $N_I^j$ sur $u_s$=$(\cos\,\theta_s^j)^2$ ×(qualité de représentation 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'ACP du seul nuage $N_I^j$, même si $u_s^j$ est une composante principale de $N_I^j$ .

En d'autres termes, l'indicateur [*inertie projetée* / *inertie totale*] appliqué à $N_I^j$ rend compte de façon pessimiste de la qualité de représentation en ce sens que la forme du nuage peut être bien respectée même si ce rapport est faible. Pour cette raison, cet indicateur n'est pas utilisé en pratique ; pour évaluer la qualité de représentation d'un groupe, on utilise plutôt le nuage des variables (*cf.* section 8.3.5).

➤ 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 mesure de ressemblance entre les projections des nuages $N_I^j$ sur un axe cette inertie intra. Mais cette valeur n'a de signification que comparée à l'inertie totale. On calcule donc, pour chaque axe, le rapport : [*inertie inter* / *inertie totale*].

Ce rapport, n'étant pas la quantité minimisée, ne décroît pas forcément avec l'ordre des axes. Mais il constitue un indicateur de l'utilité globale de la représentation superposée des nuages $N_I^j$. L'objet de cette représentation est, rappelons-le, une analyse détaillée des différences de forme entre les nuages $N_I^j$. Si ce rapport est proche de 1, tous les nuages $N_I^j$ ont suffisamment de caractères communs pour autoriser une étude fine de leurs différences.

## 8.3 L'AFM DANS L'ESPACE DES VARIABLES $R^I$

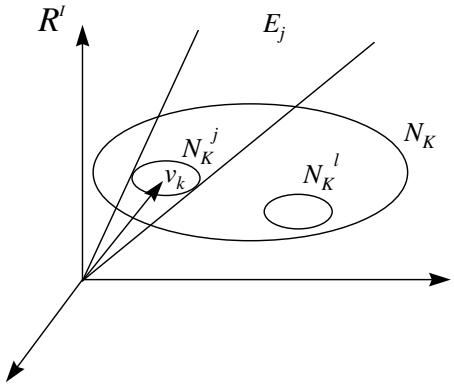
L'espace $R^I$ est l'espace des fonctions numériques définies sur l'ensemble des individus. C'est dans cet espace que sont situées les variables initiales : l'espace $R^I$ permet avant tout d'obtenir une représentation de ces variables.

Les composantes principales issues des ACP séparées de chacun des groupes peuvent aussi être situées dans $R^I$. Il est utile de visualiser leurs positions relatives au même titre que les variables initiales.

Enfin, l'exemple des vins a fait apparaître la notion de facteur commun à plusieurs ensembles de variables. En tant que fonction définie sur l'ensemble des individus, un facteur commun est un élément de $R^I$ et la problématique qui lui est associée peut être présentée dans cet espace.

Après avoir introduit les composantes principales de l'AFM comme espace de représentation des variables initiales et des composantes des analyses séparées, l'AFM est située par rapport à l'analyse multicanonique définie par Carroll.

### 8.3.1 Influence de la pondération des groupes sur les nuages des variables




$R^I$ est muni de la métrique $D$.
$v_k$ : vecteur représentant la variable $k$.
$N_K^j$ : ensemble des vecteurs $v_k$ appartenant au groupe $j$.
$N_K = \cup_j N_K^j$
$E_j$ : sous-espace engendré par les variables du groupe $j$.

**Figure 8.5** Les nuages de variables.

Chaque groupe de variables $K_j$ est représenté par un nuage $N_K^j$ (*cf.* **Figure 8.5**).

La pondération des groupes, divisant le poids de chaque variable du groupe $j$ par $\lambda_1^j$, rend égale à 1 l'inertie de la première composante principale de chaque nuage $N_K^j$.

La **figure 8.6** illustre cette pondération dans un cas simple.

En AFM, la pondération des variables d'un groupe tient compte à la fois du nombre de variables et de leurs liaisons. Remarquons qu'une pondération qui ne tiendrait pas compte des liaisons entre les variables (par exemple, en égalisant les inerties totales

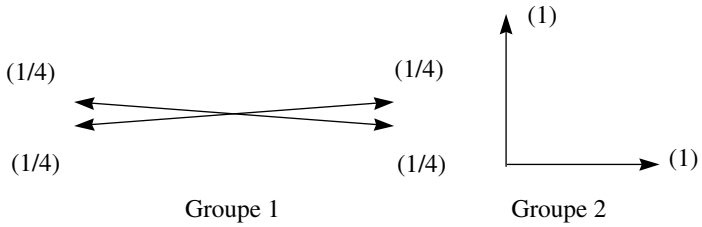



**Figure 8.6** Illustration de la pondération de l'AFM dans un cas simple. Les vecteurs représentent les variables dans $R^I$. Les nombres entre parenthèses sont les poids associés aux variables dans l'AFM. Les variables du premier groupe sont pratiquement identiques : chacune est affectée d'un poids tel que l'ensemble du groupe a un poids pratiquement égal à 1. Les variables du second groupe sont non corrélées : chacune est affectée d'un poids égal à 1.

des $N_K^j$) rendrait faible (relativement) l'inertie, dans chaque direction, d'un groupe composé de beaucoup de variables indépendantes. En revanche, une telle pondération rendrait forte (relativement) l'inertie dans une direction d'un groupe composé d'une seule variable.

### 8.3.2 Représentation des variables

Cette représentation est obtenue directement dans l'ACP du tableau complet $X$ : elle est donc duale de l'image de $N_I$ obtenue dans $R^K$. Comme en ACP, la représentation des variables peut être considérée à la fois :

1. comme une aide à l'interprétation de la représentation du nuage des individus ;
2. comme une représentation optimale des (corrélations entre) variables.

Dans le cas d'un tableau multiple, on obtient ainsi une image simplifiée des corrélations inter et intra groupe. En ce sens, la représentation des variables est un aspect de la comparaison fine des groupes de variables.

Les composantes principales rendent maximum l'inertie des projections de toutes les variables. L'inertie projetée de chaque nuage $N_K^j$ peut donc être interprétée comme la contribution d'un groupe. La pondération des groupes (par $1/\lambda_1^j$) 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 :

1. aucun groupe ne peut, à lui seul, déterminer le premier axe (sauf situation de symétrie exceptionnelle) ;
2. un groupe influe sur d'autant plus d'axes qu'il est de grande dimensionalité.

### 8.3.3 Représentation des composantes principales de chaque groupe

Dans l'exemple des vins, la première composante de l'AFM (de l'ensemble des variables) était très corrélée avec la première composante de chaque groupe. Une étude systématique des corrélations entre les premières composantes de chaque groupe apporte des éléments intéressants pour la comparaison de ces groupes.

Une telle étude peut être réalisée par une ACP des composantes principales de tous les groupes. Les composantes principales du tableau $X_j$ étant les projections du nuage d'individus sur une base orthonormée, les nuages d'individus définis dans l'ACP de $X_j$ et dans celle du tableau des composantes de $X_j$ sont identiques. Mais ceci à condition de conserver les valeurs brutes de ces composantes. Une ACP non normée des composantes principales des groupes aboutit donc aux mêmes composantes qu'une ACP de l'ensemble des variables. Une autre façon de respecter l'inertie $\lambda_s^j$ de la composante de rang $s$ du groupe $j$ consiste à normer cette composante et à lui affecter le poids $\lambda_s^j$.

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 (via l'indicateur usuel) d'une composante d'un groupe à la construction des axes.

On peut aussi adopter la démarche inverse : ACP des composantes principales avec les variables en supplémentaire.

### 8.3.4 Recherche de facteurs communs aux groupes de variables : AFM et analyse multicanonique

*a) 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, étudié par Hotelling en 1936 à l'aide de ce qu'il appela l'analyse canonique.

Dans cette analyse, on recherche simultanément une combinaison linéaire des variables du premier groupe (notée $\phi$) et une combinaison linéaire des variables du second groupe (notée $\psi$) telles que le coefficient de corrélation entre $\phi$ et $\psi$ soit maximum. Ce couple étant trouvé, on en recherche un deuxième orthogonal au premier qui satisfait le même critère, etc.

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 entre elles le plus possible. L'existence de plusieurs variantes d'Analyse Canonique Généralisée (on dit aussi : Analyse Multicanonique) tient en particulier à la multiplicité des façons de définir un critère de liaison entre plusieurs variables.

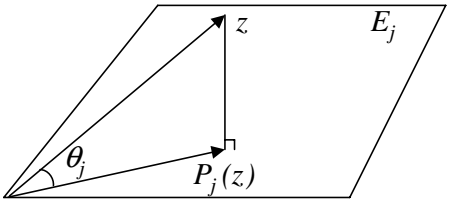
*b) L'Analyse Canonique Généralisée de Carroll*

Le principe de cette analyse est de chercher d'abord des variables liées à l'ensemble des groupes. Ces variables, qui résument les tendances générales des groupes, sont appelées variables générales. Puis, une variable générale étant obtenue, on cherche dans chaque groupe une combinaison linéaire des variables liée à cette variable générale. Ces combinaisons linéaires, qui sont en quelque sorte les représentations de la variable générale dans les groupes, sont appelées variables canoniques.

L'un des avantages de cette approche est qu'il n'est pas nécessaire de définir une mesure de liaison entre deux groupes de variables mais entre une variable et un groupe. Celle utilisée par Carroll est le carré du coefficient de corrélation multiple.

Par définition, le coefficient de corrélation multiple entre une variable $z$ et un groupe de variables $K_j$ est le coefficient de corrélation entre $z$ et la combinaison linéaire des variables du groupe $j$ la plus corrélée à $z$. Géométriquement, dans $R^I$, cette combinaison linéaire est la projection orthogonale $P_j(z)$ de $z$ sur le sous-espace

$E_j$ engendré par les variables du groupe $j$ (*cf.* **Figure 8.7**). Ainsi, le coefficient de corrélation multiple est le cosinus de l'angle $\theta_j$ entre $z$ et sa projection sur $E_j$.




$E_j$ : sous-espace engendré par le groupe $K_j$
$P_j$ : projecteur sur $E_j$
$\cos\theta_j$ : coefficient de corrélation multiple entre $z$ et $K_j$.

**Figure 8.7** Le coefficient de corrélation multiple dans $R^I$.

Si $z$ est une variable normée, on a (en notant $\langle u, v\rangle$ le produit scalaire entre les vecteurs $u$ et $v$) :

$$\cos^2\theta_j = \langle z, P_j(z)\rangle$$

Dans l'analyse multicanonique de CARROLL, on recherche une suite de variables générales $z_s$ qui rendent maximum la somme des carrés des coefficients de corrélation multiple entre $z_s$ et les $J$ groupes $K_j$ (avec la contrainte d'orthogonalité : $z_s \perp z_t$ si $s \neq t$). Cette quantité s'écrit :

$$\sum_j \cos^2\theta_j = \sum_j \langle z_s, P_j(z_s)\rangle = \langle z_s, \sum_j P_j(z_s)\rangle$$

L'opérateur $\Sigma_j P_j$ étant une somme d'opérateurs de projection orthogonale, il est symétrique, diagonalisable et ses vecteurs propres sont orthogonaux deux à deux. Une suite de vecteurs propres normés de cet opérateur, ordonnée par les valeurs propres, définit donc les variables générales (*cf.* section 5.2.4).

La variable canonique du groupe $j$, associée à une variable générale $z_s$, est sa projection $P_j(z_s)$ sur le sous-espace $E_j$.

Cette analyse multicanonique est intéressante sur le plan théorique car elle donne un cadre commun à plusieurs méthodes d'analyse.

1. L'ACP est une analyse multicanonique de groupes de variables réduits chacun à un élément. Les composantes principales sont les variables générales et les variables canoniques sont confondues avec les variables initiales.
2. L'ACM est une analyse multicanonique dans laquelle les groupes de variables sont composés chacun par les indicatrices des classes d'une même variable qualitative.

Mais la mesure de liaison utilisée, le coefficient de corrélation multiple, n'est pas sans inconvénient dans le cas où les variables du groupe sont corrélées entre elles. Lorsque les variables du groupe $K_j$ sont liées, le sous-espace $E_j$ est instable (c'est-à-dire sensible à de petites variations des variables) et l'on peut se trouver confronté à des situations paradoxales (*cf.* **Figure 8.8**).

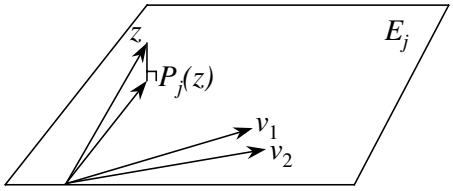



**Figure 8.8** Inadaptation du coefficient de corrélation multiple dans le cas de variables liées. La variable $z$ est presque orthogonale à chacune des variables $v_1$ et $v_2$ du groupe $K_j$. Or, son coefficient de corrélation multiple avec $E_j$ vaut presque 1.

### *c) Une mesure de liaison entre une variable $z$ et un groupe $K_j$ : l'inertie du nuage pondéré projeté sur $z$*

Avec la pondération des groupes, l'inertie de la projection de $N_K^j$ sur une variable $z$ est toujours comprise entre 0 et 1.

Elle atteint sa valeur maximum 1 lorsque $z$ est dans la direction d'inertie maximum de $N_K^j$, c'est-à-dire lorsque $z$ est confondue avec la première composante principale de $N_K^j$. Cette inertie vaut 0 pour toute variable $z$ orthogonale au sous-espace $E_j$, c'est-à-dire non corrélée avec chacune des variables du groupe $j$. Ailleurs, sa valeur est strictement positive mais, contrairement au coefficient de corrélation multiple, elle ne dépend pas uniquement de l'angle $\theta_j$ entre $z$ et $E_j$. Par exemple, dans le cas (*cf.* **Figure 8.8**) où $z$ est proche de $E_j$ mais presque orthogonale à chacune des variables du groupe, l'inertie projetée est très petite tandis que le coefficient de corrélation multiple est très grand. On montre facilement l'égalité :

$$\text{Inertie de } N_K^j \text{ sur } z = cos^2\theta_j \times \text{Inertie de } N_K^j \text{ sur } P_j(z)$$

Toutes ces propriétés font que cette inertie projetée est une bonne mesure de liaison entre une variable et un groupe de variables. Elle présente des avantages sur le coefficient de corrélation multiple lorsque les variables du groupe $K_j$ sont corrélées entre elles. Elle a la même valeur que le carré de ce dernier lorsque les variables (normées et de poids 1) ne sont pas corrélées entre elles ; en effet, dans le cas où les variables de $K_j$ sont orthogonales deux à deux, l'inertie de la projection de $N_K^j$ vaut 1 dans toutes les directions de $E_j$.

On définit donc une mesure de liaison, notée Lg, entre la variable $z$ et le groupe de variables $K_j$ en posant (les variables du groupe $K_j$ étant pondérées au sens de l'AFM) :

$$\mathrm{Lg}(z, K_j) = \sum_{k \in K_j} \text{inertie de la projection de } v_k \text{ sur } z$$

En notant $W_j = X_j M_j X_j'$ la matrice des produits scalaires entre les individus vus par $K_j$, cette mesure de liaison s'écrit aussi :

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

Cette écriture met en évidence le fait que la mesure Lg prend en compte le groupe $K_j$ au travers de l'opérateur $W_j D$ et non pas $P_j$ comme le fait le coefficient de corrélation multiple (*cf.* section b). Cet opérateur caractérise bien le groupe $K_j$ (sa diagonalisation permet de reconstituer la forme du nuage $N_I^j$, *cf.* section 5.4.5 page 120) ; il est moins sensible que $P_j$ à de petites variations des données.

### *d) L'AFM vue comme une analyse multicanonique particulière*

Nous appliquons ici la démarche proposée par CARROLL mais en caractérisant le groupe $K_j$ non plus par le projecteur $P_j$ mais par l'opérateur $W_j D$. L'AFM ne diffère donc de l'analyse de CARROLL que lorsque les variables d'un même groupe sont corrélées entre elles.

#### ➤ Variables générales

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 $N_K^j$. Nous cherchons donc une première variable générale $z_1$ telle que la somme des liaisons (au sens du paragraphe précédent) entre $z_1$ et les $J$ groupes $K_j$ soit maximum. Cette expression s'écrit :

$$\sum_{j \in J} \mathrm{Lg}(z_1, K_j) = \sum_{k \in K} \text{inertie de la projection de } v_k \text{ sur } z_1$$

On retrouve exactement l'expression maximisée par les composantes principales de $X$.

#### ➤ Pondération des groupes

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 $\mathrm{Lg}(z_s, groupej)$ doivent *a priori* avoir le même intervalle de variation pour tous les groupes. Avec la pondération de l'AFM, 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 $\sum_j \mathrm{Lg}(z_s, K_j)$ est bornée par 1.

Pour chaque groupe $K_j$ et chaque variable générale $z_s$, la quantité $\mathrm{Lg}(z_s, K_j)$ sert d'aide à l'interprétation. Elle mesure l'inertie projetée des variables du groupe $K_j$, c'est-à-dire :

- leur contribution cumulée à la construction de l'axe $z_s$ (optique ACP) ;
- l'importance relative du facteur commun de rang $s$ dans le groupe $K_j$ (optique analyse canonique). Cette importance doit être comparée d'une part à l'inertie du nuage $N_K^j$ dans les autres directions et d'autre part à l'inertie des autres nuages pour ce même facteur.

La convergence des résultats de l'AFM en tant qu'ACP et de l'analyse multicanonique transparaît au niveau des objectifs. En ACP, on cherche notamment des variables qui « résument » l'ensemble des variables étudiées. En introduisant dans l'ACP les notions de groupe de variables et d'équilibre entre les groupes, on cherche alors aussi à résumer ces groupes. Tel est bien l'objet des variables générales de l'analyse multicanonique.

➤ Variables canoniques

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 $W_j D$ (*cf.* fin de la section c) et la variable canonique associée à une variable $z$ est son image par $W_j D$. Montrons que $W_j D(z)$ extrait du groupe $j$ une part d'inertie plus importante que la projection $P_j(z)$.

- De l'écriture matricielle de $W_j D$ et de $P_j$ (applications de $R^I$ dans $R^I$) :

$$\begin{aligned} W_j D &= X_j M X_j' D \\ P_j &= X_j (X_j' D X_j)^{-1} X_j' D \end{aligned}$$

  il résulte immédiatement que $W_j D = W_j D P_j$.
- Soit $\{l_r; r = 1, K_j\}$ une base de vecteurs propres de $W_j D$ triés par valeurs propres décroissantes ($\lambda_r \geqslant \lambda_{r+1}$). Si l'on exprime $P_j(z)$ dans cette base :

$$P_j(z) = \sum_r x_r l_r$$

  alors :

$$W_j D(z) = W_j D P_j(z) = \sum_r \lambda_r x_r l_r$$

L'application $W_j D$ renforce d'autant plus les coordonnées, ici de $P_j(z)$, dans la base des $l_r$ qu'elles correspondent à un axe de faible rang. Autrement dit, dans sa transformation à l'aide de $W_j D$, un vecteur est d'abord projeté sur $E_j$ puis est rapproché des premières directions propres de $W_j D$. Ainsi $W_j D(z_s)$ correspond à une direction de plus grande inertie que $P_j(z_s)$ (sauf dans le cas extrême où $P_j(z)$ est colinéaire à un vecteur propre, auquel cas $P_j(z)$ et $W_j D(z)$ ont des directions identiques).

***Remarque :*** On montre aisément que :

$$W_j D(z) = \sum_{k \in K_j} r(z, v_k) v_k$$

On retrouve ici l'expression de la régression PLS, à une composante, exprimant $z$ en fonction des $v_k$. La convergence entre les deux approches, AFM et régression PLS, est remarquable : par rapport aux méthodes de référence, analyse canonique et régression usuelle, dans les deux cas on prend en compte les variables du groupe $K_j$ non pas au travers du seul espace qu'elles engendrent mais de leur répartition dans cet espace.

*e) 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.

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 principale $F_s$ par la relation : $u_s = (1/\lambda_s) X' D F_s$ dans laquelle $\lambda_s$ est la valeur propre de $WD$ associée à $F_s (W = \Sigma_j W_j)$.

La projection de $N_I^j$ sur $u_s$ s'écrit :

$$F_s^j = \tilde{X}_j M u_s = \left(1/\lambda_s\right) \tilde{X}_j M X' D F_s = \left(1/\lambda_s\right) W_j D F_s$$

Cette convergence des résultats montre bien que les deux approches, *a priori* très différentes (représentation superposée et analyse canonique), constituent en fait deux formalisations différentes d'une même problématique. En effet, la représentation superposée des nuages $N_I^j$ est liée à l'existence de facteurs communs : c'est là un autre point de vue sur l'intérêt d'une représentation simultanée déjà abordé section 8.2.5.

### 8.3.5 Aides à l'interprétation

Pour juger du caractère véritablement commun (aux groupes de variables) de $F_s$, l'optique « représentation superposée des $N_I^j$ » conduit à calculer un rapport d'inertie.

L'optique « analyse canonique » suggère, quant à elle, d'évaluer le degré de ressemblance entre $F_s$ et chaque $F_s^j$ au moyen du coefficient de corrélation entre $F_s$ et $F_s^j$. Ce coefficient peut avoir une valeur élevée pour l'ensemble des groupes, pour certains d'entre eux ou même pour un seul. En ce sens, l'AFM permet de mettre en évidence les facteurs communs à l'ensemble des groupes, les facteurs communs à certains groupes et les facteurs spécifiques d'un groupe (*cf.* exemple section 7.1.6 page 159). Lors de l'interprétation, on distingue :

1. le coefficient de corrélation entre $F_s$ et $F_s^j$, qui indique dans quelle mesure le facteur commun $F_s$ est effectivement présent dans le groupe $K_j$ ;
2. la mesure de liaison $\mathrm{Lg}(z_s, K_j)$, qui indique l'importance relative dans le groupe $K_j$ du facteur commun de rang $s$.

À ces aides spécifiques, s'ajoutent les aides à l'interprétation usuelles : qualité de représentation d'une variable par un axe et contribution d'une variable à la construction d'un axe.

## 8.4 L'AFM DANS L'ESPACE DES GROUPES DE VARIABLES $R^{I^2}$

Dans l'étude de plusieurs groupes de variables, l'un des objectifs est de comparer globalement les groupes. Dans l'exemple des vins, la parenté entre les deux olfactions a pu être mise en évidence à l'aide d'un graphique sur lequel les groupes sont représentés chacun par un point.

Nous introduisons ici l'espace $R^{I^2}$, base de cette représentation qui peut apparaître comme une aide à l'interprétation de l'ACP du tableau complet $X$ (c'est ainsi qu'elle a été introduite dans l'exemple des vins) mais qui possède sa propre optimalité.

### 8.4.1 Le nuage $N_J$ des groupes de variables

Pour étudier l'ensemble des groupes, nous construisons, comme pour les individus et les variables, un nuage de points, noté $N_J$, dans un espace euclidien. Nous avons déjà présenté l'aptitude de l'opérateur $W_j D$ à représenter le groupe de variables $K_j$ ; en particulier, nous avons insisté sur les avantages de ce choix par rapport à celui du sous-espace $E_j$. L'argument essentiel est que $W_j D$ permet, par sa diagonalisation, une reconstitution parfaite de la structure du nuage $N_I^j$ des individus défini par le groupe $K_j$. En tant qu'ensemble de $I^2$ scalaires, la matrice $W_j D$ peut être considérée comme un élément d'un espace vectoriel de dimension $I^2$ noté $R^{I^2}$. Un groupe $j$ est représenté dans $R^{I^2}$ par la matrice $W_j D$. Cet espace est muni du produit scalaire classique, qui s'écrit pour les éléments $W_j D$ et $W_l D$ :

$$\langle W_j D, W_l D \rangle = \sum_i \sum_{i'} p_i p_{i'} W_j(i, i') W_l(i, i') = \mathrm{trace}(W_j D W_l D)$$

### 8.4.2 Influence de la pondération des groupes sur le nuage $N_J$

La norme, dans $R^{I^2}$, de $W_j D$ s'écrit :

$$\|W_j D\|^2 = \sum_s (\lambda_s^j)^2$$

La pondération des variables du groupe $j$ par $1/\lambda_1^j$ se traduit dans $R^{I^2}$ par une homothétie des vecteurs représentant les groupes. Après cette pondération, la norme du vecteur $W_j D$ représentant le groupe $j$ n'est pas égale à 1 mais dépend de la structure du groupe : cette norme est d'autant plus grande que cette structure est multidimensionnelle (c'est-à-dire qu'il existe de nombreux facteurs d'importance comparable à celle du premier d'entre eux). Ainsi, elle constitue un indicateur de dimensionalité d'un nuage.

À strictement parler, la dimensionalité d'un nuage est égale au nombre de directions orthogonales d'inertie non nulle, soit le nombre de valeurs propres non nulles. En pratique, il n'y a pas lieu de distinguer une valeur propre très faible d'une valeur propre nulle. C'est ce que réalise, à sa manière, la norme de $W_j D$ après pondération par l'AFM. Cet indicateur, noté alors $\mathrm{N}_{\mathrm{g}}^2$, peut donc finalement s'écrire, en faisant apparaître explicitement la pondération de l'AFM :

$$\mathrm{N}_{\mathrm{g}}^2(K_j) = \|W_j D\|^2 = \sum_s \left[\frac{\lambda_s^j}{\lambda_1^j}\right]^2$$

### 8.4.3 Interprétation du produit scalaire entre deux groupes

Le nuage $N_J$ des groupes s'apparente plus à un nuage de variables qu'à un nuage d'individus car le produit scalaire entre les vecteurs représentant deux groupes s'interprète comme une mesure de liaison entre ces groupes.

Étudions d'abord le cas le plus simple où les deux groupes sont composés d'une seule variable, puis le cas où un seul des deux groupes est unidimensionnel et enfin le cas où les deux groupes sont multidimensionnels.

#### *a) Les deux groupes comprennent chacun une seule variable*

La pondération par $1/\lambda_1^j$ donne le poids 1 à une variable centrée réduite qui constitue à elle seule un groupe. À ce groupe d'une seule variable, correspond un élément de $R^{I^2}$ dit élément de rang 1 (il est associé à une matrice symétrique de rang 1). L'écriture suivante fait apparaître, dans le cas général, $W_j$ en tant que somme d'éléments de rang 1 (en notant $v_k$ une variable, de poids $m_k$, du groupe $j$) :

$$W_j = \sum_k v_k m_k v_k'$$

Soit $z$ et $v$ deux variables centrées réduites de poids 1 constituant chacune un groupe. Les éléments de $R^{I^2}$ associés à ces groupes ont chacun pour norme 1 et leur produit scalaire est le carré du coefficient de corrélation entre $z$ et $v$ ($p_i$ : poids de l'individu $i$).

$$\begin{aligned}\langle zz'D, vv'D\rangle &= \sum_i \sum_{i'} p_i p_{i'} z(i)z(i')v(i)v(i') = \left(\sum_i p_i z(i)v(i)\right)^2 \\ &= \text{inertie de la projection de } v \text{ sur } z \\ &= [\mathrm{r}(v,z)]^2\end{aligned}$$

*b) Un groupe d'une variable et un groupe multidimensionnel*

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

$$\begin{aligned}\langle W_1 D, W_2 D\rangle &= \langle z\ z'D, \sum_k v_k m_k v_k'\ D\rangle = \sum_k m_k \langle z\, z'D, v_k v_k'\ D\rangle \\ &= \sum_k (\text{inertie de la projection de } v_k \text{ sur } z) \\ &= \mathrm{Lg}(z, K_2)\end{aligned}$$

On retrouve ici la mesure de liaison Lg entre une variable et un groupe de variables (*cf.* section c). Cette coïncidence fait que les choix, de Lg d'une part et de la métrique dans $R^{I^2}$ d'autre part, se renforcent mutuellement.

*c) Les deux groupes sont multidimensionnels*

Notons $z_l$ les variables (réduites et de poids $m_l$) du groupe $K_1$ et $v_k$ les variables (réduites et de poids $m_k$) du groupe $K_2$. Alors :

$$\begin{aligned}\langle W_1 D, W_2 D\rangle &= \sum_l m_l \sum_k m_k \langle z_l z_l' D, v_k v_k' D\rangle = \sum_l m_l \mathrm{Lg}(z_l, K_2) \\ &= \sum_k m_k \mathrm{Lg}(v_k, K_1)\end{aligned}$$

Cette quantité vaut 0 lorsque toutes les variables d'un groupe sont orthogonales à toutes les variables de l'autre groupe. Elle est d'autant plus grande que chacune des variables d'un groupe est plus liée à l'ensemble des variables de l'autre groupe. Elle constitue un indice de liaison général entre groupes de variables, qui exprime en quelque sorte « le nombre de dimensions communes (aux deux groupes) d'inertie comparable à l'inertie axiale maximum d'un groupe ». D'où l'idée de généraliser la

première définition de $\mathrm{L_g}$ à deux groupes quelconques de variables ; soit, en faisant apparaître explicitement la pondération de l'AFM :

$$\mathrm{Lg}(K_1, K_2) = \langle \frac{W_1 D}{\lambda_1^1}, \frac{W_2 D}{\lambda_1^2} \rangle$$

*d) $L_g$ et RV*

L'indicateur classique de liaison entre deux groupes de variables est le coefficient RV ; il peut être défini par :

$$\mathrm{RV}(K_1, K_2) = \langle \frac{W_1 D}{\|W_1 D\|}, \frac{W_2 D}{\|W_2 D\|} \rangle$$

Cette définition fait bien apparaître les $W_j D$ normés au sens usuel (de longueur 1) : le coefficient RV s'interprète dans $R^{I^2}$ comme un cosinus. Etant toujours positif, il varie entre 0 et 1, valeur atteinte lorsque les nuages d'individus associés aux deux groupes sont homothétiques (en comparaison, la mesure $\mathrm{L_g}$ entre deux nuages homothétiques est d'autant plus grande que ces groupes ont une dimensionalité élevée). Un exemple d'interprétation conjointe de RV et $\mathrm{L_g}$ se trouve en section 9.2.1 page 211.

### 8.4.4 Étude du nuage des groupes de variables

*a) Représentation des groupes en tant qu'aide à l'interprétation de l'ACP*

Dans l'exemple des vins, nous avons proposé un graphique dans lequel :

1. chaque axe représente un facteur de l'ACP pondérée du tableau complet $X$ ;
2. chaque groupe est représenté par un point. La coordonnée du groupe $K_j$ le long d'un axe est égale à l'inertie projetée (calculée dans $R^I$) des variables du groupe $K_j$ le long de l'axe correspondant.

L'espace $R^{I^2}$ fournit une interprétation géométrique à ce graphique (*cf.* **Figure 8.9**). En effet, considérons le groupe formé par la variable définie par l'axe $z$ de $R^I$ : sa représentation dans $R^{I^2}$ est l'opérateur de rang 1 : $W_z D = zz'D$. L'inertie projetée des variables du groupe $K_j$ sur $z$ est égale au produit scalaire entre $W_j D$ et $zz'D$, c'est-à-dire à la projection de $W_j D$ sur $zz'D$.

En outre, si dans $R^I$ deux vecteurs sont orthogonaux, alors les vecteurs associés dans $R^{I^2}$ le sont aussi. Il en résulte que la représentation graphique des groupes de variables peut être interprétée comme une projection du nuage $N_J$ sur une suite d'axes orthogonaux.

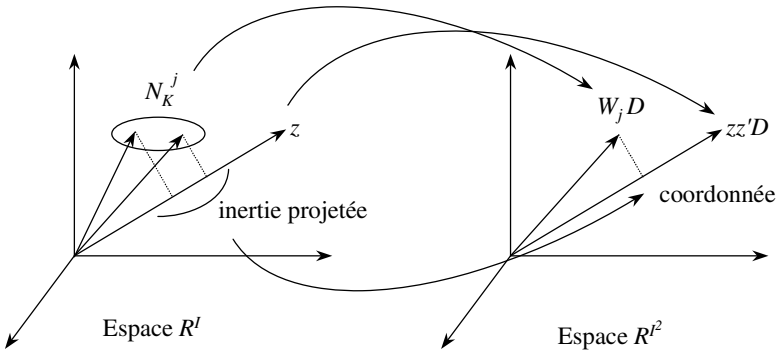



**Figure 8.9** La représentation des groupes vue comme une aide à l'interprétation de l'ACP pondérée. Au groupe de variables $j$, on associe le nuage $N_K^j$ dans $R^I$ et le vecteur $W_j D$ de $R^{I^2}$. Au vecteur $z$ de $R^I$, on associe dans $R^{I^2}$ le vecteur $zz'D$. L'inertie projetée de $N_K^j$ sur $z$ dans $R^I$ est égale à la longueur de la projection de $W_j D$ sur $zz'D$.

### *b) Représentation des groupes en tant qu'image optimale du nuage $N_J$*

Nous montrons ici que le graphique précédent peut être obtenu directement en cherchant une représentation optimale de $N_J$.

Le produit scalaire entre $W_j D$ et $W_l D$ est une mesure de liaison entre les groupes de variables $j$ et $l$. Pour comparer globalement les groupes, nous cherchons à décrire les proximités entre les $W_j D$ en les projetant sur un espace de faible dimension de $R^{I^2}$. Les angles entre les $W_j D$ doivent être bien représentés et il ne convient pas de centrer le nuage $N_J$.

En exigeant uniquement une bonne qualité de représentation (au sens de l'inertie projetée) des $W_j D$, on est conduit à une projection du nuage $N_J$ sur ses axes d'inertie, analogue à celle du nuage des variables de l'ACP. L'inconvénient de ce type d'analyse est de fournir un repère constitué d'axes difficilement interprétables car un axe quelconque de $R^{I^2}$ ne s'exprime pas clairement en fonction des données. C'est pourquoi, en AFM, on impose aux axes du repère d'être des éléments symétriques de rang 1. Ces éléments, de la forme $z_s z_s' D$, sont associés à des groupes d'une seule variable $z_s$ et s'interprètent à partir de $z_s$ et de ses liaisons avec les variables initiales.

Nous cherchons donc un repère orthonormé dans $R^{I^2}$ dont chaque composant est de la forme $zz'D$ et qui « ajuste » au mieux le nuage des $W_j D$. Nous construisons ce repère progressivement en cherchant d'abord 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. En AFM, du fait de la contrainte imposée aux vecteurs de base du repère, c'est la somme des projections et non de leurs carrés qui est maximisée.

Ce critère est plus facile à mettre en œuvre que celui des moindres carrés (souvent choisi pour les facilités de calcul qu'il implique) et possède une signification puisque les coordonnées des $W_j D$ sur des éléments de type $z_s z_s' D$ sont toujours positives. En effet, la somme des projections des $W_j D$ sur $z_s z_s' D$, qui s'écrit :

$$\sum_j \langle W_j D, z_s z_s' D \rangle$$

est égale à l'inertie dans $R^I$ des variables (de tous les groupes) projetées sur $z_s$.

La suite orthonormée d'éléments symétriques de rang 1 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 $z_s z_s' D$ dans $R^{I^2}$. Les calculs nécessités par l'analyse dans $R^{I^2}$ se déduisent directement des résultats de l'ACP de $X$ : les $z_s$ sont les composantes principales normées de $X$ et la coordonnée de $W_j D$ sur $z_s z_s' D$ est la contribution du groupe $j$ à l'inertie de la composante $z_s$.

### *c) Interprétation de la représentation des groupes*

La représentation des groupes en AFM peut être vue à la fois comme une aide à l'interprétation des autres graphiques et comme une image du nuage des groupes optimale en elle-même. La coordonnée de $W_j D$ sur l'axe factoriel $z_s z_s' D$ s'interprète comme :

1. l'inertie de la projection du nuage $N_K^j$, défini par le groupe $j$ dans $R^I$ sur la composante principale $z_s$ du tableau $X$ ; c'est la contribution (absolue, c'est-à-dire non exprimée en %) du groupe $j$ à l'axe $s$ ;
2. une mesure de liaison ($\mathrm{L_g}$) entre le groupe $j$ et la composante $z_s$ de l'AFM ;
3. la projection du groupe $j$ dans l'espace $R^{I^2}$.

Du fait de la pondération des groupes, les coordonnées des $W_j D$ sont comprises entre 0 et 1 (sur un plan, $W_j D$ est toujours situé dans un carré de côté 1 ; *cf.* exemple figure 7.6 page 162). Un groupe $j$, selon la répartition de l'inertie des nuages associés $N_I^j$ et $N_K^j$, peut avoir plusieurs coordonnées proches de 1.

En tant qu'inertie projetée du groupe, la coordonnée mesure l'importance du groupe dans la détermination de la composante. Du fait de la pondération qui rend égale à 1 l'inertie maximum de la projection sur un axe du nuage associé $N_K^j$ au groupe $j$, une coordonnée de $W_j D$ sur $z_s z_s' D$ voisine de 1 implique que la direction $z_s$ est une direction d'inertie importante pour le nuage $N_K^j$. *A contrario*, une très faible coordonnée de $W_j D$ sur $z_s z_s' D$ indique que $z_s$ est une direction de très faible inertie pour $N_K^j$. Cette dernière situation recèle une ambiguïté qui peut être levée en consultant le coefficient de corrélation entre $F_s$ et $F_s^j$ (*cf.* section 8.3.5).

En tant que projection sur un sous-espace, la représentation des groupes s'accompagne des aides à l'interprétation usuelles. Du fait de la contrainte sur les axes du

repère (éléments symétriques de rang 1), la qualité de représentation des $W_j D$ 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 $I$ alors que la dimension de l'espace est $I^2$.

**AFM et méthode Statis.** Le cœur de la méthode Statis est une analyse factorielle du nuage $N_J$, les $W_j D$ étant préalablement normés au sens usuel. Elle fournit une représentation (généralement) plane des $W_j D$, optimale du point de vue de l'inertie projetée mais dont les dimensions, n'étant pas des éléments de rang 1, ne sont pas interprétables.

Enfin, la représentation de $N_J$ fournie par l'AFM peut aussi être interprétée dans le cadre du modèle INDSCAL (*cf.* § 8.6).

## 8.5 AFM ET MODÈLE INDSCAL

L'approche du modèle INDSCAL est différente des précédentes : un modèle est proposé dont il faut estimer les paramètres. Le modèle INDSCAL (Analysis of Individual Differences in Multidimensional Scaling), dû à Carroll et Chang, 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 similarités 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 distances 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, le poids affecté à chaque facteur différant suivant les groupes. Plus précisément, en notant :

1. $z_s(i)$ la valeur du $s^{\text{e}}$ facteur pour l'individu $i$ $(s = 1, ..., S)$ ;
2. $q_s^j$ le poids affecté à $z_s$ par le $j^{\text{e}}$ groupe ;
3. $\mathrm{d}_j(i,l)$ la distance entre les individus $i$ et $l$ induite par le $j^{\text{e}}$ groupe ;

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

Remarquons enfin que, dans ce modèle, tous les individus ont le même poids. Même si l'AFM suggère une généralisation en affectant des poids quelconques aux individus, nous restons conformes ici au modèle original.

### 8.5.1 Interprétation du modèle INDSCAL dans $R^K$

Lorsque les données vérifient exactement le modèle INDSCAL, ce dernier exprime une décomposition de chacun des nuages $N_I^j$ suivant $S$ projections axiales $q_s^j z_s$ homothétiques aux $z_s$.

Les données ne vérifient jamais exactement le modèle. Les paramètres (facteurs et poids) doivent être calculés par un algorithme satisfaisant un critère d'ajustement (plusieurs critères sont possibles).

Dans ce cadre, un facteur est une image de dimension 1 des individus telle qu'il existe une direction de chaque $N_I^j$ « presque » homothétique à ce facteur. Il représente une direction « presque » commune aux nuages $N_I^j$.

L'ajustement du modèle INDSCAL dans $R^K$ se présente comme la recherche d'une suite de directions de $R^K$ telle que, pour chacune, les projections des nuages $N_I^j$ soient le plus homothétique possible. Cette présentation est très proche de celle de la représentation superposée des $N_I^j$ de l'AFM (*cf.* la condition C2 de la section 8.2.5).

### 8.5.2 Interprétation du modèle INDSCAL dans $R^I$

En raison de la dualité entre ces deux nuages, une question concernant les nuages des individus dans $R^K$ peut être traduite à propos du nuage des variables dans $R^I$.

Les facteurs du modèle INDSCAL sont des éléments de $R^I$. Le $s^{\text{e}}$ facteur du groupe $j$ peut être noté $F_s^j$ par analogie aux variables canoniques. En effet, la problématique du modèle INDSCAL est analogue à celle de l'analyse multicanonique : on recherche, au rang $s$, un ensemble $\{F_s^j ; j = 1, J\}$ de facteurs se ressemblant entre eux. Dans le cas du modèle INDSCAL, la contrainte imposée aux $F_s^j$ est très stricte :

$$F_s^j = q_s^j F_s$$

L'estimation des paramètres du modèle est orientée principalement sur le facteur commun $F_s$. Cette démarche est celle de l'analyse multicanonique au sens de Carroll.

### 8.5.3 Interprétation du modèle INDSCAL dans $R^{I^2}$

À la décomposition des distances dans le modèle INDSCAL

$$d_j^2(i,l) = \sum_s q_s^j \left[z_s(i) - z_s(l)\right]^2$$

correspond celle des produits scalaires :

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

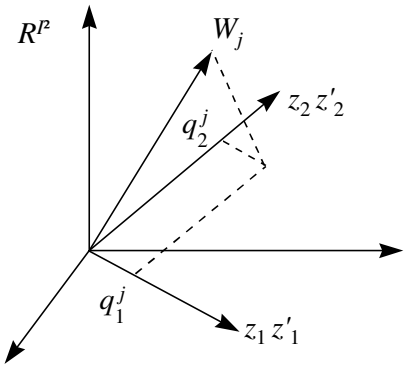



**Figure 8.10** Interprétation géométrique du modèle INDSCAL dans $R^{I^2}$. Selon ce modèle, la matrice $W_j$ des produits scalaires entre individus induite par le groupe j est une somme d'éléments symétriques de rang 1.

Soit, matriciellement :

$$W_j = \sum_s q_s^j z_s z_s'$$

Cette formule, traduite dans $R^{I^2}$, exprime que les $W_j$ sont décomposés sur un même repère formé d'éléments symétriques de rang 1. Le poids $q_s^j$ est la coordonnée de $W_j$ sur l'élément $z_s z_s'$ de ce repère. Chercher des paramètres $z_s$ et $q_s^j$ qui ajustent le modèle revient à chercher dans $R^{I^2}$ une suite orthogonale de vecteurs, représentant chacun une matrice symétrique de rang 1, qui ajuste le nuage des $W_j$ (*cf.* **Figure 8.10**).

### 8.5.4 Estimation des paramètres du modèle INDSCAL par l'AFM

L'interprétation du modèle INDSCAL dans $R^{I^2}$ et le calcul des paramètres qui en découle conduisent exactement à l'ajustement du nuage $N_J$ tel qu'il est réalisé dans l'AFM. Il en résulte que cet ajustement fournit une estimation des paramètres du modèle INDSCAL. Comparée aux méthodes d'estimation usuelles (correspondant aux programmes INDSCAL et SINDSCAL), l'AFM présente les avantages suivants.

1. Pour tout $(j, s) : 0 \leqslant q_s^j \leqslant 1$. Les poids sont toujours interprétables et même comparables d'un groupe à l'autre, d'un axe à l'autre, d'une analyse à l'autre. Il en résulte, entre autres, que la quantité $\Sigma_j q_s^j$ mesure l'importance du facteur de rang $s$.
2. Lorsque les données vérifient parfaitement le modèle, l'AFM donne systématiquement la bonne solution et hiérarchise les facteurs par ordre d'importance décroissante (au sens de la somme des poids). En effet, si les nuages $N_I^j$ correspondent au modèle, alors le nuage moyen $N_I$ le vérifie aussi.

3. L'algorithme qui fournit l'estimation est une diagonalisation et ne pose aucun problème de convergence.
4. L'interprétation de l'estimation comme une projection permet d'introduire très simplement des groupes de variables supplémentaires.
5. Cette estimation est incluse dans une analyse complète : les résultats associés aux autres points de vue de l'AFM 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.

### 8.5.5 Cas des tableaux de distances et de similarités (AFMTD)

Lorsque l'on souhaite étudier au travers du modèle INDSCAL des données recueillies sous la forme de tableaux de distances, l'AFM ne s'applique qu'indirectement. À chaque tableau de distances, on peut associer un tableau de type *individus*×*variables* en lui appliquant une Analyse Factorielle sur Tableau de Distances (*cf.* section 5.4.5 page 120) : les variables dans ce nouveau tableau sont les facteurs de son AFTD. Les tableaux ainsi obtenus peuvent être juxtaposés et l'ensemble soumis à une AFM dans laquelle chaque groupe comprend les facteurs issus d'un même tableau de distances.

Si les distances sont des distances euclidiennes et si, de plus, on considère tous les facteurs, l'AFM analyse dans $R^{I^2}$ les matrices de produits scalaires qui correspondent exactement aux distances initiales. En revanche, si les distances ne sont pas euclidiennes, on ne conserve que les facteurs associés aux valeurs propres positives (*cf.* section 5.4.5) ; l'AFM analyse alors des approximations euclidiennes des données. Les tableaux de similarités peuvent être traités de la même façon en les transformant préalablement en tableaux de distances.

Cette méthodologie, AFTD par tableau de distances puis AFM sur les facteurs associés aux valeurs propres positives des AFTD, est appelée AFM sur tableaux de distances (AFMTD).

## 8.6 CAS DES VARIABLES QUALITATIVES ET DES TABLEAUX MIXTES

La problématique analysée dans le cas de variables quantitatives s'étend sans modification fondamentale aux variables qualitatives : équilibre entre les groupes, recherche de facteurs communs, comparaison globale des groupes, construction d'une représentation superposée, etc. Nous montrons dans cette section que l'AFM s'applique aux tableaux disjonctifs complets dans lesquels les variables sont structurées en groupes.

L'essentiel de la démonstration réside dans l'équivalence entre l'ACM d'une part et l'ACP appliquée aux variables indicatrices pondérées de manière adéquate d'autre part.

Le fait qu'une même technique, l'AFM, s'applique aussi bien à des variables quantitatives que qualitatives, suggère à son tour le traitement simultané de variables des deux types (le tableau des données est dit alors « mixte »). Ce type de tableau est très répandu : son analyse pose un certain nombre de problèmes inhérents à la différence de nature entre les objets que l'on souhaite étudier simultanément. Il a déjà été traité lorsque seules les variables quantitatives sont actives (ACP, *cf.* § 1.10 page 27), lorsque seules les variables qualitatives sont actives (ACM *cf.* section 4.5 page 101) ou lorsque les deux le sont mais sans structure de groupes (AFDM, *cf.* section 4.6 page 104).

### 8.6.1 Équivalence entre ACM et ACP pondérée des indicatrices

Les résultats de l'ACM peuvent être obtenus à partir d'une ACP normée des variables indicatrices (c'est-à-dire du TDC), à condition d'associer à celles-ci des poids adéquats. L'argumentation s'opère en trois temps, détaillés dans les paragraphes suivants.

1. Mise en avant de deux propriétés du nuage des modalités en ACM. A ce niveau, le nuage est considéré par rapport à l'origine (et non par rapport à son centre de gravité).
2. Construction d'un nuage des indicatrices de modalités, ayant les mêmes propriétés inertielles que le nuage des modalités en ACM, en vue de son traitement par ACP normée. À ce niveau, les indicatrices ne sont pas centrées.
3. Équivalence entre les opérations de centrage de l'ACM et de l'ACP lorsqu'elles s'effectuent sur les nuages définis précédemment.

#### *a) Propriétés des modalités en ACM*

Du fait de la transformation des colonnes en profils, de la métrique dans $R^I$ (proportionnelle à la métrique identité) et des poids des éléments, les modalités en ACM possèdent les propriétés suivantes lorsqu'on les considère par rapport à l'origine :

1. les modalités d'une même variable sont orthogonales entre elles ; la transformation en profil ne change pas leur direction ;
2. chaque modalité possède la même inertie par rapport à l'origine ($I_k$ : nombre d'individus possédant la modalité $k$ ; $x_{ik} = 0$ ou 1) :

$$\text{Inertie de } k \text{ par rapport à } O = \frac{I_k}{IJ} \sum\nolimits_i I \left( \frac{x_{ik}}{I_k} \right)^2 = \frac{1}{J}$$

*b) Pondération des indicatrices en vue d'un traitement par ACP normée*

Considérons, dans $R^I$, le nuage des indicatrices non centrées mais divisées par leur écart-type. Si l'on affecte à chaque indicatrice $k$ le poids $(I - I_k)/I$, alors le nuage ainsi défini possède les mêmes propriétés inertielles que celui analysé en ACM (*cf.* paragraphe précédent). Soit :

1. la métrique de l'espace $R^I$ est aussi la métrique identité au facteur $1/I$ près ;
2. la direction des indicatrices n'est pas modifiée par la division par l'écart-type (de même qu'en ACM, elle n'est pas modifiée par la transformation en profil) ;
3. chaque indicatrice possède la même inertie par rapport à l'origine :

$$\text{Inertie de } k \text{ par rapport à } O = \frac{I - I_k}{I} \sum\nolimits_i \frac{1}{I} x_{ik} / \left( \frac{I_k(I - I_k)}{I^2} \right) = 1$$

*c) Équivalence entre les deux centrages*

En ACP, le centrage des variables s'interprète dans l'espace $R^I$ comme une projection du nuage des variables sur l'hyperplan orthogonal à la première bissectrice.

En ACM, vue comme une AFC appliquée à un TDC, le nuage des indicatrices est centré en un autre sens : l'origine est placée au centre de gravité $G_K$ du nuage $N_K$.

Or, en ACM, le nuage $N_K$ des modalités présente les propriétés suivantes :

1. le centre de gravité $G_K$ est situé sur la première bissectrice (la marge sur les lignes est constante) ;
2. le nuage $N_K$ est contenu dans un hyperplan orthogonal à la première bissectrice.

Il en résulte que, appliqué à ce nuage des modalités, le centrage en ACM s'interprète comme en ACP : une projection sur l'hyperplan orthogonal à la première bissectrice.

En conclusion, une ACP normée des indicatrices pondérées conduit aux mêmes facteurs sur $I$ qu'une ACM (les inerties des facteurs des deux analyses sont égales, au coefficient $J$ près).

### 8.6.2 Variables qualitatives et tableaux mixtes en AFM

*a) Principe général*

De l'équivalence précédente, il résulte que l'on peut appliquer des méthodes factorielles construites pour des variables quantitatives à des variables qualitatives à condition de faire intervenir ces dernières à l'aide de leurs indicatrices pondérées de façon adéquate. Dans cet esprit, l'AFM peut traiter des tableaux d'indicatrices pondérées : il est ainsi possible d'étendre aux variables qualitatives la méthodologie liée aux groupes de variables proposée initialement pour les variables quantitatives.

Cette extension du champ d'application de l'AFM est renforcée par le résultat suivant. Sur un ensemble de variables qualitatives, il est équivalent de réaliser :

1. une ACM ;
2. une AFM dans laquelle chaque groupe est constitué par l'ensemble des indicatrices associées à une même variable.

Ce résultat s'obtient en montrant que, lorsqu'on applique l'AFM à un ensemble d'indicatrices préalablement pondérées comme indiqué en section 8.6.1, les coefficients de pondération spécifiques de l'AFM (inverses des premières valeurs propres de chacun des groupes) sont tous égaux à 1.

En effet, le nuage des indicatrices associées à une même variable possède une inertie de 1 dans toutes les directions du sous-espace qu'elles engendrent. Cela se montre directement lorsqu'on se place avant le centrage : les indicatrices sont deux à deux orthogonales et l'inertie de chacune vaut 1. Ce résultat est conservé après le centrage puisque, au niveau du sous-espace engendré par les modalités d'une même variable, cette opération ne fait que retirer une dimension (*i.e.* la première bissectrice) au sous-espace précédent.

Il en résulte que les valeurs propres des ACP de chacun des groupes sont toutes égales à 1. Sur un tel tableau, l'AFM conduit alors aux mêmes facteurs que l'ACP et est donc équivalente à une ACM (au coefficient $J$ près pour les valeurs propres, celles-ci étant la moyenne des rapports de corrélation en ACM – *cf.* § 4.3.6 page 96 – et leur somme en AFM).

L'inertie du nuage des indicatrices d'une même variable $j$ valant 1 dans toutes les directions du nuage qu'elles engendrent, l'opérateur $W_j D$ est égal au projecteur $P_j$ et l'AFM dans ce cas (et donc aussi l'ACM) se confond avec une analyse multicanonique.

La représentation des groupes de l'AFM correspond alors à celle des variables en ACM proposée en section 4.3.7 page 98. puisque la contribution d'une variable qualitative à un axe est son rapport de corrélation avec cet axe.

En ce sens, l'AFM généralise l'ACM et est susceptible de traiter des variables qualitatives. Elle apporte une solution technique pour aborder la problématique associée à l'étude simultanée de plusieurs groupes de variables lorsque ces dernières sont qualitatives.

De façon analogue, l'AFM généralise l'ACP : si chaque groupe est réduit à une seule variable quantitative, l'AFM est équivalente à une ACP normée. La même technique qui s'applique à des variables qualitatives ou à des variables quantitatives s'applique donc aussi à un mélange des deux à condition de résoudre le problème de leurs pondérations respectives. Or la pondération de l'AFM, qui équilibre l'inertie et donc l'influence *a priori* des groupes, élimine les déséquilibres éventuels entre groupes induits aussi bien par des différences de structure que par des différences de types de variables. Cet équilibre peut être vu au travers du critère optimisé par la composante

$z_s$ de l'AFM, écrit ci-après dans le cas de deux groupes, le premier comportant $K$ variables quantitatives $v_k$ et le second $Q$ variables qualitatives $V_q$.

$$\frac{1}{\lambda_1^1}\sum_{k\in K} r^2(z_s,k) + \frac{1}{Q\lambda_1^2}\sum_{q\in Q}\eta^2(z_s,V_q)$$

avec $\lambda_1^1$, la première valeur propre de l'ACP du groupe 1, $\lambda_1^2$ la première valeur propre de l'ACM du groupe 2.

L'écriture des inerties via les coefficients et rapports de corrélation montre, comme en AFDM, l'équilibre entre les deux types de variables ; les valeurs propres des analyses séparées assurent l'équilibre entre les groupes (comme dans toute AFM). Il est clair que si chaque groupe est réduit à une seule variable, quantitative ou qualitative, l'AFM est équivalente à une AFDM. Cette nouvelle équivalence suggère à son tour la possibilité de traiter des groupes de variables mixtes, dont l'analyse séparée est une AFDM. L'AFM permet ainsi de traiter simultanément en actif des groupes quantitatifs, qualitatifs ou mixtes.

### *b) Représentation des modalités en AFM*

Remarquons que, dans une ACP normée des indicatrices, les projections des colonnes sont les corrélations entre les indicatrices et les facteurs sur $I$. Elles ne représentent pas, comme en ACM, les centres de gravité des classes d'individus définies par les modalités.

Cette dernière représentation étant essentielle dans les interprétations, il est nécessaire de l'ajouter. En pratique, seule cette représentation des modalités, en tant que centre de gravité d'individus, est utilisée, parce qu'elle est habituelle (*cf.* ACM) mais aussi parce qu'elle s'intègre dans la représentation superposée. En effet on peut calculer le centre de gravité d'un ensemble d'individus vus par l'ensemble des variables mais aussi par chacun des groupes (*cf.* Figure 7.9 page 168).

Cette dernière représentation est très importante car elle permet à l'AFM d'aborder des fichiers d'enquête assez volumineux dans lesquels les individus ne sont pas intéressants en eux-mêmes mais uniquement au travers des modalités qu'ils possèdent. La projection de ces centres de gravité est accompagnée des aides à l'interprétation usuelles en particulier la contribution à l'inertie de chaque facteur. La somme de ces contributions, pour les modalités d'une même variable, est égale au carré du rapport de corrélation entre la variable et le facteur $F_s$.

### *c) Données manquantes et modalités de faible poids*

En ACM, les modalités de très faible effectif sont souvent une source de perturbation des résultats. En outre, le problème des données manquantes se pose, là comme dans tout traitement de données. En effet, la construction d'une modalité supplémentaire

*donnée manquante*, outre le fait qu'elle conduit fréquemment à des modalités de faible effectif, n'est véritablement satisfaisante que si la donnée manquante possède une signification, comme c'est souvent le cas des non-réponses volontaires dans les enquêtes d'opinion.

Pour résoudre ce problème, nous avons indiqué en section 6.3.2 page 131 une méthode dérivée de l'ACM qui en améliore les possibilités en présence de données manquantes et de modalités de faible poids. Elle traite des tableaux disjonctifs incomplets dans lesquels les données manquantes sont codées par des zéros et les modalités rares supprimées. La marge de ces tableaux, contrairement aux tableaux disjonctifs complets, n'est pas constante. Or la plupart des propriétés intéressantes de l'analyse des tableaux disjonctifs complets sont liées à cette marge constante. Le principe de cette variante de l'ACM est de remplacer la marge du tableau incomplet par une marge constante, partout où elle intervient (en particulier dans les poids des individus et donc dans la métrique dans $R^I$ ).

Lorsqu'il y a des données manquantes, il est possible d'appliquer une ACP pondérée sur le tableau incomplet d'indicatrices. On montre que cette ACP, et par conséquent l'AFM, est équivalente à la variante de l'ACM esquissée ci-dessus (l'inertie du nuage des colonnes et la métrique de l'espace sont identiques entre ces deux analyses).

## 8.7 ELÉMENTS SUPPLÉMENTAIRES

### 8.7.1 Individus supplémentaires

Comme dans toute ACP ou toute ACM, des individus peuvent intervenir en tant qu'éléments supplémentaires dans une AFM, 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 $I(J + 1)$, soit le nombre d'individus multiplié par le nombre de groupes de variables augmenté de 1 (pour le nuage moyen). La lecture des aides à l'interprétation 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.

### 8.7.2 Groupes de variables supplémentaires

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 œuvre 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 :

1. ***normalisation du nuage*** $N_K^j$ : pour comparer aux autres nuages le nuage associé à un groupe supplémentaire, il faut le normaliser de la même façon en surpondérant les variables du groupe par l'inverse de la première valeur propre de son analyse séparée ;
2. ***projection des composantes principales du groupe*** : elle permet de comparer la forme générale du nuage $N_K^j$ avec celle du nuage moyen $N_K$ et celles des nuages associés aux autres groupes de variables ;
3. ***projection des $W_j D$ :*** 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 $W_j D$ sur l'axe de rang $s$ coïncide avec la mesure de liaison entre le $s^{\text{e}}$ 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).

Par contre, il n'est pas possible d'obtenir une représentation superposée des nuages d'individus associés à des groupes de variables supplémentaires : cela reviendrait à projeter un nuage$N_I^j$ sur un axe de $R^K$ orthogonal au sous-espace qui contient ce nuage.

Si l'on étudie les raisons qui conduisent à introduire un groupe en supplémentaire, on peut voir que cette limite n'est pas très gênante. Si l'on craint qu'il 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.

## 8.8 MISE EN ŒUVRE DE L'ANALYSE FACTORIELLE MULTIPLE

La mise en œuvre de l'AFM comprend deux étapes.

Dans la première étape, on analyse chaque groupe séparément ; lorsqu'il s'agit d'un groupe de variables qualitatives (resp. mixtes), on réalise une ACP pondérée équivalente à une ACM (resp. AFDM). Cette première étape est nécessaire pour calculer :

1. l'inverse de la première valeur propre de l'ACP de chaque groupe, qui pondère (ou surpondère) les variables dans la seconde étape ;
2. les facteurs de chaque groupe.

La seconde étape est une ACP de l'ensemble des variables de tous les groupes pondérés ; en pratique, on réalise cette analyse à partir des facteurs des analyses séparées. En effet, il est équivalent de considérer un tableau du point de vue de ses données brutes ou du point de vue de ses facteurs. On se limite aux facteurs associés à une valeur propre non nulle, ce qui réduit la dimension de la matrice à diagonaliser. Dans la perspective du traitement de très grands tableaux, on peut aussi ne pas prendre en compte les facteurs associés à des petites valeurs propres, ce qui conduit à une analyse approchée, la qualité de l'approximation étant liée au seuil en dessous duquel on écarte les valeurs propres.

Chapitre **9**

# Méthodologie de l'AFM

Ce chapitre regroupe d'abord plusieurs aspects utiles dans la mise en œuvre de l'AFM. Ils s'articulent autour de deux thèmes : tactique méthodologique et aides à l'interprétation. Le chapitre se termine par une présentation synthétique d'une extension de l'AFM, l'AFM hiérarchique (AFMH), dédiée aux tableaux dans lesquels les variables sont structurées selon plusieurs partitions emboîtées.

## 9.1 TACTIQUE MÉTHODOLOGIQUE

### 9.1.1 AFM et analyses séparées

Les résultats de l'AFM et ceux des analyses séparées des groupes de variables se complétant, il est souvent utile de les effectuer toutes. L'expérience montre que, dans ce cas, il est préférable de réaliser d'abord l'AFM afin d'avoir une vision globale des données et des relations entre les groupes. L'exemple de l'enquête Ouest-France montre l'intérêt de cette démarche : par le jeu des indices de comparaison entre groupes, l'AFM indique d'emblée (*cf.* section 7.2 page 164) qu'il existe un facteur commun aux deux groupes et que le groupe « profil de lecture » possède un facteur spécifique important. En outre, l'AFM relie ces facteurs à ceux des analyses séparées et donne les éléments nécessaires à leur interprétation. À ce stade, on peut décider en toute connaissance de cause de réaliser ou non l'une et/ou l'autre de ces analyses et éviter des tâtonnements. Si par exemple un groupe de variables se révèle très différent des autres, il est généralement inutile de le mêler aux autres et nécessaire de l'étudier séparément.

### 9.1.2 Cas dans lequel les variables sont homologues d'un groupe à l'autre

*a) Analyses factorielles de tableaux juxtaposés et AFM*

Pour fixer les idées, nous considérons le cas d'une suite de $J$ tableaux, dans lesquels les mêmes $K_g$ variables quantitatives sont mesurées sur les mêmes $I$ individus, indicée par le temps. L'ACP usuelle offre deux voies pour décrire l'évolution des données (*cf.* **Figure 9.1**).

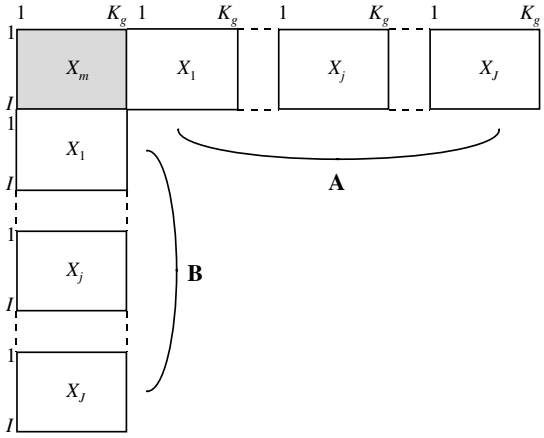



**Figure 9.1** Les deux juxtapositions d'un ensemble de tableaux doublement appariés. La juxtaposition A (resp. B) en ligne (resp. colonne) est possible du fait de l'homologie entre les lignes (resp. colonnes) des tableaux. $K_g$ : nombre de variables par groupe ; $X_m$ : tableau moyen.

L'ACP des $J$ tableaux juxtaposés en ligne (**A**) fournit une représentation des individus, chacun considéré du point de vue de l'ensemble des $J$ dates. Elle fournit une représentation des variables dans laquelle chacune donne lieu à un point par date. Elle permet ainsi de suivre l'évolution des $K_g$ variables au cours du temps. Dans cette approche, l'homologie entre les variables n'est pas utilisée dans les calculs mais seulement lors de l'interprétation (en reliant sur les plans factoriels, par exemple, les points relatifs à une même variable). En revanche, l'observation des mêmes individus au cours du temps est ici essentielle.

L'ACP des $J$ tableaux juxtaposés en colonne (**B**) permet de suivre l'évolution des individus au cours du temps puisqu'une ligne correspond à un individu à une date donnée. On peut centrer chaque tableau avant la juxtaposition si l'on souhaite ne pas faire apparaître l'évolution globale des individus. Dans cette analyse, l'homologie entre les variables est utilisée. En revanche, le fait que ce sont toujours les mêmes individus qui ont été observés au cours du temps n'est pas nécessaire dans cette ACP.

L'ACP usuelle offre ainsi la possibilité d'étudier l'évolution des individus et celle des variables. Chaque évolution est décrite dans un cadre différent puisque issu d'un traitement différent. Cette méthodologie est, à juste titre, très utilisée (on retrouve cette démarche, pour les tableaux de fréquence, au chapitre 10). Par rapport à cette méthodologie, l'AFM des $J$ tableaux juxtaposés en ligne (dans laquelle chaque ensemble des mesures à une date donnée constitue un groupe de variables) offre les caractéristiques intéressantes suivantes :

1. les groupes de variables étant pondérés, l'influence des différentes dates de mesure est équilibrée ;
2. grâce à la représentation superposée, on dispose d'une visualisation de l'évolution des individus et des variables au sein d'une même analyse ;
3. du fait de la pondération et de la prise en compte explicite de la structure en groupes des variables, on dispose d'un large éventail d'aides à l'interprétation (représentation des groupes, des facteurs des analyses séparées, etc.).

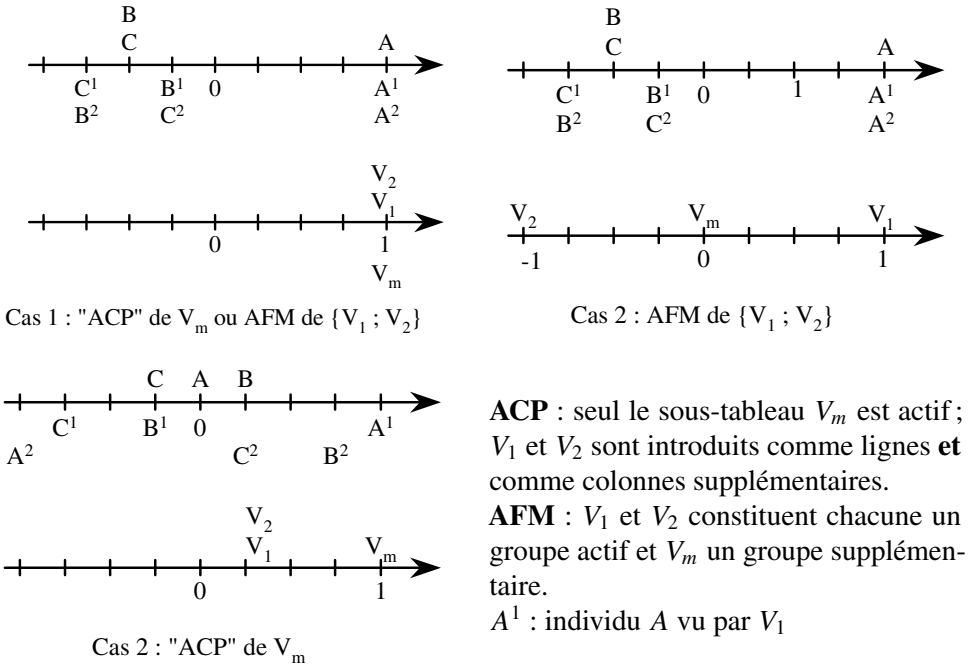
### *b) Analyse factorielle d'un tableau moyen et AFM*

L'ACP usuelle permet aussi dans certains cas de représenter l'évolution des individus et celle des variables au sein d'une même analyse. Pour cela, on construit le tableau $X_m$, moyenne des tableaux $X_j$ (on prendra soin de vérifier que ce tableau des moyennes a un sens ; en particulier, si les écarts-types diffèrent entre variables homologues, il peut être nécessaire de centrer et réduire les tableaux $X_j$ avant d'en faire la moyenne). L'ACP peut alors être appliquée au tableau $X_m$ en actif, les tableaux $X_j$ étant introduits à la fois en tant que lignes et colonnes supplémentaires. Cette méthodologie est surtout utilisée dans le cadre des tableaux de fréquence (*cf.* chapitre 10 en particulier figure 10.3 page 230). Elle s'appuie sur l'homologie entre les variables, c'est-à-dire que la structure commune aux tableaux n'est mise en évidence que si les variables homologues sont corrélées positivement entre elles.

**Tableau 9.1** Deux cas de données choisies. Le premier (resp. second) groupe de variables se limite à la variable $V_1$ (resp. $V_2$). $V_m$ : moyenne entre $V_1$ et $V_2$.

| Individu | Cas 1 | | | Cas 2 | | |
|---|---|---|---|---|---|---|
| | $V_m$ | $V_1$ | $V_2$ | $V_m$ | $V_1$ | $V_2$ |
| A | 4 | 4 | 4 | 0 | 4 | -4 |
| B | -2 | -1 | -3 | 1 | -1 | 3 |
| C | -2 | -3 | -1 | -1 | -3 | 1 |

Le cas de deux groupes réduits chacun à une seule variable quantitative centrée est commode pour illustrer la comparaison entre AFM de groupes de variables homologues et ACP du tableau moyen (*cf.* **Tableau 9.1** et **Figure 9.2**). Notons $V_1$ et $V_2$ ces




**ACP** : seul le sous-tableau $V_m$ est actif ; $V_1$ et $V_2$ sont introduits comme lignes **et** comme colonnes supplémentaires.
**AFM** : $V_1$ et $V_2$ constituent chacune un groupe actif et $V_m$ un groupe supplémentaire.
$A^1$ : individu $A$ vu par $V_1$

**Figure 9.2** Axe unique de l'ACP du tableau moyen ou premier axe de l'AFM du tableau 9.1. Pour chaque cas, représentation des individus (en haut) et des variables.

deux variables et $r_{12}$ leur coefficient de corrélation. Le cœur de l'AFM de $\{V_1; V_2\}$ est une ACP normée des deux variables. On vérifie aisément que $V_1 + V_2$ et $V_1 - V_2$ sont vecteurs propres de la matrice des corrélations (*cf.* section 5.3.1 page 112) et donc que cette ACP admet comme composantes principales la somme $V_1 + V_2$ (inertie associée : $1 + r_{12}$) et la différence $V_1 - V_2$ (inertie associée : $1 - r_{12}$).

Le signe de $r_{12}$ détermine l'ordre de ces deux composantes.

1. $r_{12} > 0$ : la structure commune respecte l'homologie entre les variables. Exemple : cas 1 du tableau 9.1, dans lequel les variables $V_1$ et $V_2$ mettent toutes deux en évidence la forte valeur de A. L'AFM de $\{V_1, V_2\}$ et l'ACP du tableau moyen (i.e. de $V_m$) conduisent aux mêmes représentations. Exemple : *cf.* **Figure 9.2** cas 1.

2. $r_{12} < 0$ : la structure commune ne respecte pas l'homologie entre les variables. Exemple : cas 2 du tableau 9.1, dans lequel les variables $V_1$ et $V_2$ mettent toutes deux en évidence l'éloignement de A mais chacune dans un sens différent. L'ACP du tableau moyen ne peut déceler cette structure : elle place le point A à l'origine

des axes (*cf.* **Figure 9.2** cas 2). L'AFM de $\{V_1, V_2\}$, pour son premier axe, prend en quelque sorte l'opposée de l'une des variables avant de les superposer.

Dans l'ACP de $V_m$, qui se réduit bien sûr à la représentation de $V_m$, $V_1$ et $V_2$ étant introduits à la fois en lignes et colonnes supplémentaires, on prend en compte l'homologie entre les variables, ce qui inclut le sens de variation des $V_1$ et $V_2$ : ainsi, dans le cas 2, la structure commune se limite alors à des valeurs de B généralement plus élevées que celles de C. C'est bien ce que met en évidence l'axe unique de l'« ACP » de $V_m$ (*cf.* **Figure 9.2** cas 2). La représentation des variables initiales ($V_1$ et $V_2$), qui dans cette méthodologie ne peuvent apparaître qu'en fonction de leur liaison avec la structure commune, est ici proche de l'origine.

Dans l'AFM de $\{V_1, V_2\}$, $V_m$ étant introduite en supplémentaire, on ne prend pas en compte l'homologie entre les variables ; la structure commune majeure est alors le particularisme de l'individu A, ce que met bien en évidence le premier axe de l'AFM (*cf.* **Figure 9.2**). Les points partiels restituent bien les données : A est extrême du point de vue des 2 groupes ; C est extrême du point de vue du groupe 1 et non du point de vue du groupe 2. La variable moyenne $V_m$ est non corrélée à l'axe 1. En revanche, elle est parfaitement corrélée à l'axe 2 de l'AFM qui correspond donc à l'axe 1 de l'ACP.

En conclusion, on réservera l'ACP du tableau moyen en actif au cas où l'on s'intéresse aux seules structures communes respectant l'homologie des variables. Autrement, lorsque toutes les structures communes sont dignes d'intérêt, on réalisera une AFM, en introduisant le tableau moyen en tant que groupe supplémentaire.

### 9.1.3 Définition et statut des groupes de variables

#### *a) Définition des groupes*

Dans la plupart des cas, le regroupement des variables s'impose, tout simplement parce que la notion de groupe s'insère directement dans la problématique et a été utilisée dans la définition des données à recueillir. L'enquête *Ouest-France* en est un exemple simple mais typique : la mise en relation de la lecture et de repères sociaux fait partie des objectifs définis préalablement à l'étude, guide la rédaction du questionnaire et apparaît dont tout naturellement dans les traitements. Il en est de même dans la plupart des questionnaires, presque toujours structurés en thèmes.

Des hésitations peuvent toutefois apparaître lorsque les thèmes sont eux-mêmes structurés en sous-thèmes. Il n'est bien sûr pas possible de donner de règles générales quant au choix des groupes dans ce cas. Il faut rappeler toutefois la possibilité d'introduire plusieurs fois les données dans l'analyse : on peut ainsi réaliser l'AFM sur les données structurées en thèmes et introduire les données structurées en sous-thèmes en supplémentaire et voir ainsi apparaître les sous-thèmes dans le carré des liaisons.

Un exemple simple de données introduites deux fois dans une analyse est fourni par les données *vins de Loire* examinées au chapitre 7.

Les variables *appellation* et *terroir* ont été introduites au sein d'un même groupe, *origine des vins*, mais peuvent aussi être considérées séparément. La représentation de ces deux nouveaux groupes (*cf.* **Figure 7.6 page 162**) a montré que la liaison entre l'origine des vins et les deux premiers facteurs est due au terroir et non à l'appellation.

Un autre exemple dans lequel plusieurs définitions des groupes sont possibles est fourni par les observations répétées d'un même ensemble de variables, cas déjà évoqué section 9.1.2. On peut, dans ce cas, regrouper les variables de deux façons et la formulation de l'AFM en tant que méthode de recherche de facteurs communs aide à choisir entre les deux partitions des variables.

1. Partition 1 : un groupe rassemble les variables d'une même date (cas envisagé section 9.1.2). L'AFM cherche alors les facteurs communs aux structures sur les individus définies par les différentes dates (question : qu'y a-t-il de commun aux différentes dates ?).
2. Partition 2 : un groupe rassemble les variables de même nature, toutes dates confondues ; il représente l'évolution de la variable tout au long de la période étudiée. L'AFM cherche alors les facteurs communs à ces évolutions (question : qu'y a-t-il de commun aux évolutions des différentes variables ?).

### *b) Statut des groupes*

Comme dans les analyses factorielles usuelles, le statut des éléments, actif ou supplémentaire, s'impose dans beaucoup de cas mais mérite quelquefois une discussion. La démarche en AFM est identique à celle des autres analyses factorielles lorsque l'on considère l'AFM comme… une analyse factorielle. Ainsi, dans l'exemple des vins, la volonté de rechercher les principaux facteurs de variabilité sensorielle conduit d'emblée à introduire les variables *appellation* et *terroir* en supplémentaire. Le cas des variables d'ensemble (t*ypicité* et *qualité d'ensemble*) est plus nuancé puisque ces variables peuvent être considérées comme sensorielles, mais leur spécificité a finalement conduit à les écarter des éléments actifs.

Des points de vue spécifiques apparaissent lorsque l'on considère l'AFM comme méthode de recherche de facteurs communs. Une illustration en est fournie par un autre exemple, issu lui aussi du domaine des vins.

Pour un ensemble de vins, on dispose de variables :

1. physico-chimiques mesurées sur la vendange ;
2. physico-chimiques mesurées sur les vins ;
3. sensorielles.

On peut vouloir donner à ces trois groupes le statut *actif*. Ce faisant, on recherche les facteurs communs, à la vendange, à la physico-chimie du vin et à la description sensorielle. Cette problématique est ambitieuse en ce sens qu'elle vise des facteurs

communs aux trois groupes. Aussi peut-on préférer se limiter à introduire seulement deux groupes en actif. Ce qui revient à chercher des facteurs communs :

1. soit à la physico-chimie des vendanges et à celle des vins ;
2. soit à la physico-chimie des vins et à la description sensorielle des vins.

Ces deux dernières problématiques sont moins ambitieuses mais plus faciles à concevoir. Aussi, pour de telles données, une démarche empirique mais raisonnable consiste à commencer par une AFM avec les trois groupes actifs en s'attendant à conserver comme analyse(s) finale(s) celle(s) avec deux groupes actifs.

## 9.2 AIDES À L'INTERPRÉTATION

L'AFM fournit un grand nombre d'aides à l'interprétation. Certaines de ces aides ne lui sont pas spécifiques et se retrouvent dans toute analyse factorielle. Soit, principalement :

1. l'inertie et le pourcentage d'inertie associés à chaque axe ;
2. les contributions des lignes et des colonnes à l'inertie de chaque axe ;
3. les qualités de représentation ( = cosinus carré) des lignes et des colonnes par chaque axe ;
4. la distance (dans l'espace complet) entre chaque individu et l'origine ;
5. les valeurs-tests associées aux modalités des variables qualitatives.

Parmi les aides spécifiques, certaines ont déjà été définies et commentées dans l'un et/ou l'autre exemple. Soient :

1. les corrélations entre les facteurs du nuage moyen et les facteurs des nuages partiels (*cf.* section 7.1.6 page 159) ;
2. les contributions des groupes de variables à l'inertie des axes (*cf.* tableau 7.1 et section 7.1.7 page 161) ;
3. le rapport [*inertie inter* / *inertie intra*] associé à la représentation superposée (*cf.* section 7.2.4 page 167).

Enfin certaines aides n'ont pas été illustrées. Elles font l'objet des sections suivantes.

### 9.2.1 Mesures globales de liaison entre deux groupes de variables

Deux mesures de liaisons entre groupes de variables sont présentées en section 8.4.3 page 189 : $L_g$ et RV. Elles se complètent bien, comme l'illustre l'exemple des vins (*cf.* **Tableau 9.2**). Dans cet exemple :

1. $\mathrm{RV}(1,3) = .71 \approx \mathrm{RV}(2.4) = .75$ ; les groupes 1 et 3 d'une part et 2 et 4 d'autre part ont des structures voisines, également proches de l'homothétie ;
2. $\mathrm{L_g}(1,3) = 1.05 > \mathrm{L_g}(2,4) = .80$ ; la structure commune aux groupes 1 et 3 est plus riche que la structure commune aux groupes 2 et 4.

L'indicateur $\mathrm{L_g}(j,\, j)$ n'est rien d'autre que l'indicateur de dimensionalité $\mathrm{N_g^2}(j)$ défini en section 8.4.2 page 189 et illustré en 9.2.2.

**Tableau 9.2** Exemple des vins : mesures $\mathrm{L_g}$ et RV de liaison entre groupes.

| Groupe | $\mathrm{L}_g$ | | | | | RV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | $\sum$ | 1 | 2 | 3 | 4 | $\sum$ |
| 1 : olfaction au repos | 1.61 | | | | | 1 | 1 | | | |
| 2 : vision | .55 | 1.00 | | | | .44 | 1 | | | |
| 3 : olfaction après agitation | 1.05 | .70 | 1.37 | | | .71 | .60 | 1 | | |
| 4 : gustation | .68 | .80 | .94 | 1.12 | | .51 | .75 | .76 | 1 | |
| $\sum = \{1,2,3,4\}$ | 1.13 | .88 | 1.17 | 1.02 | 1.22 | .81 | .80 | .91 | .88 | 1 |

Ces indicateurs s'appliquent aussi à un groupe rassemblant les variables de plusieurs groupes $j$, chaque variable du groupe $j$ étant au préalable pondérée par $1/\lambda_1^j$. Ces calculs sont surtout intéressants lorsque l'on considère l'ensemble des groupes actifs (noté $\sum$ dans le **tableau 9.2**), chaque variable étant pondérée selon l'AFM. Dans l'exemple des vins :

1. la dimensionalité de l'ensemble de la dégustation (1.22) est plus faible que celle des deux olfactions (1.61 et 1.37) ; ceci, qui peut paraître paradoxal, est dû au fait que l'écart entre l'inertie du premier facteur et celle des autres facteurs est plus important dans l'AFM globale que dans ces deux groupes ;
2. ce sont, d'après les mesures $\mathrm{L_g}$, les deux olfactions qui ont la plus riche structure commune avec l'ensemble de la dégustation ;
3. mais ce sont, d'après les mesures RV, l'olfaction après agitation et la gustation qui ont la structure la plus proche de celle de l'ensemble de la dégustation.

### 9.2.2 Aides relatives à la représentation des groupes de variables

Les diverses interprétations des coordonnées des groupes de variables sont rassemblées à la section 8.4.4 page 191 et un exemple se trouve en section 7.1.7 page 161. Ces coordonnées sont accompagnées de plusieurs indicateurs présentés ci-après.

#### *a) Distance entre un groupe et l'origine*

Le carré de la distance entre un groupe et l'origine, somme des carrés des valeurs propres du groupe après pondération de l'AFM, constitue le critère de dimensionalité du groupe noté $\mathrm{N_g^2}$. Dans l'exemple des vins (*cf.* **Tableau 9.3**), cet indicateur met

clairement en évidence la faible dimensionalité de chacun des groupes quantitatifs, en particulier des groupes *gustation*, *vision* et *appréciation d'ensemble*. La multidimensionalité du groupe *origine* tient bien sûr au codage disjonctif complet, caractéristique classique de l'ACM (*cf.* section 4.3.5 page 95).

**Tableau 9.3** Exemple des vins : distance $d(0, j)$ des groupes à l'origine dans $R^{I^2}$. $d^2(0, j) = N_g^2$ mesure « le nombre de directions d'inertie comparable à celle de la première direction ».

| groupe $j$ | | % d'inertie des analyses séparées | | |
|---|---|---|---|---|
| | $d^2(0, j)$ | F1 | F2 | F3 |
| Olfaction au repos | 1.610 | 44.8 | 30.3 | 16.3 |
| Vision | 1.003 | 94.5 | 5.0 | .5 |
| Olfaction après agitation | 1.369 | 47.0 | 24.8 | 10.5 |
| Gustation | 1.123 | 62.7 | 19.9 | 7.5 |
| Appréciation d'ensemble | 1.007 | 92.5 | 7.5 | - |
| Origine | 2.645 | 29.0 | 25.6 | 20.0 |

$d^2(0, j)$ mesure, en quelque sorte, « le nombre de directions dont l'inertie est proche de l'inertie axiale maximum ». Ainsi, selon ce critère, le groupe *olfaction au repos* est de dimensionalité plus importante que *olfaction après agitation* du fait du deuxième axe, d'inertie plus proche de celle du premier axe dans le cas de l'olfaction au repos.

Le **tableau 9.4** illustre ce phénomène dans deux cas de référence. Le sous-espace engendré dans le cas 1 est à 3 dimensions (3 valeurs propres non nulles). Dans le cas 2, il est à 6 dimensions. Malgré cela, l'indicateur de dimensionalité est plus important dans le cas 1 car il accorde une importance très faible aux dimensions de faible inertie.

**Tableau 9.4** Distance d'un groupe $j$ à l'origine dans 2 cas de référence décrits par leurs pourcentages d'inertie.

| | % d'inertie | | | | | | |
|---|---|---|---|---|---|---|---|
| | F1 | F2 | F3 | F4 | F5 | F6 | $d^2(0, j)$ |
| Cas 1 | .5 | .4 | .1 | 0 | 0 | 0 | 1.68 |
| Cas 2 | .5 | .1 | .1 | .1 | .1 | .1 | 1.20 |

Plus précisément :

1. la valeur 1.68 (proche de 2) traduit la présence de 2 dimensions prépondérantes d'inerties comparables ;
2. la valeur 1.20 (proche de 1) traduit la présence d'une seule dimension prépondérante.

### b) Contribution d'un groupe à l'inertie d'un axe

Cette notion est définie clairement dans l'espace des variables comme la somme des contributions (inerties projetées) des variables d'un même groupe. Dans l'espace des groupes de variables, cet indicateur est la coordonnée d'un groupe (dans l'espace des groupes, la quantité maximisée est la somme des coordonnées ; *cf.* section 8.4.4 page 191). Ces contributions (absolues) sont souvent exprimés en % (contributions relatives).

Dans l'exemple des vins (*cf.* **Tableau 9.5**), ces contributions quantifient le rôle équilibré des quatre groupes dans la construction du premier axe et la prépondérance des deux olfactions dans la construction du second et du troisième.

**Tableau 9.5** Exemple des vins : contribution des groupes de variables à la construction de chacun des trois premiers axes de l'AFM.

| | Contributions absolues | | | Contributions relatives | | |
|---|---|---|---|---|---|---|
| Groupe | F1 | F2 | F3 | F1 | F2 | F3 |
| Olfaction au repos | .78 | .62 | .37 | 22.6 | 45.3 | 60.7 |
| Vision | .85 | .04 | .01 | 24.7 | 2.9 | 2.3 |
| Olfaction après agitation | .92 | .47 | .18 | 26.7 | 34.3 | 29.3 |
| Gustation | .90 | .24 | .05 | 26.0 | 17.4 | 7.7 |
| $\sum$ | 3.46 | 1.37 | .62 | 100 | 100 | 100 |

Cet indicateur peut aussi être calculé pour les groupes supplémentaires, auquel cas il ne s'interprète pas comme une contribution mais sert simplement à situer les groupes supplémentaires par rapport à l'ensemble des groupes actifs.

### c) Qualité de représentation d'un groupe par un axe

La qualité de représentation du groupe $j$ (en tant que point du nuage $N_J$ défini en section 8.4) par l'axe $w_s$ peut être quantifiée à l'aide de l'indicateur classique :

$$\cos^2(j, w_s) = \left[\frac{P_{w_s}(j)}{\|j\|}\right]^2$$

en notant $P_{w_s}(j)$ la projection de $j$ sur $w_s$.

Cet indicateur s'utilise, dans l'étude de la représentation des groupes, de la même façon que pour les représentations usuelles des analyses factorielles : la proximité sur le graphique peut être considérée comme une proximité globale lorsque les points sont bien représentés (*cf.* section 7.1.7). En outre, lorsque l'AFM est utilisée selon le point de vue du modèle INDSCAL, il s'interprète comme une mesure d'adéquation, axe par axe ou pour plusieurs axes, d'un groupe au modèle.

L'interprétation de cet indicateur comme un rapport *inertie projetée* / *inertie totale* suggère de le calculer pour l'ensemble des groupes actifs. La qualité de représentation du nuage $N_J$ ainsi obtenue est utile :

1. pour juger globalement de l'adéquation des données au modèle INDSCAL ;
2. pour comparer la représentation de $N_J$ fournie par l'AFM et celle fournie par d'autres méthodes.

Cet indicateur peut bien sûr être calculé pour un groupe supplémentaire et même pour l'ensemble de ces derniers lorsque les considérer dans leur ensemble a un sens.

Dans l'exemple des vins, la qualité de représentation globale des quatre groupes actifs sur le premier plan de l'AFM vaut .72 ; la qualité maximum de représentation de ces points par un plan, obtenue par la méthode Statis, soit une ACP directe de $N_J$ (*cf.* § 8.4.4), est de .82 ; la perte de 10 %, « prix à payer » pour pouvoir interpréter les axes de projection (les axes optimaux, issus de l'ACP de $N_J$, ne sont pas interprétables ; *cf.* section c), peut être considérée comme raisonnable dans ce cas.

### 9.2.3 Qualité de représentation du nuage des variables d'un groupe

La qualité de représentation d'un nuage de points par un axe s'apprécie usuellement au travers du rapport [*inertie projetée sur l'axe* / *inertie totale*]. C'est bien là une interprétation essentielle du pourcentage d'inertie en analyse factorielle.

On peut appliquer ce critère aux nuages des variables associés à un seul groupe (notés $N_K^j$ au chapitre 8), actif ou supplémentaire. On obtient ainsi, pour chaque groupe $j$, une suite de valeurs qui, comparées aux pourcentages d'inertie de l'analyse séparée du groupe $j$, quantifie ce que l'on perd en qualité de représentation du nuage $N_K^j$ en le projetant sur les axes de l'AFM plutôt que sur ses axes principaux. Appliqué à l'exemple des vins, cet indicateur conduit aux valeurs rassemblées dans le **tableau 9.6**.

**Tableau 9.6** Exemple des vins : qualités de représentation cumulées des nuages des variables de chaque groupe, dans l'AFM et dans les analyses séparées.

| | AFM | | | ACP séparées | | |
|---|---|---|---|---|---|---|
| Groupe | F1 | F2 | F3 | F1 | F2 | F3 |
| 1 : olfaction au repos | 35.1 | 62.9 | 79.6 | 44.8 | 75.2 | 91.5 |
| 2 : vision | 80.8 | 84.6 | 86.0 | 94.5 | 99.5 | 100.0 |
| 3 : olfaction après agitation | 43.5 | 65.5 | 74.0 | 47.0 | 71.8 | 82.3 |
| 4 : gustation | 56.4 | 71.3 | 74.3 | 62.7 | 82.6 | 90.1 |
| 5 : jugement d'ensemble | 57.2 | 80.1 | 81.6 | 92.5 | 100.0 | - |

Le premier plan de l'ACP séparée du groupe 1 exprime 75.2 % de l'inertie de ce groupe. Le premier plan de l'AFM extrait 62.9 % de l'inertie des variables de ce groupe.

La qualité de représentation de chacun des groupes de variables par l'AFM est nécessairement inférieure (égale si tous les groupes ont les mêmes premières composantes principales) à celle des analyses factorielles séparées. Dans l'exemple, cette

diminution de qualité de représentation par un plan varie entre 6.3 % et 14.9 %. Elle quantifie le « prix à payer » en contrepartie de la représentation simultanée de tous les $N_K^j$. Dans l'exemple, on peut considérer que ce « prix à payer » n'est pas exagéré.

Cet indicateur peut aussi être calculé pour les variables qualitatives auquel cas, pour être comparable à celui d'une ACM (vue comme une AFC du Tableau Disjonctif Complet), il doit prendre en compte les modalités au travers de leurs indicatrices (et non de leurs centres de gravité). Le **tableau 9.7** rassemble ces indicateurs pour l'enquête *Ouest-France*.

**Tableau 9.7** Enquête *Ouest-France* : qualités de représentation cumulées des nuages des variables de chaque groupe.

| | AFM | | ACM séparées | |
|---|---|---|---|---|
| | F1 | F2 | F1 | F2 |
| 1 : signalétique | 8 % | 10 % | 9 % | 16 % |
| 2 : rubriques | 12 % | 25 % | 15 % | 28 % |

Ce tableau confirme que, entre les ACM séparées et l'AFM :

1. la qualité de représentation des rubriques est presque inchangée, ce qui était pressenti puisque le premier plan de l'AFM est très proche, à une rotation près, de celui de l'analyse des rubriques ;
2. la qualité de représentation du signalétique est presque inchangée pour le premier axe mais sensiblement diminuée pour le deuxième axe et donc le premier plan ; ceci aussi était pressenti dès lors que le deuxième axe de l'AFM a été considéré comme spécifique des rubriques.

### 9.2.4 Aides relatives aux facteurs partiels

Les exemples présentés (*cf.* Figures 7.7 et 7.8) montrent l'intérêt de relier les résultats d'une AFM à ceux des analyses factorielles séparées des groupes de variables. Pour cela, on introduit les composantes principales des analyses séparées (dits axes - ou facteurs - partiels) dans l'AFM comme des variables réduites pondérées (*cf.* section 8.3.3 page 181).

Pour chaque composante principale partielle et chaque axe de l'AFM, on calcule les mêmes indicateurs que pour les variables initiales, à savoir :

1. le rapport [*inertie de la composante* / *inertie de l'axe*] qui s'interprète, pour les groupes actifs, comme une contribution au sens usuel (*cf.* section 8.3.3) ;
2. la qualité de représentation (ou cosinus carré).

Ces indicateurs, calculés dans l'exemple des vins, sont rassemblés **tableau 9.8**. Ils sont complétés par des indicateurs relatifs aux groupes et à l'ensemble des groupes.

Pour l'ensemble des $S_j$ premières composantes principales du groupe $j$ et pour chaque axe $s$ de l'AFM, on peut effectuer plusieurs calculs.

**Tableau 9.8** Exemple des vins : aides à l'interprétation relatives aux facteurs partiels.

| | | Contribution | | Qualité de représentation | | |
|---|---|---|---|---|---|---|
| | | F1 | F2 | F1 | F2 | {F1,F2} |
| Olfaction au repos | F1 | 0.222 | 0.008 | 0.770 | 0.011 | 0.781 |
| | F2 | 0.002 | 0.441 | 0.009 | 0.891 | 0.899 |
| | {F1,F2} | 0.224 | 0.449 | 0.463 | 0.366 | 0.829 |
| Vision | F1 | 0.246 | 0.024 | 0.852 | 0.033 | 0.884 |
| | F2 | 0.001 | 0.006 | 0.057 | 0.143 | 0.199 |
| | {F1,F2} | 0.247 | 0.029 | 0.811 | 0.038 | 0.850 |
| Olfaction après agitation | F1 | 0.260 | 0.049 | 0.899 | 0.068 | 0.966 |
| | F2 | 0.007 | 0.286 | 0.044 | 0.739 | 0.783 |
| | {F1,F2} | 0.266 | 0.335 | 0.603 | 0.300 | 0.903 |
| Gustation | F1 | 0.259 | 0.004 | 0.898 | 0.005 | 0.903 |
| | F2 | 0.000 | 0.157 | 0.005 | 0.678 | 0.683 |
| | {F1,F2} | 0.26 | 0.161 | 0.683 | 0.167 | 0.850 |
| Ensemble | {F1,F2} | 0.997 | 0.974 | 0.619 | 0.239 | 0.858 |

1. La contribution des $S_j$ composantes à l'axe de rang $s$, somme des contributions de chacune d'elles ; cet indicateur montre dans quelle mesure les axes de l'AFM sont engendrés par les premières composantes principales du groupe $j$ ;
2. La qualité de représentation des $S_j$ composantes, qui rapporte leur inertie projetée à leur inertie totale ; cet indicateur montre dans quelle mesure l'ensemble des $S_j$ composantes principales du groupe $j$ est bien représenté. Ainsi, dans l'AFM sur les données *Ouest-France*, le premier plan représente 86 % des deux premières composantes des rubriques, ce qui est une façon de quantifier la ressemblance entre les premiers plans de ces deux analyses.

Ces deux indicateurs peuvent enfin être calculés pour l'ensemble des composantes principales tous groupes confondus.

1. La contribution montre dans quelle mesure les premiers axes de l'AFM peuvent être reconstitués avec les seules premières composantes principales des différents groupes. Ainsi, dans l'exemple des vins, les 4 premiers axes partiels (i.e. le premier de chaque groupe) contribuent globalement à 99 % de l'inertie du premier axe de l'AFM, ce qui quantifie le rôle majeur de ces axes dans l'analyse (remarque : si l'on conserve toutes les composantes principales de chaque groupe actif, cet indicateur vaut 100).

2. La qualité de représentation montre dans quelle mesure, les premières composantes principales partielles sont, dans leur ensemble, bien représentées par les premiers axes de l'AFM ; ainsi, dans l'exemple des vins, la qualité de représentation de l'ensemble des 2 premiers axes de toutes les analyses séparées par les 2 premiers axes de l'AFM vaut .858, ce qui quantifie globalement l'excellente représentation (pressentie figure 7.7 page 164) de ces axes partiels dans l'AFM. Cet indicateur est particulièrement précieux dans les applications de l'AFM ayant pour objet principal la comparaison de résultats de différentes analyses factorielles. Remarque : si l'on conserve toutes les composantes principales de chaque groupe, les valeurs de cet indicateur sont égales aux pourcentages d'inertie de l'AFM.

### 9.2.5 Décomposition de l'inertie associée à la représentation superposée des nuages partiels

La section 8.2.5 page 175 présente la décomposition de l'inertie de la représentation superposée selon le théorème de Huygens, associée à la partition des $I \times J$ points partiels en $I$ groupes comportant chacun les $J$ points partiels correspondant à un même individu. Cette décomposition est réalisée pour chaque axe. L'inertie intra de cette décomposition mesure la ressemblance entre les nuages partiels, mise en évidence par un axe donné (attention : cette inertie ne s'additionne pas d'un axe à l'autre).

En pratique, on calcule le rapport [*inertie inter* / *inertie totale*]. Appliqué à l'exemple des vins (*cf.* **Tableau 9.9**), cet indicateur quantifie l'étroite proximité, sur l'axe 1, entre les points partiels relatifs à un même vin. L'intérêt de ce critère pour les axes 2 et 3 est faible puisque ces axes ne sont communs respectivement qu'à 3 et 2 groupes : il montre toutefois une proximité entre points partiels bien plus importante pour ces deux axes que pour les quatre suivants.

**Tableau 9.9** Exemple des vins : rapports [inertie inter / inertie totale] relatifs à la représentation superposée.

| Axe | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
|---|---|---|---|---|---|---|---|
| [*inertie inter* / *inertie totale*] | .87 | .58 | .38 | .14 | .17 | .14 | .17 |

Cette inertie intra peut à son tour être décomposée par individu ; ainsi, dans l'exemple des vins, les contributions (en %) des vins 1DAM et 1POY à l'inertie intra pour l'axe 1 valent respectivement 11,4 % et 8,1 %, valeurs qui quantifient la plus grande variabilité des coordonnées des points représentant 1DAM (*cf.* Figure 7.5 page 160).

En pratique, on trie les individus par inertie intra croissante. Les premiers individus présentent les ensembles d'images partielles les plus homogènes du point de vue de

l'axe : ils illustrent bien le caractère commun (aux groupes de variables) du facteur. À l'opposé, les derniers individus présentent les ensembles d'images partielles les plus hétérogènes du point de vue de l'axe. Ainsi, dans l'exemple, ce tri selon le premier axe fait apparaître 1VAU (respectivement 2ING) comme l'individu ayant une des plus faibles (respectivement fortes) inertie intra. On retrouve bien ce phénomène sur la figure 7.5.

La quantification de la variabilité axe par axe des points partiels relatifs à un même individu présente un intérêt en soi. Mais son apport le plus important réside dans le tri qui permet, lorsque les individus sont nombreux, de sélectionner les individus les plus remarquables selon ce critère sans les examiner tous.

La part d'inertie intra de chaque individu peut à son tour être décomposée selon ses points partiels. Ainsi, la part de 1DAM1 vaut 6.5 %, ce qui montre bien le rôle important de l'olfaction au repos dans l'hétérogénéité des perceptions de 1DAM.

En pratique, on sélectionne les individus partiels ayant les plus fortes inerties intra. Cela permet de mettre en évidence des points partiels « non concordants » avec les autres images associées à ces mêmes points.

## 9.3 ANALYSE FACTORIELLE MULTIPLE HIÉRARCHIQUE

Dans de nombreux tableaux *individus*×*variables*, les variables sont structurées selon plusieurs partitions, généralement emboîtées. L'exemple le plus classique est celui des questionnaires d'enquêtes d'opinion déjà évoqué, dont les questions sont souvent structurées en thèmes et en sous-thèmes. Ainsi, dans un questionnaire d'étude de marchés, on regroupera les questions relatives aux opinions d'une part et celles relatives au comportement d'autre part et, au sein de ce second groupe, on distinguera comportement d'achat et comportement de consommation. Un second exemple est fourni par les données « vins » du chapitre 7, en considérant que l'on dispose de ces mêmes données pour plusieurs millésimes : on regroupera les descripteurs sensoriels d'abord par millésimes puis, au sein de chaque millésime, on distinguera, comme nous l'avons fait au chapitre 7, les quatre phases de la dégustation. Dans ces deux exemples, les variables sont structurées selon deux partitions emboîtées ; plus généralement, on peut considérer une structure hiérarchique sur les variables (*cf.* Figure 9.3).

Pour analyser de telles données, en prenant en compte la structure hiérarchique des variables, on peut utiliser l'analyse factorielle multiple hiérarchique (AFMH) développée par Sébastien Lê. Cette extension de l'AFM présente suffisamment de spécificités pour justifier son statut de méthode à part entière. Nous en décrivons les principales ci-après.

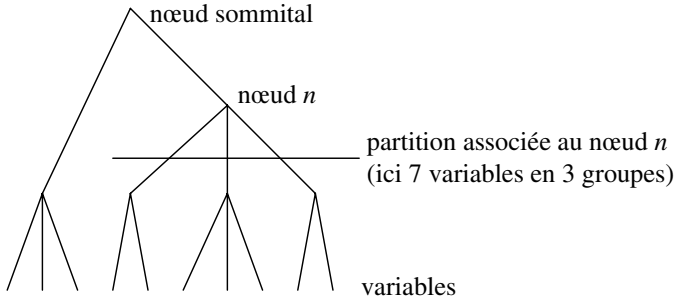



**Figure 9.3** Exemple de hiérarchie structurant les variables d'un tableau.

### 9.3.1 Pondération des variables

De même qu'en AFM, les groupes de variables doivent voir leurs influences respectives équilibrées ; mais cet équilibre doit ici être réalisé pour chaque nœud de la hiérarchie. Ainsi, dans l'exemple des vins esquissé ci-dessus, les deux millésimes doivent être équilibrés entre eux et, au sein de chacun d'eux, les quatre phases de la dégustation doivent être équilibrées entre elles.

Comme en AFM, ces équilibres sont assurés par l'intermédiaire de poids affectés aux variables. Ces poids sont calculés en réalisant les analyses séparées, ACP ou AFM (ou AFMH lorsqu'il y a plus de deux partitions emboîtées) des variables correspondant à chaque nœud de la hiérarchie. Ainsi, dans l'exemple des vins, on réalise d'abord les huit ACP des quatre phases de dégustation pour chacun des deux millésimes, puis une AFM pour chaque millésime. Dans l'analyse finale (AFMH) des deux millésimes, le poids d'une variable est égal à son poids dans l'AFM du millésime auquel elle correspond divisé par la première valeur propre de cette AFM. En procédant ainsi, dans l'analyse correspondant à chaque nœud $n$, les groupes de variables définis par la partition associée au nœud $n$ sont équilibrés.

De façon plus formelle, en considérant les nœuds auxquels appartient successivement la variable $k$ dans l'arbre hiérarchique, et en ordonnant ces nœuds de la base au sommet de la hiérarchie, le poids $P_k^n$ d'une variable $k$ dans l'analyse des variables du nœud $n$ est défini par la formule de récurrence suivante :

$P_k^1$ est fixé *a priori* ; généralement 1

$$P_k^n = P_k^{n-1} / \lambda_1^{n-1}$$

en notant $\lambda_1^n$, la première valeur de l'analyse (des variables) du nœud $n$.

Cette définition des poids de variables correspond à la façon dont ils sont calculés en pratique.

### 9.3.2 Représentation des groupes de variables.

En AFM, dans la représentation des groupes de variables (dite « carré des liaisons »), la coordonnée du groupe $j$ le long de l'axe $s$ s'interprète principalement de deux façons (*cf.* § c page 193) :

– la contribution des variables du groupe $j$ à la construction de l'axe $s$ ;

– la mesure de liaison $\mathrm{L_g}$ entre le groupe $j$ et l'axe $s$.

En AFMH, ces deux notions ne coïncident que pour les groupes de l'analyse d'ensemble (définis par la partition associée au nœud sommital). En pratique, on privilégie la mesure $\mathrm{L}_g$ : on calcule l'inertie projetée des variables du groupe défini par un nœud, en utilisant les poids de ces variables dans l'analyse du nœud immédiatement supérieur.

Le carré des liaisons ainsi obtenu s'interprète comme la projection du nuage $N_J$ des groupes de variables définis par la partition associée à chaque nœud de la hiérarchie et pondérés aux sens de l'AFM (première valeur propre égale à 1). Il bénéficie, entre autres, de la propriété suivante : quelles que soient leurs positions dans l'arbre hiérarchique, deux groupes identiques sont confondus dans le carré des liaisons (ce qui ne serait pas vrai avec l'optique « contribution »).

### 9.3.3 Représentation des nuages partiels

À chaque individu, outre le point moyen, on peut faire correspondre autant de points partiels qu'il y a de nœuds dans la hiérarchie. En AFM, cette représentation est obtenue en projetant le nuage $N_I^j$ de points défini par chacun des $J$ groupes de variables $j$ (nuages dits partiels) sur les axes principaux du nuage moyen $N_I$ (*cf.*§ 8.2.5 page 175). Cette représentation est généralisée en AFMH. Pour chaque nœud $n$, le nuage partiel $N_I^n$ est construit à partir des seules variables regroupées par le nœud $n$. Les $N_I^n$ sont projetés sur les axes principaux de $N_I$.

En AFM, cette représentation bénéficie d'une propriété importante : chaque point moyen $i$ est au barycentre de ses $J$ points partiels $i^j$. Cette propriété est obtenue en dilatant le nuage $N_I^j$ par une homothétie de rapport $J$. En AFMH, cette propriété est étendue de la façon suivante : le point $i^n$ (individu $i$ considéré du point de vue des variables incluses dans le nœud $n$) est au barycentre des individus partiels associés aux groupes de variables rassemblés par le nœud $n$. Ainsi, dans l'exemple des vins, chaque vin (moyen) $i$ est au barycentre de ses deux représentations « annuelles » (i.e. associée à un millésime). À son tour, chaque représentation « annuelle » est au barycentre de ses représentations par groupe de variables sensorielles (*cf.* Figure 9.4).

Cette propriété est obtenue en dilatant le nuage $N_I^n$. En notant $J_n$ le nombre de classes de la partition associée au nœud $n$, le coefficient de la dilatation est égal aux

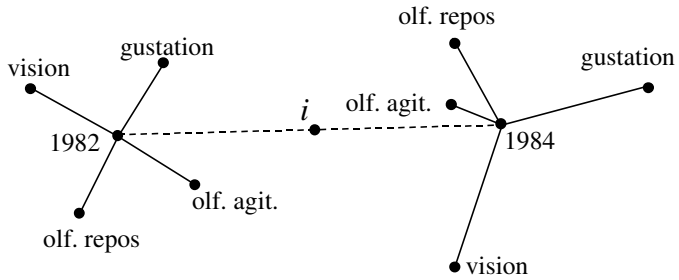



**Figure 9.4** Exemple de représentations de points partiels en AFMH. Pour chaque vin $i$, on distingue son image pour chaque millésime et pour chaque phase de dégustation au sein d'un millésime.

produits des $J_l$ associés aux $L_n$ nœuds englobant le nœud $n$, ce qui peut s'écrire :

$$\prod_{l=1}^{l=L_n} J_l$$

Dans l'exemple des vins, les points partiels relatifs à un millésime sont dilatés avec le coefficient 2 (cas 2 millésimes) ; les points partiels relatifs à une phase de dégustation (dans un millésime) sont dilatés avec le coefficient $2 \times 4 = 8$ (car 2 millésimes$\times$4 phases).

Chapitre **10**

# Comparaison de tableaux de fréquence binaire

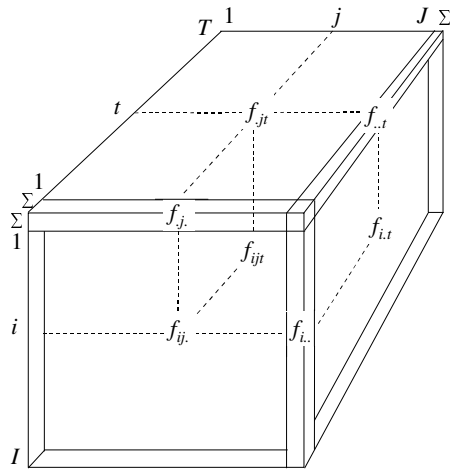
## 10.1 DONNÉES ET PROBLÈMES

L'AFC est extrêmement efficace dans l'analyse des tableaux de fréquence binaire et dans celle de tableaux de type assez semblable. Très naturellement, on souhaite disposer de techniques analogues pour traiter des tableaux non plus binaires mais ternaires, c'est-à-dire des tableaux définis par le croisement de trois variables qualitatives. Dans ce chapitre, nous abordons l'étude de tels tableaux. Plus généralement, nous nous intéressons à l'étude conjointe de plusieurs tableaux binaires définis à partir d'un même couple de variables sur des populations différentes sans que ces tableaux dérivent nécessairement du même tableau ternaire. De telles suites apparaissent fréquemment lorsque des observations identiques sont effectuées à des moments différents. On a alors une suite de tableaux indicés par le temps et l'on souhaite faire intervenir cette dimension temporelle dans l'analyse.

Pour certaines méthodes, il n'est pas nécessaire que l'ensemble des lignes **et** l'ensemble des colonnes soient identiques pour tous les tableaux ; il suffit que l'un de ces deux ensembles soit commun à tous les tableaux. Ceci dit, nous choisissons de rester plutôt dans le cadre et le vocabulaire des tableaux ternaires.

### 10.1.1 Notations

Notons $I, J, T$ les ensembles des modalités des trois variables (la notation $T$ fait référence au temps).




$$\sum_{ijt} f_{ijt} = 1$$
$$f_{ij.} = \sum_t f_{ijt}$$
$$f_{i.t} = \sum_j f_{ijt}$$
$$f_{.jt} = \sum_i f_{ijt}$$
$$f_{i..} = \sum_{jt} f_{ijt}$$
$$f_{.j.} = \sum_{it} f_{ijt}$$
$$f_{..t} = \sum_{ij} f_{ijt}$$

**Figure 10.1** Le parallélépipède des données et ses marges.

Les données peuvent être présentées sous forme d'un parallélépipède (*cf.* **Figure 10.1**) de terme général noté $f_{ijt}$. Les $f_{ijt}$, obtenus en divisant les effectifs par leur total, peuvent être considérés comme une mesure de probabilité sur le produit des trois ensembles $I$, $J$ et $T$.

Les marges binaires de ce parallélépipède sont les trois tableaux de contingence binaire, obtenus en sommant sur l'un des trois indices. Leur terme général est noté respectivement $f_{ij.}$, $f_{.jt}$ et $f_{i.t}$. Chacune peut être représentée par une face du parallélépipède. On parlera aussi des trois marges unaires, vecteurs obtenus en sommant sur deux indices et notés $f_{i..}$, $f_{.j.}$ et $f_{..t}$ : chacune peut être représentée par une arête du parallélépipède. L'arête $f_{i..}$ (resp. $f_{.j.}$ ou $f_{..t}$) est dite souvent « marge sur $I$ (resp. sur $J$ ou $T$) ».

On peut présenter aussi les données comme une suite de tableaux binaires (*cf.* **Figure 10.2**). C'est d'ailleurs ainsi qu'elles se présentent concrètement. L'une des dimensions, $T$ par exemple, joue alors un rôle différent des deux autres. Les $T$ tableaux binaires croisant $I$ et $J$ sont des « tranches » du parallélépipède. Leur somme n'est autre que la marge binaire sur ce même produit.

### 10.1.2 Exemples

Dans ce chapitre, nous appliquons la plupart des méthodes exposées à un tableau de très petite dimension issu de données de l'INSEE (« Bilan formation-emploi 1973 »,

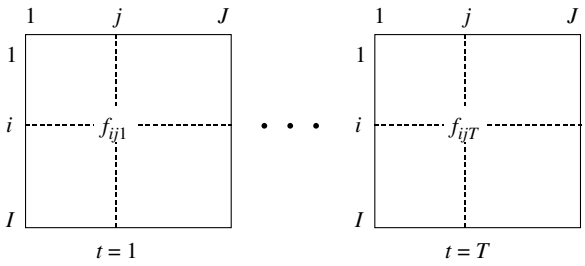



**Figure 10.2** L'ensemble des T tableaux binaires.

**Tableau 10.1** Élèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi : sexe masculin.

| | Niveaux de diplôme | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Emploi occupé | sans diplôme | BEPC | BEP/ CAP | BAC général | BAC technique | DEUG/ ENT | DUT/ BTS | SUP | Total |
| Agriculteur | 15068 | 2701 | 5709 | 297 | 1242 | - | 322 | - | 25339 |
| Ingénieur | - | 337 | 309 | 917 | - | 308 | - | 4383 | 6254 |
| Technicien | 302 | 1697 | 2242 | 1969 | 1399 | 357 | 1943 | 381 | 10290 |
| Ouvrier qualifié | 10143 | 3702 | 30926 | 314 | 1861 | - | - | 337 | 47283 |
| Ouvrier non qualifié | 59394 | 8087 | 17862 | 2887 | 1696 | - | - | 323 | 90249 |
| Cadre supérieur | 596 | 298 | 892 | 1227 | 298 | 2362 | 318 | 6781 | 12772 |
| Cadre moyen | 2142 | 2801 | 672 | 6495 | 924 | 2807 | 2301 | 4030 | 22172 |
| Employé qualifié | 5445 | 7348 | 4719 | 4353 | 1280 | 614 | 982 | - | 24741 |
| Employé non qualifié | 4879 | 4987 | 1514 | 3478 | 886 | 1326 | - | 661 | 17731 |
| Total | 97969 | 31958 | 64845 | 21937 | 9586 | 7774 | 5866 | 16896 | 256831 |

CEREQ, INSEE, SEIS, volume D 59 des Collections de l'INSEE, p. 102 et 103). Il répartit, suivant trois critères, la population des élèves scolarisés en 1972-1973 ayant trouvé un emploi en 1973. Le premier critère est le niveau de diplôme qui comprend 8 modalités ; le second critère est le type d'emploi qui a 9 modalités ; le troisième, le sexe, sépare les hommes et les femmes. Il est clair que les trois dimensions de ce tableau ternaire ne jouent pas le même rôle et qu'il est naturel de présenter des données sous la forme de deux tableaux binaires concernant l'un les hommes et l'autre les femmes (*cf.* **Tableaux 10.1** et **10.2**).

Certains objectifs de l'étude des tableaux ternaires ne s'imposent pas de manière très naturelle sur ces données. Pour les illustrer nous évoquerons deux autres exemples.

Le deuxième exemple croise l'ensemble des cantons de Bretagne, l'ensemble des causes de mortalités et l'âge réparti en 10 classes. Le terme général du parallélépipède est le nombre de décès durant la période 1960-1970 dans le canton $t$, dans la classe d'âge $i$, par la cause de mortalité $j$.

Le troisième et dernier exemple comprend une dimension temporelle. Ce n'est pas un tableau ternaire stricto sensu mais une suite de tableaux contenant, année par

**Tableau 10.2** Élèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi : sexe féminin.

| | Niveaux de diplôme | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Emploi occupé | sans diplôme | BEPC | BEP/ CAP | BAC général | BAC technique | DEUG/ ENT | DUT/ BTS | SUP | Total |
| Agriculteur | 5089 | 1212 | 1166 | - | - | **-** | - | **-** | 7467 |
| Ingénieur | **-** | - | - | 316 | **-** | - | 304 | 1033 | 1653 |
| Technicien | 281 | - | 320 | 320 | 283 | - | 683 | - | 1887 |
| Ouvrier qualifié | 7470 | 1859 | 4017 | 1752 | 657 | **-** | 285 | - | 16040 |
| Ouvrier non qualifié | 29997 | 4334 | 4538 | 1882 | - | **-** | - | - | 40751 |
| Cadre supérieur | - | - | - | 2236 | 595 | 911 | 569 | 6788 | 11099 |
| Cadre moyen | 1577 | 1806 | 4549 | 17063 | 875 | 4152 | 15731 | 3991 | 49744 |
| Employé qualifié | 21616 | 19915 | 32452 | 16137 | 5865 | 1256 | 3332 | 1286 | 101859 |
| Employé non qualifié | 19849 | 7325 | 6484 | 5111 | 898 | 294 | 635 | - | 40596 |
| Total | 85879 | 36451 | 53526 | 44817 | 9173 | 6613 | 21539 | 13098 | 271096 |

année, pour 40 entreprises, le nombre total d'emplois dans chacune des 10 catégories d'emplois qui apparaissent dans ces entreprises.

### 10.1.3 Réduction à des tableaux binaires

Les techniques proposées consistent d'abord à construire des tableaux binaires. À ces tableaux, on applique soit une AFC classique, en utilisant largement la technique des éléments supplémentaires, soit une méthode moins classique spécifique des tableaux ternaires qui dérive de l'AFC.

La réduction des problèmes à l'étude de tableaux binaires est inévitable. On pourrait penser généraliser l'AFC au croisement de trois variables. Mais le concept de « trinité », qui remplacerait celui de dualité, s'est avéré inaccessible : il n'a pas été possible d'obtenir une analyse factorielle de tableaux de contingence ternaire dans laquelle on puisse faire jouer un rôle symétrique aux trois ensembles en traitant toute l'information contenue dans le parallélépipède des données.

Cette limite théorique peut paraître tout à fait regrettable mais elle ne l'est guère car **les problèmes réels ne se posent jamais en termes symétriques suivant les trois variables**. Le plus souvent, un tableau ternaire est considéré comme un ensemble (ou une suite) de tableaux binaires croisant les mêmes variables. Le problème de la **comparaison entre ces tableaux binaires** est presque toujours la préoccupation essentielle.

Dans les trois exemples cités, cette dissymétrie entre les trois variables dans la formulation des objectifs est évidente. On cherche à comparer : dans le premier, les deux tableaux concernant les hommes et les femmes ; dans le deuxième, les tableaux de mortalité (causes$\times$classes d'âge) des différents cantons ; dans le troisième, l'évolution des effectifs de l'ensemble des catégories d'emplois des différentes entreprises.

Cette comparaison elle-même recouvre des objectifs très divers décrits dans la section suivante en référence aux exemples cités.

### 10.1.4 Quelques questions sur la comparaison de tableaux binaires

Dans une suite de tableaux binaires, tous les tableaux ne sont pas absolument identiques mais de grandes tendances peuvent s'y retrouver. Par exemple, la liaison entre *emploi* et *diplôme*, qu'il s'agisse des hommes ou des femmes, doit avoir des points communs. Dans la comparaison entre ces tableaux, un des objectifs peut être la recherche de ces tendances communes qui forment, si elles existent, la **structure commune**. L'objectif complémentaire est l'analyse des **écarts entre ces tableaux** (ou de leur évolution s'il s'agit d'une suite temporelle). Il est bien sûr utile de pouvoir mesurer **l'importance relative des écarts entre tableaux** et **de la structure commune aux tableaux**. Par exemple, peut-on considérer que les répartitions croisées des emplois et des diplômes chez les hommes et chez les femmes sont globalement analogues ou, au contraire, très différentes ?

Dans cette comparaison, on peut s'intéresser plus particulièrement aux **profils des lignes homologues** ou aux **profils des colonnes homologues** : soit les pourcentages des différents niveaux de diplôme chez les hommes et chez les femmes occupant un même type d'emploi ; soit les pourcentages des différents emplois auxquels conduit un même diplôme, pour les hommes et pour les femmes.

On peut aussi comparer les **facteurs de l'AFC** des différents tableaux puisqu'ils en schématisent leurs grandes tendances : lorsqu'il existe une structure commune assez forte, les premiers facteurs se ressemblent.

On peut s'intéresser aussi à un phénomène plus complexe : une liaison conditionnelle. Ce type de problème est posé dans les deux derniers exemples cités. L'étude de la mortalité dans les différents cantons de Bretagne a pour but de mettre en évidence d'éventuelles disparités géographiques des causes de mortalité. Or les causes de mortalité sont très liées à l'âge. Une comparaison brute de ces causes dans tous les cantons ne fait ressortir que la différence entre leur structure d'âge. Il faut étudier la liaison entre deux variables (*canton* et *cause de mortalité*) en neutralisant (en un certain sens) l'influence de la troisième (*classe d'âge*). Une solution couramment utilisée consiste à redresser les pourcentages de mortalité de chaque canton en tenant compte des différences entre les répartitions en classes d'âge. Mais cette technique élimine toutes les informations concernant ces répartitions. Pour conserver la richesse initiale des données, nous posons le problème différemment, en cherchant à mettre en évidence des disparités géographiques valables pour l'ensemble des classes d'âge.

Le problème de l'évolution de l'emploi dans un ensemble d'entreprises est analogue : il faut éliminer l'influence d'une variable. Sachant que la répartition des différentes catégories d'emplois varie beaucoup d'une entreprise à l'autre, comment comparer les évolutions de ces répartitions ?

### 10.1.5 Conclusion

En passant du binaire au ternaire, le niveau de complexité croît considérablement. L'étude d'une liaison ternaire est vaste et il ne peut être question, même pour un tableau de très petite taille, d'en étudier tous les aspects. Aussi, nous n'avons la prétention dans ce chapitre, ni de donner des réponses à toutes les questions posées, ni de faire un bilan exhaustif des traitements. Notre but est d'orienter la réflexion sur ce type de données et de proposer quelques outils que chacun peut adapter à ses problèmes.

Nous évoquons d'abord l'analyse des marges binaires d'un tableau ternaire. Puis nous proposons trois méthodes illustrées par le même exemple. La faible dimension de ces données permet de fournir les résultats complets de chaque analyse. L'intérêt de cette étude systématique est essentiellement pédagogique. Pour préciser ce qu'apporte chacune des techniques proposées, nous mettons l'accent sur les différences entre leurs résultats.

1. La première analyse est une AFC de la somme des tableaux, avec les différents tableaux en éléments supplémentaires.
2. La seconde analyse est une AFC de tableaux juxtaposés complétée par de multiples indices.
3. La troisième analyse, baptisée « analyse intra », permet d'étudier des liaisons conditionnelles.

Nous allons de la plus simple à la plus complexe et il est raisonnable de respecter cet ordre dans les applications. Pour chaque méthode, nous indiquons les grandes lignes des techniques d'interprétation ; puis nous évaluons leur efficacité pour répondre à chacune des questions soulevées concernant la comparaison des tableaux binaires.

## 10.2 ÉTUDE DES MARGES BINAIRES

L'analyse de chacune des trois marges binaires est la première étape indispensable dans l'étude d'un tableau ternaire dès que les dimensions $I, J, T$ sont assez grandes. Ces marges sont des tableaux binaires classiques et l'AFC est tout à fait adaptée à leur étude.

Dans le premier exemple, seule l'analyse de la marge *Emplois*×*Diplômes* présente de l'intérêt (puisque la troisième variable, le sexe, n'a que deux modalités). L'analyse de cette marge permet de dégager les liens entre les emplois occupés et les diplômes possédés, hommes et femmes cumulés, sans tenir compte du sexe.

Dans le troisième exemple où $I$, $J$ et $T$ représentent respectivement les catégories d'emplois, les entreprises et les années, il est utile d'étudier les trois marges. La première marge, qui cumule les années, donne la répartition moyenne (sur la période

étudiée) des catégories d'emplois dans chaque entreprise. Son analyse met en évidence les différences de répartition des emplois suivant les entreprises dans la période considérée et de comparer les entreprises suivant ce critère. La deuxième marge, qui cumule les entreprises, permet d'étudier l'évolution de la répartition des emplois dans l'ensemble du secteur auquel appartiennent les entreprises. La troisième marge croise les années et les entreprises en cumulant les catégories d'emplois ; elle donne l'évolution du nombre total d'emplois dans chacune des entreprises.

Cette méthodologie permet de dégager d'abord les grandes tendances des données, avant de s'attaquer à la description précise de phénomènes plus fins. Les nuances que l'analyse du parallélépipède complet permet de dégager n'ont en effet de sens qu'à l'intérieur de structures plus grossières, mais plus fortes, impliquées par les marges. Cette démarche correspond à la philosophie générale de l'analyse des données.

De plus, on ne le répétera jamais trop, **une étude n'est pas faite par une seule séquence d'analyses**. Chaque résultat remet en question le tableau traité, notamment le codage et les éléments pris en compte. Si des valeurs excentrées, appelées couramment aberrantes, qui apparaissent déjà au niveau de l'analyse des marges ne sont pas éliminées ou recodées avant une analyse fine, les résultats de cette dernière risquent de ne présenter aucun intérêt ou d'être mal interprétés. L'analyse de tableaux binaires se maîtrise bien, les phénomènes perturbateurs se repèrent aisément et on peut les neutraliser beaucoup plus facilement que lors de l'analyse d'un tableau ternaire.

## 10.3 PREMIÈRE ANALYSE : LES TABLEAUX EN SUPPLÉMENTAIRE DANS L'AFC DE LEUR SOMME

### 10.3.1 Principe

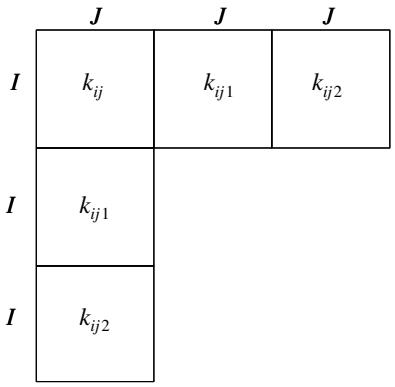
La méthode classique consiste à traiter par l'AFC la somme des $T$ tableaux, en mettant ces $T$ tableaux à la fois en lignes et en colonnes supplémentaires (*cf.* **Figure 10.3**).

Comment cette analyse permet-elle de comparer les $T$ tableaux ? Voyons d'abord la représentation géométrique des colonnes actives et supplémentaires dans l'espace $R^I$ (*cf.* **Figure 10.4**).

Comme le montre la formule ci-après, la colonne $j$ de la marge étudiée, étant la somme des $T$ colonnes homologues $(j,t)$ des $T$ tableaux, son profil $f_{ij.}/f_{.j.}$ est situé au barycentre des $T$ profils $f_{ijt}/f_{.jt}$ (chaque profil étant muni du poids affecté en AFC) puisque :

$$\sum_t \frac{f_{.jt}}{f_{.j.}} \frac{f_{ijt}}{f_{.jt}} = \frac{f_{ij.}}{f_{.j.}}$$

**L'analyse de la somme des *T* tableaux** est donc l'analyse d'un nuage moyen : celui des barycentres des profils des colonnes homologues des $T$ tableaux. Les facteurs




Le tableau actif est de dimension $I \times J$
$I$ = catégories d'emplois
$J$ = niveaux de diplôme
$T$ = sexes
En grisé, les tableaux mis en supplémentaire. Le rectangle vide en bas à droite, qui n'intervient pas dans les calculs, peut contenir des zéros.

**Figure 10.3** Structure des données pour l'AFC du tableau « somme ».

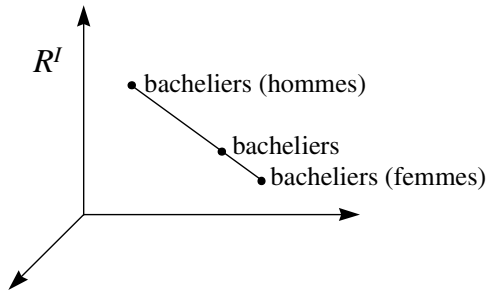
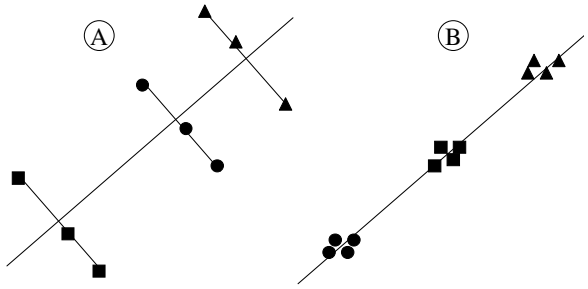



**Figure 10.4** Le profil de la colonne j de la marge étudiée est au barycentre des T profils des colonnes [j,t] des T tableaux. Le profil d'emplois de la classe des bacheliers, hommes et femmes cumulés, est au barycentre des profils d'emplois des bacheliers hommes et des bachelières.

mettront donc en évidence des tendances communes aux $T$ tableaux (si elles existent). Dans l'exemple, les oppositions entre diplômes qui se retrouvent à la fois chez les hommes et chez les femmes apparaissent clairement ; par contre, les différences entre les profils d'emplois des deux sexes, à diplôme égal, sont éliminées.

**Mettre les *T* tableaux en colonnes supplémentaires** dans l'AFC de leur somme consiste à projeter les profils de leurs colonnes sur les axes d'inertie de leurs barycentres. Ceci permet d'étudier, sur chaque facteur, l'écart entre le profil de la colonne $j$ de chaque tableau $t$ et le profil moyen de ces colonnes $j$. Si ces écarts sont tous faibles, le facteur représente une tendance commune à tous les tableaux. Cette projection des profils des colonnes des différents tableaux sur un référentiel commun permet de les comparer, au moins dans ce qui apparaît dans ce référentiel. Mais attention, les différences entre profils homologues ne sont pas forcément visibles sur cette projection, soit parce que les écarts entre ces profils sont orthogonaux aux structures moyennes, soit




Les points d'un même tableau sont représentés par un même symbole. La droite représente la dispersion inter. La dispersion intra est orthogonale à la dispersion inter (A) ou très faible par rapport à la dispersion inter (B).

**Figure 10.5** Exemples dans lesquels la dispersion intra-tableaux n'est pas visible dans l'étude de la dispersion inter-tableaux.

parce que ces écarts sont très faibles par rapport aux écarts entre les profils différents (*cf.* **Figure 10.5**). La situation est exactement identique pour les lignes.

Les deux paragraphes suivants illustrent cette technique par un commentaire détaillé des résultats de l'analyse des données croisant emplois, diplômes et sexes.

### 10.3.2 Interprétation

Le dépouillement des résultats commence par l'étude des éléments actifs, les lignes et les colonnes de la somme des $T$ tableaux ; ce tableau croise 9 catégories d'emplois et 8 niveaux de diplôme.

#### *a) Répartition parabolique sur le plan 1-2 et effet Guttman*

Le graphique des deux premiers facteurs de cette analyse (*cf.* **Figure 10.8**) montre les ensembles de diplômes et d'emplois répartis approximativement sur une courbe de forme parabolique. Ce phénomène, assez courant en AFC, est appelé communément « effet Guttman ». Il apparaît lorsqu'il existe une structure d'ordre à la fois sur l'ensemble des lignes et sur celui des colonnes et que ces structures sont associées. Plus précisément, si l'on réordonne les lignes et les colonnes dans l'ordre du premier facteur, on obtient un tableau dont les éléments proches de la diagonale ont de fortes valeurs tandis que les éléments éloignés sont nuls ou presque nuls. Nous profitons de cet exemple pour présenter quelques résultats généraux concernant cette structure.

➤ Le modèle de l'effet Guttman

Il a été démontré que l'AFC de tableaux modèles, ayant tous leurs éléments nuls en dehors d'une bande diagonale et constants sur cette bande, aboutit au résultat suivant : le deuxième facteur est une fonction polynôme du second degré du premier facteur et, sur le plan 1-2, les points sont situés exactement sur une parabole. De même, le troisième facteur est une fonction du troisième degré du premier et, sur le plan 1-3, les

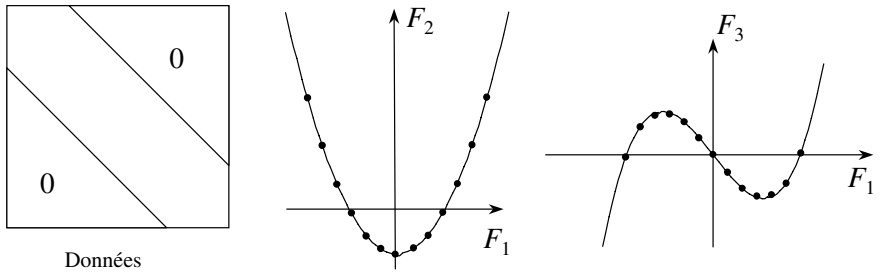



**Figure 10.6** L'effet Guttman : données et premiers plans de l'AFC. Le tableau des données comporte la même valeur sur la bande (en grisé) autour de la diagonale et 0 ailleurs. Ce tableau est appelé « scalogramme ». Sur le plan ($F_1$, $F_s$) de l'AFC d'un tel tableau, les points (lignes et colonnes) sont répartis sur une courbe de degré s.

points sont situés sur une courbe qui coupe trois fois l'axe 1 (*cf.* **Figure 10.6**). Plus généralement, le facteur de rang *s* est un polynôme de degré *s* du premier.

Dans ce cas, la liaison entre les deux variables peut se résumer à la double structure d'ordre mise en évidence par le premier plan. Les facteurs suivants continuent à traduire ce même phénomène. Notons que les pourcentages d'inertie extraite des nuages par les premiers facteurs sont, dans ce cas, faibles alors que l'information sur la structure des données est complète (ce n'est pas le seul exemple qui illustre le fait que, dans le choix du nombre de facteurs à considérer, ces pourcentages n'ont qu'une valeur indicative).

Lorsque l'on observe ainsi une répartition parabolique sur un plan, on interprète globalement le plan qui traduit l'association ordonnée des lignes et des colonnes. Deux questions se posent assez naturellement concernant l'apport du deuxième facteur, puisque la double structure d'ordre est déjà visible sur le premier facteur. La première est relativement théorique : pourquoi observe-t-on une répartition parabolique et pourquoi plus généralement y a-t-il d'autres facteurs que le premier, suffisant pour traduire l'ordre ? La deuxième question est plus pratique : le deuxième facteur peut-il apporter des résultats complémentaires concernant les données étudiées ? Pour répondre à la première question, nous ne donnons pas de démonstration[1], mais remarquons simplement qu'un facteur unique ne peut traduire correctement les distances entre profils dans le cas d'un effet Guttman.

En effet, sur une droite les distances s'ajoutent et, sur l'axe 1, la distance de la première ligne à la dernière est plus importante que sa distance à n'importe quelle ligne intermédiaire. Or, dans le nuage des profils, ceci est faux car les deux lignes extrêmes sont rapprochées par un caractère commun : les zéros qui apparaissent en leur

1. *L'analyse des données*. J.-P. Benzécri et collaborateurs, Dunod, 1973, Tome 2 p.192.

milieu. Le premier facteur donne une bonne représentation de l'ensemble des distances entre tous les couples de points et traduit bien ainsi la structure générale – les distances entre lignes successives sont faibles – mais traduit mal la distance entre les extrêmes. Le deuxième facteur corrige donc et affine l'approximation relativement grossière des distances traduite par le premier facteur. L'intérêt pratique du deuxième facteur est avant tout de caractériser par la forme parabolique une situation type à laquelle on peut se référer pour décrire les données. En outre, il permet de voir si certains points s'écartent de la parabole, ce qui se produit dès que les données s'écartent un tant soit peu du modèle. Nous verrons dans le commentaire de l'exemple, au niveau de la projection des points supplémentaires, l'interprétation de ces écarts.

➤ Reconstitution des données dans un effet Guttman

La reconstitution du tableau de données et son approximation par les premiers facteurs de l'AFC (*cf.* **section** 3.7.4 page 78) se schématisent assez bien dans le cas d'un effet Guttman. C'est pour illustrer cet aspect de l'AFC que nous l'examinons ici. Rappelons que l'approximation d'un tableau $f_{ij}$ par ses $S$ premiers facteurs est la somme du tableau de terme général $f_{i.}\, f_{.j}$ (correspondant à l'hypothèse d'indépendance) et de $S$ tableaux de terme général :

$$\frac{1}{\sqrt{\lambda_s}} f_{i.}\ f_{.j}\ F_s(i)\ G_s(j)$$

Dans le cas d'un effet Guttman, le tableau défini par le premier facteur possède une structure très particulière (*cf.* **Figure 10.7**). Les éléments situés en haut à gauche et en bas à droite sont très fortement positifs tandis que les éléments situés dans les coins opposés sont fortement négatifs ($F_1(i)$) et $G_1(j)$ sont alors de signes opposés). Les autres cases du tableau, qui correspondent aux lignes ou aux colonnes moyennes dont les projections valent presque zéro, ont des valeurs très faibles. Dans cette reconstitution d'ordre 1, le profil des lignes ou des colonnes moyennes est presque proportionnel à la marge du tableau.

Le tableau défini de la même façon par le deuxième facteur a des termes positifs au centre et aux quatre coins, et des termes négatifs ailleurs. Le cumul de ces deux tableaux s'approche de la structure en bande diagonale, caractéristique de l'effet Guttman.

➤ Interprétation du plan des deux premiers facteurs

Le modèle de l'effet Guttman n'apparaît jamais exactement dans des données concrètes, les éléments hors de la bande diagonale n'étant jamais tous nuls, mais une répartition à peu près parabolique traduit un phénomène assez proche. Ici, le **premier facteur** (*cf.* **Figure 10.8**) classe les diplômes et les emplois du plus qualifié au moins qualifié. Or, il y a très peu d'individus non diplômés qui occupent un emploi qualifié ; et réciproquement, peu de diplômés de l'enseignement supérieur occupent

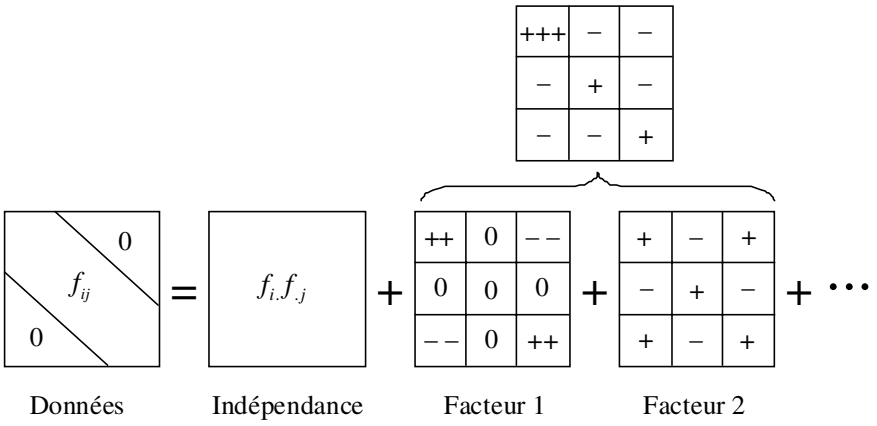



**Figure 10.7** Reconstitution des données, à partir des premiers facteurs de l'AFC, dans le cas d'un effet Guttman.

un emploi d'ouvrier : si le tableau est réordonné suivant le premier facteur, seule la bande diagonale est chargée. Cette structure, qui associe préférentiellement les emplois et les diplômes de même niveau, est celle qui apparaît de la manière la plus évidente dans les résultats, ce qui n'est pas pour nous surprendre !

La parabole est ici asymétrique car les effectifs des emplois très qualifiés (et des diplômes élevés) sont beaucoup plus faibles que ceux des emplois non qualifiés (et des sans-diplôme).

Le **deuxième facteur** s'interprète comme une opposition entre modalités extrêmes et moyennes : les ouvriers non qualifiés ont, comme les ingénieurs, une coordonnée positive alors que les techniciens ont une coordonnée négative. Il met en évidence le point commun aux deux extrêmes d'une même variable : *études supérieures* aussi bien que *sans diplôme* correspondent peu à des emplois moyens et réciproquement les cadres supérieurs comme les ouvriers non qualifiés sont rarement titulaires d'un diplôme moyen. Ce facteur représente, beaucoup mieux que le premier, les diplômes et les emplois moyens, proches de l'origine sur le premier axe.

Le nombre de points actifs est ici trop faible pour juger des écarts de ces points à une direction parabolique parfaite. L'étude de la projection des éléments supplémentaires faite dans la section 10.3.3 montre comment ces écarts peuvent s'interpréter.

### *b) Plan des facteurs 3 et 4*

Contrairement au cas modèle, on constate dans notre exemple, à l'aide des graphiques des plans 1-3 et 1-4, que les facteurs 3 et 4 ne sont pas des fonctions polynômes du premier. On en déduit que la liaison entre diplômes et emplois ne se résume pas au

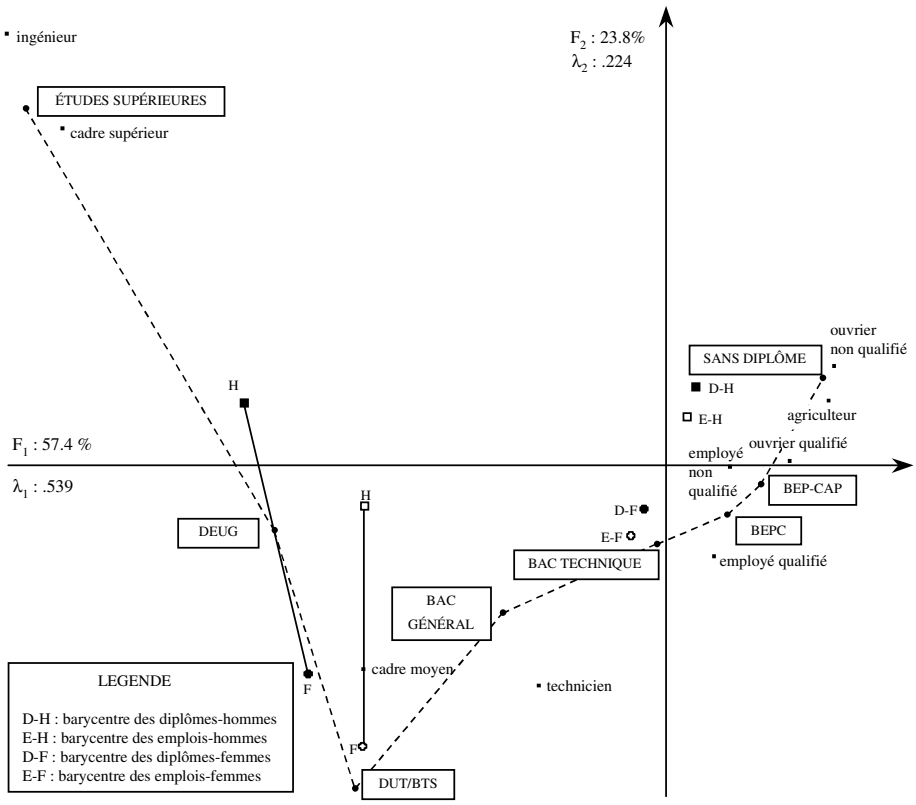



**Figure 10.8** Plan des deux premiers facteurs de l'analyse du tableau somme.

double ordre et que d'autres phénomènes s'y ajoutent. Leur importance est moindre puisque l'inertie de ces facteurs est beaucoup plus faible. Le plan 1-2 donne une image globale de la liaison tandis que les facteurs 3 et 4 montrent des phénomènes beaucoup plus ponctuels, *i.e.* concernant peu d'éléments.

Le **troisième facteur** (*cf.* **Figure 10.9**) traduit la forte association entre les ouvriers qualifiés et le CAP/BEP. Ces deux points, chacun dans leur nuage, ont une contribution à l'inertie du troisième facteur très importante (40 % et 51 % respectivement *cf.* **Tableau 10.3**). Ils déterminent donc en grande partie la direction de l'axe d'inertie ; le fait qu'ils soient situés du même côté signifie qu'ils s'associent « trop ».

Ce troisième facteur différencie entre eux les diplômes (resp. les emplois) de faible qualification très proches sur le premier plan. Il montre une nuance très nette entre les *sans diplôme* et les titulaires d'un diplôme de faible niveau (CAP/BEP) : par rapport à l'ensemble de la population étudiée, les premiers aboutissent beaucoup plus à des

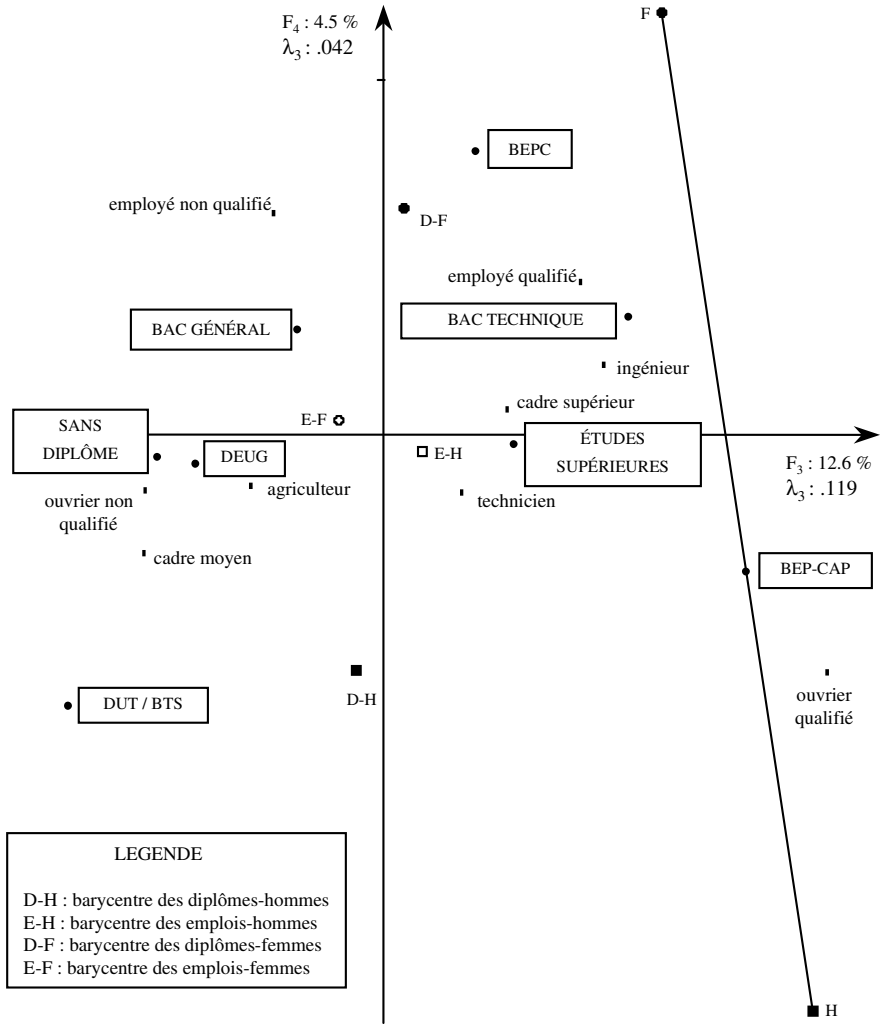



**Figure 10.9** Le plan des facteurs 3 et 4 de l'analyse du tableau somme.

emplois d'ouvrier ou d'employé non qualifié tandis que les seconds mènent beaucoup plus fréquemment à des emplois d'ouvrier qualifié.

Le **quatrième facteur**, schématiquement, oppose les employés (qualifiés et non qualifiés) aux ouvriers qualifiés. Par rapport à l'ensemble de la population étudiée, les premiers sont plus souvent titulaires d'un BEPC et les seconds d'un CAP/BEP.

**Tableau 10.3** Contributions des profils-lignes et des profils colonnes.

| | E. complet | Facteur 1 | Facteur 2 | Facteur 3 | Facteur 4 |
|---|---|---|---|---|---|
| Inertie totale : Brute | .940 | .539 | .224 | .119 | .042 |
| : En % | 1.00 | .574 | .238 | .126 | .045 |
| Agriculteur | 29 | .035 | .020 | .019 | .008 |
| Ingénieur | 120 | .139 | .147 | .013 | .040 |
| Technicien | 27 | .008 | .061 | .003 | .030 |
| Ouvrier qualifié | 89 | .039 | .000 | .403 | .323 |
| Ouvrier non qualifié | 146 | .145 | .126 | .241 | .040 |
| Cadre supérieur | 270 | .353 | .269 | .013 | .002 |
| Cadre moyen | 247 | .265 | .310 | .129 | .093 |
| Employé qualifié | 52 | .006 | .067 | .158 | .273 |
| Employé non qualifié | 21 | .010 | .000 | .022 | .255 |
| | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Sans diplôme | 178 | .183 | .142 | .303 | .018 |
| BEPC | 36 | .010 | .016 | .019 | .518 |
| CAP/BEP | 99 | .044 | .005 | .507 | .182 |
| BAC général | 83 | .071 | .149 | .013 | .074 |
| BAC Technique | 15 | .000 | .012 | .037 | .023 |
| DEUG/ENT | 61 | .090 | .006 | .016 | .001 |
| DUT/BTS/Santé | 150 | .107 | .286 | .088 | .183 |
| Études supérieures | 378 | .494 | .384 | .017 | .000 |
| | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

### 10.3.3 Interprétation des tableaux supplémentaires (hommes et femmes)

Le dépouillement des résultats se poursuit par l'étude des deux tableaux mis en supplémentaire dans l'AFC de leur somme.

#### *a) Profils d'emplois de chaque diplôme, pour les hommes et pour les femmes*

Les deux tableaux ont été mis deux fois en supplémentaire, en colonne et en ligne (*cf.* Figure 10.3 page 230). Etudions d'abord la projection sur le plan 1-2 des **colonnes supplémentaires**, c'est-à-dire des profils d'emplois des hommes et des femmes pour chacun des diplômes. Pour tous les diplômes, sur le premier axe, les deux points représentant les hommes et les femmes sont presque confondus. Sur le deuxième axe, la plupart d'entre eux sont aussi très proches de leur barycentre ; pour des raisons de clarté seuls les DEUG, qui font exception à cette règle, sont représentés sur le graphique de

**Tableau 10.4** Profils d'emplois des hommes et des femmes titulaires d'un DEUG, réordonnés suivant le premier facteur.

| | DEUG | |
|---|---|---|
| | Hommes | Femmes |
| Agriculteur | 0 | 0 |
| Ouvrier non qualifié | 0 | 0 |
| Ouvrier qualifié | 0 | 0 |
| Employé non qualifié | 17 | 4 |
| Employé qualifié | 9 | 19 |
| Technicien | 5 | 0 |
| Cadre moyen | 35 | 63 |
| Cadre supérieur | 30 | 14 |
| Ingénieur | 4 | 0 |
| Total | 100 | 100 |

la **figure 10.8**. L'égalité des coordonnées des deux points *DEUG-Hommes* et *DEUG-Femmes* sur le premier axe implique que, en moyenne, le niveau des emplois des hommes et des femmes qui ont un DEUG est le même (en l'occurrence un niveau moyen). Sur le deuxième axe, la différence est très importante : la coordonnée du point *DEUG-Femmes* est très fortement négative tandis que celle du point *DEUG-Hommes* est très fortement positive. Les propriétés barycentriques indiquent que les femmes occupent plus que les hommes les emplois de coordonnées négatives, donc les emplois moyens. Réciproquement, les hommes occupent plus que les femmes les emplois de coordonnées positives, c'est-à-dire les emplois extrêmes. Cette propriété, décelée sur le graphique se retrouve dans les données initiales (*cf.* **Tableau 10.4**).

### *b) Profils de diplômes de chaque emploi, pour les hommes et pour les femmes*

L'étude des projections des **lignes supplémentaires** permet de comparer les profils de diplômes des hommes et des femmes à emploi égal. Sur le plan 1-2, la situation est tout à fait analogue à celle des profils d'emplois : les coordonnées des couples de points représentant le même emploi sont presque identiques sur le premier axe et, pour la plupart d'entre eux, très proches sur le deuxième axe ; *cadre moyen*, qui fait exception, est représenté sur le graphique. En moyenne, le niveau de diplôme des cadres moyens diffère peu entre hommes et femmes, puisque leur coordonnée sur le premier axe est quasiment la même. Par contre, dans cet emploi, la proportion d'hommes qui possèdent des diplômes extrêmes (*sans-diplôme*, *études supérieures*) est supérieure à celle des femmes qui ont généralement des diplômes moyens (*cf.* **Tableau 10.5**).

**Tableau 10.5** Profils de diplômes des cadres moyens, réordonnés suivant le premier facteur.

| | Cadres moyens | |
|---|---|---|
| | Hommes | Femmes |
| Sans diplôme | 10 | 3 |
| CAP/BEP | 3 | 9 |
| BEPC | 13 | 4 |
| BAC technique | 4 | 2 |
| BAC général | 29 | 34 |
| DUT/BTS/Santé DEUG | 10 | 32 |
| Études supérieures | 13 | 8 |
| Total | 100 | 100 |

### *c) Barycentre des deux tableaux*

Pour faciliter l'interprétation des résultats, on peut ajouter, en supplémentaire, deux lignes et deux colonnes : les sommes des lignes (et des colonnes) de chacun des deux tableaux (hommes et femmes). Ceci fournit les barycentres des profils d'emplois (resp. de diplômes) des hommes et des femmes. Sur les deux premiers facteurs, les quatre barycentres sont très proches de l'origine. Sur le premier facteur, les deux barycentres des femmes ont une coordonnée légèrement positive : les profils d'emplois et de diplômes des femmes sont, en moyenne, très légèrement supérieurs à ceux des hommes (cette situation s'explique-t-elle par le fait que les hommes qui effectuent leur service national ne sont pas pris en compte ?). Sur le deuxième facteur, ces deux barycentres sont nettement négatifs : les profils des femmes sont surtout un peu moins extrêmes.

Sur le quatrième axe, qui oppose ouvriers et employés, les points *hommes* et *femmes* représentant les profils d'emplois d'un même diplôme sont très séparés ; les femmes se dirigent plus vers des emplois d'employés que d'ouvriers tandis que les hommes sont plutôt ouvriers. Cela est particulièrement marqué pour les titulaires de CAP/BEP et seuls les diplômes les plus qualifiés font exception à cette règle. Les hommes et les femmes (tous diplômes cumulés) diffèrent beaucoup de par leur profil d'emplois alors que, tous emplois confondus, ils ont des profils de diplômes analogues : autrement dit, la tendance moyenne des hommes à être plus souvent ouvriers et moins souvent employés que les femmes n'est pas liée à une différence de diplômes.

## 10.3.4 Bilan

Récapitulons dans quelle mesure cette première analyse, l'AFC de la somme des tableaux avec ces tableaux en supplémentaires, répond aux questions posées par la comparaison des tableaux binaires.

**Étude de la structure commune aux deux tableaux : oui.**

S'il existe des tendances communes à tous les tableaux, elles apparaissent dans le nuage moyen. L'AFC de la somme des tableaux permet alors d'analyser cette structure commune. Mais, s'il n'existe pas de structure commune assez forte, la somme peut ne traduire qu'un mélange de tendances diverses. Elle peut aussi être influencée de manière prépondérante par un tableau particulièrement typé. Pour juger du caractère « commun » des facteurs de la somme, on peut examiner la dispersion des profils des lignes ou des colonnes homologues.

Dans l'exemple commenté, le premier facteur est visiblement une structure commune, puisque les profils d'emplois ou de diplômes des hommes et des femmes sont presque confondus. Pour le deuxième facteur et surtout le quatrième, la conclusion doit être plus nuancée.

**Comparaisons inter-tableaux entre profils des lignes ou entre profils des colonnes : un peu**.

Nous avons comparé sur les graphiques la position des profils d'emplois et de diplômes des hommes et des femmes ; mais attention : cette comparaison est incomplète car elle est faite uniquement sur les axes du nuage moyen (*cf.* **Figure 10.5**).

**Analyse et mesure de l'importance relative des différences entre tableaux : non.**

Les différences entre les tableaux ne sont pas particulièrement bien mises en évidence par cette technique. L'importance relative des structures communes et des différences n'est pas mesurée.

**Comparaison entre les facteurs : non.**

Les facteurs de chacun des tableaux n'apparaissent pas du tout dans cette analyse.

**Variante**

Au lieu de prendre, comme base de l'analyse, la somme de tous les tableaux, il est possible de prendre l'un d'entre eux. Par exemple le premier ou le dernier d'une suite temporelle s'il s'impose de manière naturelle comme base de référence. La méthodologie d'interprétation est tout à fait analogue mais ce n'est plus un nuage de barycentres qui est analysé.

## 10.4 DEUXIÈME ANALYSE : AFC DE VARIABLES CROISÉES OU DE TABLEAUX JUXTAPOSÉS

### 10.4.1 Tableau traité et problèmes spécifiques aux tableaux composés de sous-tableaux

Le tableau analysé (*cf.* **Figure 10.10**) est la juxtaposition des deux tableaux « hommes » et « femmes » ; les lignes du tableau sont, comme dans le paragraphe précédent, les

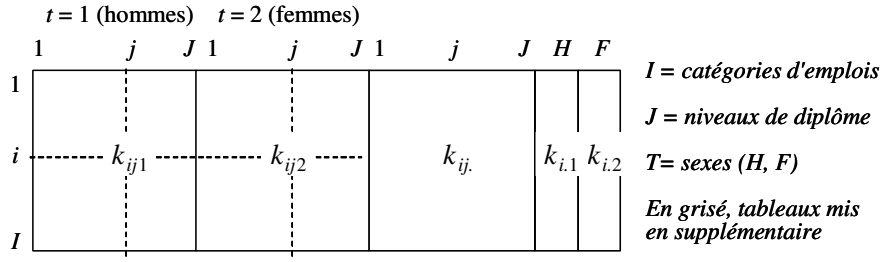



**Figure 10.10** Structure des données dans l'AFC des tableaux juxtaposés. Le tableau actif est de dimensions I et JT.

catégories d'emplois et les colonnes sont les modalités de la variable obtenue en croisant les diplômes et le sexe.

On met en colonnes supplémentaires les marges définies par les deux variables : le tableau à 8 colonnes cumulant les 2 sexes (analysé dans la première méthode) et le tableau à deux colonnes cumulant les diplômes pour chaque sexe. Notons que le rôle des emplois et des diplômes n'est pas symétrique : chaque diplôme apparaît deux fois (à travers les profils d'emplois des hommes et des femmes) tandis que les emplois n'apparaissent qu'une fois. Nous aurions pu juxtaposer les deux tableaux « hommes » et « femmes » en prenant comme dimension commune les diplômes, ce qui aurait inversé la situation respective des emplois et des diplômes. Nous évoquons cette analyse, fondamentalement différente, dans le bilan sur la méthode à la fin de cette section.

L'AFC s'applique sans problème à ce tableau. Cependant, l'une de ces deux variables étant une variable croisée, la nature du tableau est complexe et il est nécessaire de compléter les résultats par des indices d'aide à l'interprétation concernant ce croisement. Nous consacrons l'essentiel de cette section à l'introduction de ces indices, valables d'ailleurs pour tout tableau se décomposant naturellement en sous-tableaux. Leur définition étant fondée sur une décomposition de l'inertie suivant le principe de Huygens, nous précisons d'abord cette décomposition qui permet, de plus, de comparer formellement les résultats de cette deuxième analyse à ceux de la première. Cette première série d'indices est complétée par des indices de comparaison des facteurs de tous les tableaux apparaissant dans le tableau traité. Une dernière section introduit le principe de l'analyse par sous-tableaux qui permet, entre autres, de calculer tous ces indices et plus généralement traite de tableaux de fréquence composés de plusieurs sous-tableaux juxtaposés. Enfin, avant de faire le bilan de cette deuxième analyse, nous comparons les objectifs des analyses de tableaux composés de sous-tableaux suivant qu'il s'agit de tableaux de fréquence ou de tableaux de variables.

### 10.4.2 Comparaison avec l'analyse de la somme et décomposition de l'inertie

Nous comparons ici les deux nuages (celui des colonnes et celui des lignes) étudiés dans cette analyse des tableaux juxtaposés aux nuages construits dans l'analyse de leur somme commentée dans la section 10.3.

*a) Nuage des colonnes*

Si l'on considère l'ensemble des colonnes, actives et supplémentaires, de cette deuxième analyse, on retrouve exactement tous les points du nuage construit dans la première analyse de la section 10.3 ; mais les axes d'inertie sont calculés sur 16 points (les profils d'emplois des femmes et des hommes pour chaque diplôme) et non plus sur les 8 barycentres (profils d'emplois, hommes et femmes cumulés).

Le principe de Huygens indique que l'inertie d'un nuage de points composé de plusieurs sous-nuages peut se décomposer en une somme de l'inertie inter nuages (inertie des barycentres des sous-nuages) et des inerties intra nuages (inertie de chaque sous-nuage autour de son barycentre). La formule ci-dessous résume cette décomposition :

$$\text{Inertie Totale} = \text{Inertie Inter} + \sum \text{Inertie Intra}$$

Dans l'analyse proposée dans cette section, la dispersion intra-diplôme intervient donc dans la détermination des axes. Dans le cas limite (illustré dans la **figure 10.5**) où la dispersion intra, orthogonale à la dispersion inter, est invisible sur les axes d'inertie du nuage des barycentres, les écarts entre les profils homologues peuvent déterminer un axe et ainsi apparaître sur les graphiques de cette analyse.

*b) Nuage des lignes*

Le nuage des emplois n'est pas le même que dans l'analyse de la somme puisque la distance entre deux catégories d'emplois est induite par une répartition en 16 modalités et non plus en 8 modalités : les écarts entre hommes et femmes jouent maintenant un rôle. Plus précisément, on peut montrer que le carré de la distance entre deux emplois $i$ et $l$ (ou entre un emploi $i$ et le barycentre $G$ du nuage) se décompose en une somme de deux termes. Le premier n'est autre que le carré de leur distance dans le tableau somme, actif dans l'analyse de la section 10.3 ; c'est la part inter-diplômes de la distance. Le second terme est aussi le carré d'une distance : il exprime la part intra-diplôme. Pour démontrer cette égalité, il suffit d'écrire formellement les distances entre profils :

$$d^2(i,l) = d^2_{\text{somme}}(i,l) + \sum_j d^2_{\text{intra } j}(i,l)$$

$$d^2(i,G) = d^2_{\text{somme}}(i,G) + \sum_j \mathrm{d}^2_{\text{intra } j}(i,G)$$

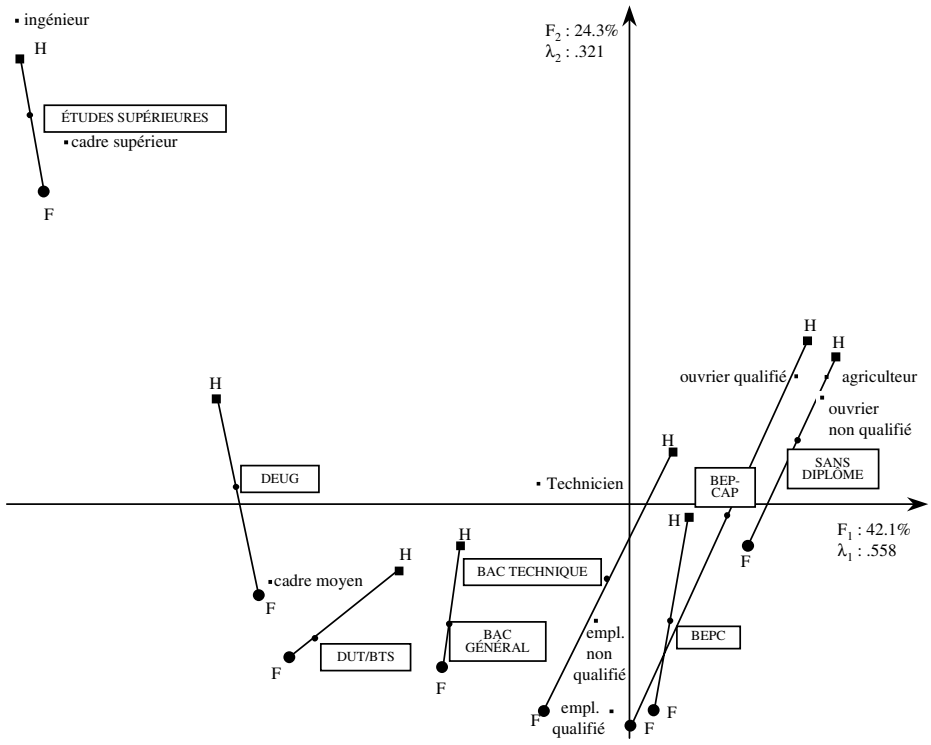



**Figure 10.11** Le plan des deux premiers facteurs de l'AFC des tableaux juxtaposés.

Ainsi, que ce soit dans le nuage des lignes ou celui des colonnes, l'inertie se décompose en deux parts : la part inter-diplômes qui est celle de la première analyse (celle de la somme) et la part intra-diplôme qui s'y ajoute. $d_{\text{intra } j}$ est spécifié en 10.5.4.

Le graphique des deux premiers facteurs (*cf.* **Figure 10.11**) est assez semblable à celui obtenu dans l'analyse de la marge. Ceci n'a rien pour nous étonner car la double structure d'ordre, qui est une structure inter-diplômes, est très forte. Les paragraphes suivants permettent de comparer plus précisément les résultats de ces deux analyses. Notons seulement ici que, si l'inertie de ce premier facteur est légèrement supérieure à celle du premier facteur de la somme, le pourcentage d'inertie extrait est beaucoup plus faible, l'inertie totale étant plus élevée.

### 10.4.3 Indices de contribution à l'inertie de la marge et des sous-tableaux

Lorsque la structure des données définit une partition « naturelle » sur l'ensemble des points d'un nuage, il est intéressant de calculer des indices qui généralisent, aux sous-nuages et au nuage des barycentres, l'indice de contribution à l'inertie défini pour un élément. Précisons bien que la décomposition de l'inertie se fait suivant le principe de Huygens. La contribution à l'inertie d'un sous-nuage n'est pas la somme des contributions à l'inertie de ses éléments mais son inertie intra (rapportée, comme toujours dans le calcul des contributions, à l'inertie totale). La somme des contributions des sous-nuages est la part « intra » de l'inertie et la contribution du nuage des barycentres est la part « inter ». Une discussion analogue, dans le contexte de l'interprétation conjointe d'une ACP et d'une CAH, est développée en 2.5.2.

L'ensemble des colonnes étant, dans notre exemple, formé par les modalités de deux variables croisées, **deux décompositions sont possibles**. Dans la première, celle indiquée dans la section précédente pour comparer cette analyse à celle de la somme, l'ensemble des 16 profils d'emplois est décomposé en 8 sous-ensembles des deux éléments correspondant au même diplôme. Dans la seconde, il se décompose en 2 sous-ensembles de 8 éléments correspondant aux 8 diplômes pour chacun des deux sexes. Ces deux décompositions apportent des résultats complémentaires.

#### *a) Contribution à l'inertie : décomposition en 8 sous-tableaux*

Schématiquement, cette première décomposition permet de répondre à la question suivante : dans le choix d'un emploi, quel est le plus déterminant ? le diplôme obtenu ou bien, à diplôme égal, le sexe ? Commentons le **tableau 10.6** qui donne, dans l'espace et sur chacun des 4 premiers facteurs :

1. le pourcentage d'inertie inter (inertie des 8 profils d'emplois de chacun des diplômes hommes et femmes cumulés rapportée à l'inertie totale) ;
2. les 8 pourcentages d'inertie intra de chacun des diplômes (inertie du nuage des deux profils d'emplois des hommes et des femmes rapportée à l'inertie totale).

**Dans l'espace tout entier**, l'inertie inter-diplômes est de l'ordre de 2/3 de l'inertie totale : le diplôme joue un rôle prépondérant dans l'emploi occupé ; mais la part restante, presque $1/3$ de la totalité, montre bien que le sexe, à diplôme égal, n'est pas sans importance. La différence entre les profils d'emplois des hommes et des femmes à diplôme égal est importante surtout au niveau des CAP/BEP : l'inertie de ce seul sous-nuage représente 12.2 % de l'inertie totale et presque la moitié de l'inertie intra.

Sur le premier facteur $F_1$, l'inertie est presque exclusivement une inertie inter (0.944). Cela montre clairement qu'il s'agit d'un facteur inter-diplômes : le niveau moyen d'un profil d'emplois est déterminé par le niveau de diplôme sans que le sexe intervienne.

**Tableau 10.6** L'inertie totale et sa décomposition inter-diplômes et intra-diplôme dans l'analyse des profils d'emplois.

| | Espace entier | $F_1$ | $F_2$ | $F_3$ | $F_4$ |
|---|---|---|---|---|---|
| Inertie totale | 1.326 | .558 | .321 | .170 | .140 |
| I. inter (en %) | .709 | .944 | .513 | .660 | .811 |
| I. intra (en %) | .291 | .056 | .487 | .340 | .189 |
| Sans diplôme | .057 | .013 | .105 | .050 | .070 |
| BEPC | .018 | .001 | .040 | .045 | .004 |
| CAP/BEP | .122 | .037 | .288 | .184 | .091 |
| BAC général | .013 | .000 | .013 | .000 | .001 |
| Bac technique | .015 | .003 | .021 | .036 | .001 |
| DEUG | .008 | .000 | .010 | .020 | .003 |
| DUT/BTS | .032 | .002 | .002 | .001 | .017 |
| Supérieur | .027 | .000 | .008 | .005 | .001 |

Dans notre exemple où le jeu de données est de dimension très faible, on décèle très rapidement cette structure sur les graphiques. L'intérêt de cet indice est de la quantifier. Dans l'étude de données de dimension plus importante, un tel indice peut apporter un gain de temps précieux : une inertie inter aussi importante montre que l'interprétation de l'axe doit s'appuyer uniquement sur les barycentres.

Le deuxième facteur est mixte : son inertie est pour moitié inter et pour moitié intra. Son interprétation est plus complexe car elle nécessite de prendre en compte les deux dispersions. L'écart entre les hommes et les femmes joue un rôle important sur cet axe, surtout au niveau des CAP/BEP et des sans-diplôme. La parabole traduisant l'effet Guttman est moins régulière que dans l'AFC de la somme. Le troisième facteur est encore mixte tandis que le quatrième est plutôt inter-diplômes.

### *b) Contribution à l'inertie : décomposition en 2 sous-tableaux*

Schématiquement, cette seconde décomposition des mêmes données et des mêmes facteurs permet de répondre à la question suivante : dans le choix d'un emploi, quel est le plus déterminant ? Le sexe ou bien, pour un sexe donné, le diplôme possédé ?

Commentons le **tableau 10.7** qui donne, dans l'espace entier et sur chacun des 4 premiers facteurs :

1. le pourcentage d'inertie inter (inertie des 2 profils des hommes et femmes tous diplômes cumulés rapportée à l'inertie totale) ;
2. les 2 pourcentages d'inertie intra de chacun des sexes (inertie du nuage des 8 profils d'emplois des hommes ou des femmes rapportée à l'inertie totale).

**Tableau 10.7** L'inertie totale et sa décomposition inter-sexes et intra-sexe dans l'analyse des profils d'emplois.

| | Espace entier | $F_1$ | $F_2$ | $F_3$ | $F_4$ |
|---|---|---|---|---|---|
| Inertie totale | 1.326 | .558 | .321 | .170 | .140 |
| I. inter (en %) | .172 | .070 | .506 | .092 | .001 |
| I. intra (en %) | .828 | .930 | .494 | .907 | .999 |
| Hommes | .465 | .544 | .244 | .313 | .824 |
| Femmes | .363 | .386 | .267 | .594 | .175 |

La différence entre les profils d'emplois des hommes et des femmes, tous diplômes cumulés, représentent 17 % de l'inertie du nuage. Elle n'influe que sur le deuxième facteur. Le premier facteur, comme les facteurs 3 et 4, est dû exclusivement à l'écart entre profils d'emplois correspondant à des diplômes différents, tant chez les hommes que chez les femmes. Notons que le facteur 4 montre une dispersion beaucoup plus importante chez les hommes que chez les femmes. Sur les autres facteurs, ces dispersions sont plus équilibrées.

#### *c) Complémentarité des deux décompositions*

Les résultats des deux décompositions de l'inertie, suivant le sexe ou le diplôme, ne sont pas directement liés. Ainsi, la dispersion inter-sexes du **tableau 10.7** est plus faible que la dispersion intra-diplôme du **tableau 10.6** ; la seconde correspond aussi à une dispersion due au sexe, mais à diplôme constant.

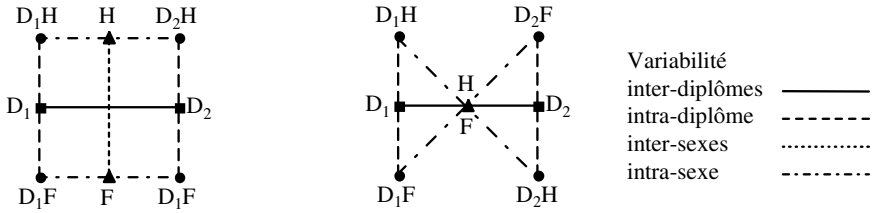
Dans ces données, il y a une interaction entre le sexe et le diplôme sur le profil d'emplois ; en cumulant les diplômes pour un même sexe, des écarts qui ne jouent pas dans le même sens se neutralisent.

La **figure 10.12** illustre ces deux décompositions dans deux cas schématiques comportant chacun deux diplômes. Dans le cas 1, la variabilité intra-diplôme est la même pour chacun des diplômes : elle est donc identique à la variabilité inter-sexes. Dans le cas 2, la variabilité intra-diplôme est très différente d'un diplôme à l'autre : la variabilité inter-sexes est nulle.

On retiendra que les deux décompositions sont **deux regards différents** sur les données qui peuvent avoir chacun leur intérêt ; la décomposition la plus intéressante (ici celle en 8 sous-tableaux) n'est pas forcément celle qui vient la première à l'esprit.

### 10.4.4 Indices de qualité de représentation des différents nuages

La qualité de représentation d'un nuage sur un axe (ou sur un sous-espace) est le rapport entre l'inertie du nuage projeté et l'inertie du nuage dans l'espace. Comme dans la section précédente, nous suivons la décomposition de Huygens pour calculer




Cas 1 : deux décompositions de l'inertie totale (8)
inter-diplômes : 4 inter-sexes : 4
intra-diplômes : 4 intra sexes : 4

Cas 2 : deux décompositions de l'inertie totale (8)
inter-diplômes : 4 inter-sexes : 0
intra-diplômes : 4 intra sexes : 8

**Figure 10.12** Deux exemples très simples de décomposition de l'inertie totale en inerties inter-diplômes, intra-diplômes et inter-sexes. *$D_1H$ : diplôme 1 pour les hommes.*

**Tableau 10.8** Qualité de représentation du nuage des 8 barycentres des profils d'emplois et de deux des 8 sous-nuages.

| | $F_1$ | $F_2$ | $F_3$ | $F_4$ |
|---|---|---|---|---|
| Barycentres | 0.560 | 0.175 | 0.119 | 0.121 |
| CAP/BEP | 0.127 | 0.571 | 0.193 | 0.079 |
| Supérieur | 0.002 | 0.074 | 0.024 | 0.005 |

les qualités de représentation des nuages définis par les sous-tableaux : l'inertie de ces sous-nuages est calculée par rapport à leur barycentre, comme dans la décomposition inter et intra de l'inertie. Les indices concernant le nuage des barycentres et chacun des sous-nuages s'utilisent de la même manière que la qualité de représentation d'un point : ils permettent de repérer le ou les facteurs sur lesquels ces nuages sont bien représentés et, inversement, de repérer le ou les nuages qui caractérisent un facteur.

Les **tableaux 10.8** et **10.9** donnent les qualités de représentation du nuage des barycentres dans les deux décompositions du nuage des 16 profils d'emplois et la qualité de représentation de quelques sous-nuages.

Commentons d'abord le **tableau 10.8**.

C'est sur le premier facteur que le nuage des 8 barycentres est le mieux représenté ; nous avons déjà indiqué que ce facteur est un facteur inter-diplômes.

Par contre, c'est sur le deuxième facteur que le petit sous-nuage de deux points défini par le niveau de diplôme CAP/BEP est le mieux représenté. Si nous voulons préciser la différence entre les profils d'emplois des hommes et des femmes titulaires de ce diplôme, c'est donc sur le deuxième facteur qu'il faut surtout se pencher.

**Tableau 10.9** Qualité de représentation du nuage des deux barycentres hommes et femmes et des deux sous-nuages.

| | $F_1$ | $F_2$ | $F_3$ | $F_4$ |
|---|---|---|---|---|
| Barycentres | .174 | .724 | .069 | 0 |
| Hommes | .500 | .120 | .087 | .190 |
| Femmes | .455 | .181 | .212 | .052 |

Pour les diplômés du Supérieur, pour lesquels d'ailleurs la différence de profil d'emplois des hommes et des femmes est faible (*cf.* Tableau 10.6), la qualité de représentation est mauvaise sur les quatre premiers facteurs.

Etudions maintenant le **tableau 10.9**. Les facteurs sur lesquels les deux sous-nuages *hommes* et *femmes* sont les mieux représentés sont le premier puis le troisième pour les femmes, le premier puis le quatrième pour les hommes. Cela, ajouté au fait que le pourcentage d'inertie du sous-nuage *hommes* sur le quatrième facteur est très important (*cf.* Tableau 10.7), montre que l'interprétation de ce dernier doit être axée essentiellement sur la dispersion des profils d'emplois des hommes.

### 10.4.5 Indices de comparaison des facteurs des différents tableaux

La première décomposition de l'inertie suivant les 8 diplômes (*cf.* section a page 244) a permis d'affirmer que le premier facteur du tableau juxtaposant les deux tableaux *hommes* et *femmes* est un facteur inter-diplômes. Mais est-il confondu avec le premier facteur de l'analyse inter-diplômes, c'est-à-dire celui de l'AFC de la somme de ces deux tableaux ? La ressemblance entre les graphiques de ces deux analyses tend à le montrer mais il est utile de quantifier cette ressemblance par un indice numérique.

La deuxième décomposition de l'inertie, suivant le sexe et non plus le diplôme, montre que le premier facteur est aussi un facteur intra-hommes et intra-femmes. La même question se pose : est-il confondu avec le premier facteur de chacun de ces deux tableaux ?

De même, on peut chercher à comparer entre eux les facteurs des deux tableaux hommes et femmes et plus généralement les facteurs de tous les tableaux binaires étudiés conjointement.

L'expérience de l'ACP montre qu'il est beaucoup plus efficace de comparer un ensemble de variables (ici de facteurs) sur un référentiel commun que de les comparer deux à deux. Nous effectuons cette comparaison des facteurs dans l'AFC des tableaux juxtaposés car les facteurs de cette dernière forment un référentiel commun bien adapté. Notons d'ailleurs que la juxtaposition des deux tableaux suivant les catégories d'emplois permet de comparer les facteurs définis sur cet ensemble mais ne permet pas de comparer les facteurs définis sur l'ensemble des diplômes.

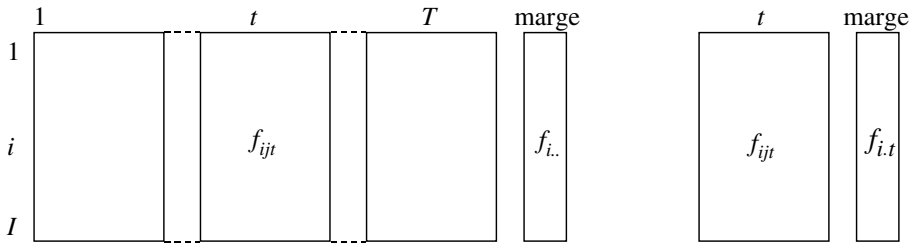



**Figure 10.13** Marges du tableau juxtaposé et du sous-tableau t.

### *a) Problème du poids des lignes et choix d'un référentiel commun*

Les facteurs de tous les tableaux étant des fonctions numériques définies sur le même ensemble $I$, il paraît naturel, pour les comparer deux à deux, de calculer leurs coefficients de corrélation et par conséquent de les comparer tous simultanément à l'aide d'une ACP. Mais le calcul des corrélations fait intervenir le poids des individus (ici les lignes $i$) ; or ces facteurs proviennent d'AFC dans lesquelles les lignes ont des poids imposés. Ces poids sont définis par la marge sur $I$ du tableau analysé ; ils interviennent dans le calcul des facteurs et ces facteurs sont centrés pour ces poids. Or, si la marge sur $I$ du tableau juxtaposant les tableaux hommes et femmes est égale à celle de la somme de ces deux tableaux, les marges des tableaux pris séparément sont différentes (*cf.* **Figure 10.13**). Cette différence de marge et par suite de poids ne fait que traduire une différence plus fondamentale : les populations qui définissent une même ligne $i$ ne sont pas les mêmes (*hommes* et *femmes* par exemple). En toute rigueur, les facteurs de ces tableaux étant définis sur des objets différents, on ne peut définir leur corrélation.

Cependant il est utile de disposer d'indices mesurant la ressemblance entre facteurs qui représentent la projection du même ensemble de modalités.

On peut songer à affecter un poids identique à toutes les lignes $i$. Dans ce cas, les facteurs sont recentrés à l'isobarycentre des points, et les barycentres pondérés, références fondamentales de la situation d'indépendance, ne jouent pas leur rôle ce qui diminue considérablement l'intérêt des résultats.

Il est plus logique d'affecter aux lignes $i$ les poids $f_{i.}$ définis par la population entière. Ceci résout le problème de la comparaison des facteurs du tableau juxtaposé et de ceux des deux tableaux somme $f_{ij.}$ et $f_{i.t}$ puisque ces tableaux ont tous deux pour marge $f_{i..}$.

Il ne reste alors que le problème de la comparaison des facteurs des sous-tableaux. Prenons par exemple le tableau défini en fixant $t$. Sa marge sur $I$ vaut $f_{i.t}$ (*cf.* **Figure 10.13**). Pour comparer ses facteurs sur $I$ notés $F_s^t(i)$ aux facteurs définis sur la population entière où la ligne $i$ a le poids $f_{i..}$, nous allons les « redresser » en les multipliant par le rapport $f_{i.t}/f_{i..}$. Cette transformation s'appuie sur trois arguments :

**a)** Ce « redressement » permet d'obtenir des fonctions centrées pour les poids $f_{i..}$ :

$$\sum_i f_{i..} \left( \frac{f_{i.t}}{f_{i..}} F_s^t(i) \right) = \sum_i f_{i.t} F_s^t(i) = 0$$

**b)** Le facteur redressé apparaît comme une mise en perspective du facteur $F_s^t(i)$, en tant que terme d'écart à l'indépendance, dans le cadre du tableau juxtaposé. En effet, la formule de reconstitution des données appliquée au tableau $t$ de terme général $f_{ijt}/f_{..t}$ fait apparaître le modèle de référence défini par l'indépendance des deux caractères sur la sous-population $t$ :

$$\frac{f_{ijt}}{f_{..t}} = \frac{f_{i.t}}{f_{..t}} \frac{f_{.jt}}{f_{..t}} \left( 1 + \sum_s \frac{1}{\sqrt{\lambda_s}} F_s^t(i) G_s^t(j) \right)$$

Dans cette formule, $F_s^t(i)$ apparaît comme un terme de l'écart au modèle d'indépendance. Une transformation simple de cette formule fait apparaître d'une part le modèle de référence défini par l'indépendance sur la population entière ($f_{i..} f_{.jt}$) et, d'autre part, le facteur redressé :

$$\frac{f_{ijt}}{f_{..t}} = \frac{f_{i..}}{f_{..t}} \frac{f_{.jt}}{f_{..t}} \left( \frac{f_{i.t}}{f_{i..}} + \sum_s \frac{1}{\sqrt{\lambda_s}} \left\{ \frac{f_{i.t}}{f_{i..}} F_s^t(i) \right\} G_s^t(j) \right)$$

**c)** On peut montrer[2] que les facteurs sur $I$ du tableau juxtaposé sont les composantes principales de l'ensemble des variables suivantes :

1. les facteurs redressés des $T$ sous-tableaux ;
2. les facteurs du tableau somme $f_{i.t}$.

Dans cette ACP non normée, les facteurs du sous-tableau $t$ ont un poids égal à $f_{..t}$, ceux du tableau somme un poids égal à 1 et les individus ont un poids égal à $f_{i..}$. L'équivalence entre cette ACP et l'AFC est importante. Outre le fait que les facteurs redressés s'introduisent naturellement dans cette ACP, elle montre que les facteurs sur $I$ du tableau juxtaposé forment le référentiel commun adapté à la comparaison de tous ces facteurs. Elle offre aussi une possibilité de calcul exploitée dans l'analyse par sous-tableaux (*cf.* Section 10.4.6 page 252).

La représentation des facteurs normés des sous-tableaux et du tableau somme sur le cercle des corrélations s'obtient facilement à partir des résultats de l'AFC du tableau juxtaposé par de simples calculs de corrélation.

2. *Cluster Analysis and Data Analysis*. M. Jambu and M.O. Lebeaux, North–Holland, 1983, p.481.

**Tableau 10.10** Corrélations entre les facteurs de l'analyse du tableau juxtaposé et ceux des autres analyses.

| Tableau analysé | | Tableau juxtaposé | | | |
|---|---|---|---|---|---|
| | | $F_1$ | $F_2$ | $F_3$ | $F_4$ |
| Hommes + Femmes | $F_1$ | -.986 | -.158 | .060 | -.012 |
| | $F_2$ | -.088 | .784 | .607 | -.071 |
| | $F_3$ | .008 | -.056 | .182 | .969 |
| | $F_4$ | .136 | -.557 | .745 | -.139 |
| Hommes | $F_1$ | .949 | -.120 | -.178 | -.036 |
| | $F_2$ | .015 | .756 | .385 | .126 |
| | $F_3$ | -.032 | .014 | -.355 | .926 |
| | $F_4$ | .006 | -.284 | .595 | .169 |
| Femmes | $F_1$ | .888 | -.214 | .316 | .063 |
| | $F_2$ | .080 | .616 | .614 | .079 |
| | $F_3$ | -.067 | -.471 | .541 | .534 |
| | $F_4$ | -.142 | -.354 | .329 | -.376 |

### *b) Résultats*

Pour les facteurs, comme pour les autres indices, on peut décomposer le même tableau, soit en 8 tableaux de 2 colonnes, soit en 2 tableaux de 8 colonnes. Dans la première décomposition, seuls les facteurs de la marge présentent un intérêt puisque les sous-tableaux n'ont que deux colonnes et un unique facteur. Inversement, dans la deuxième décomposition, nous nous intéressons aux facteurs des deux tableaux *hommes* et *femmes* et négligeons celui du tableau marge qui ne comprend que 2 colonnes.

Commentons le **tableau 10.10** qui contient les corrélations entre :

1. d'une part les facteurs des tableaux juxtaposés ;
2. d'autre part les facteurs de la somme du tableau *hommes* et du tableau *femmes* ainsi que les facteurs redressés des deux tableaux *hommes* et *femmes*.

On constate la grande ressemblance entre le premier facteur du tableau juxtaposé, celui de la somme ainsi que celui du tableau *hommes* ; le premier facteur du tableau *femmes* est encore assez proche : la double structure d'ordre des emplois et des diplômes est assez forte pour déterminer le premier facteur de tous ces tableaux.

Le deuxième facteur du tableau juxtaposé est un compromis entre plusieurs facteurs de chacun des tableaux et de leur somme.

La représentation de ces facteurs (facteurs redressés des sous-tableaux et facteurs du tableau somme) sur le cercle des corrélations du plan 2-3 (*cf.* **Figure 10.14**) montre les ressemblances entre les facteurs d'ordre 2 des sous-tableaux et du tableau somme.

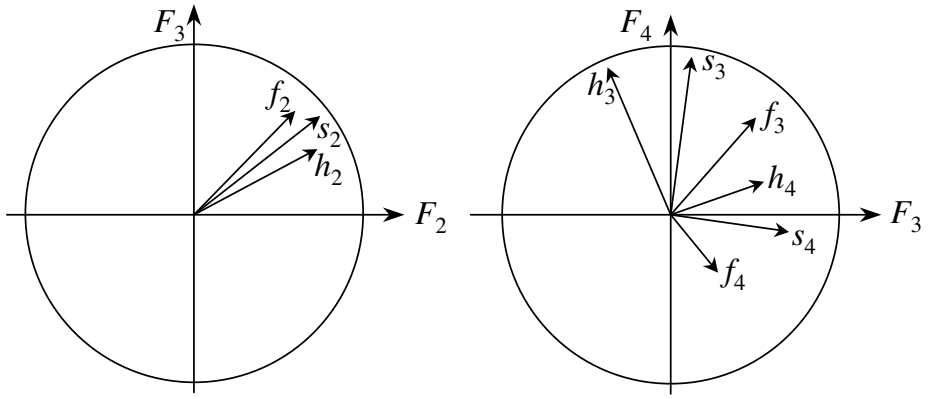



**Figure 10.14** Représentation des facteurs des tableaux somme (s), hommes (h) et femmes (f) sur les plans 2-3 et 3-4 de l'AFC du tableau juxtaposé.

Le cercle des corrélations du plan 3-4 montre que les facteurs d'ordre 3 et d'ordre 4 des tableaux *hommes* et *femmes* ne se correspondent pas.

En conclusion, la dimension principale traduite par le premier plan (double structure d'ordre) est commune aux deux tableaux. Les dimensions suivantes diffèrent.

### 10.4.6 Calcul des différents indices et AFC par sous-tableaux

Les tableaux juxtaposés issus d'un tableau ternaire entrent dans le cadre général de tableaux assimilables à des tableaux de fréquence dont l'ensemble des colonnes est structuré en plusieurs groupes. Les trois séries d'indices introduites plus haut et qui concernent les sous-tableaux ainsi définis et le tableau (dit tableau-marge) constitué par l'ensemble des marges des sous-tableaux (contributions à l'inertie, qualité de représentation des nuages associés et comparaison des facteurs) sont valables dans ce cadre général.

Le calcul de ces indices nécessite un programme spécifique. Il existe, dans la programmathèque ADDAD, un tel programme. Initialement, ce programme a été conçu pour traiter des tableaux trop grands pour les programmes classiques. Ce problème se pose de moins en moins mais cette possibilité demeure.

Le découpage en sous-tableaux est choisi par l'utilisateur. Les AFC du tableau-marge et des sous-tableaux sont effectuées d'abord. Ces analyses sont nécessaires, au moins pour le calcul des corrélations entre leurs facteurs. Les facteurs sur $I$ du tableau entier sont obtenus ensuite par une ACP non normée de l'ensemble des facteurs redressés des sous-tableaux et des facteurs du tableau-marge.

Les derniers facteurs des sous-tableaux et du tableau-marge peuvent être éliminés de cette ACP. Comme ils correspondent aux inerties les plus faibles, ils ne traduisent qu'une information souvent négligeable si ce n'est du bruit. Avec ce principe, on obtient de bonnes approximations des facteurs de tableaux dont le nombre de colonnes (qui détermine la dimension de la matrice à diagonaliser) dépasse les capacités de calcul disponibles. Des indices concernant la qualité de l'approximation complètent le programme. Le nuage de lignes analysé dans l'approximation étant une projection du nuage exact, l'indice général de qualité de l'approximation est le rapport entre les inerties de ces deux nuages.

Les contributions du tableau-marge et des sous-tableaux (inter et intra) sont données directement par la somme des contributions de leurs facteurs. Les qualités de représentation de ces nuages en dérivent immédiatement.

### 10.4.7 AFC de tableaux juxtaposés et AFM

L'AFC par sous-tableaux et l'AFM traitent de tableaux de données dont l'ensemble des colonnes est structuré en groupes. Elles présentent donc une certaine analogie mais, du fait que l'AFM a été conçue pour traiter des tableaux *individus*×*variables*, son aptitude à traiter des tableaux de contingence juxtaposés nécessite d'être discutée.

En AFM, les lignes représentent des individus ou des classes d'individus. Ce dernier point de vue est tout à fait compatible avec les tableaux de contingence dans lesquels une modalité est souvent considérée comme l'ensemble des individus qui la possèdent. En outre, chaque ligne est affectée d'un poids, identique pour tous les sous-tableaux. Ce problème a été rencontré lors de la comparaison des facteurs des AFC séparées des sous-tableaux (*cf.* section a page 249) ; il a pu alors être résolu empiriquement car il ne concernait que des indices d'aide à l'interprétation. Il ne peut être question d'appliquer cette solution à l'AFM dans laquelle les poids des individus interviennent dès le centrage des nuages d'individus. En conclusion, l'AFM ne peut être appliquée à des tableaux de contingence juxtaposés que si les marges de ces tableaux sont identiques d'un groupe de colonnes à l'autre. Cette contrainte peut sembler restreindre fortement le champ des applications : ainsi, on ne peut réaliser l'AFM du tableau de la figure 10.10, les sous-tableaux *hommes* et *femmes* constituant chacun un groupe actif. En fait, cette contrainte met en exergue la difficulté à comparer des tableaux de contingence ayant des marges différentes. Dans ce paragraphe, nous considérons donc des sous-tableaux ayant la même marge sur les lignes.

Dans l'optique de l'AFM sur tableaux de fréquence, les classes d'individus sont décrites par la répartition des leurs individus selon plusieurs variables qualitatives. En référence à l'AFC, la ressemblance entre classes doit être mesurée, au sein de chaque sous-tableau, par la distance du $\chi^2$ appliquée aux profils. Ce point de vue de l'AFC se transpose directement à l'AFM.

En revanche, la **pondération des sous-tableaux**, fondamentale en AFM, nécessite d'être discutée dans le cas des tableaux de contingence. En effet, dans un tableau de fréquence, l'influence du sous-tableau $t$ sur les facteurs de l'AFC globale dépend de deux éléments :

1. sa marge sur $I$, qui détermine la position du barycentre du nuage des colonnes de $t$ et qui intervient dans l'inertie inter ; cet aspect est éliminé par la contrainte imposée plus haut ;
2. la liaison, dans le tableau $t$, entre les deux variables dont les modalités sont les lignes pour l'une et les colonnes pour l'autre : l'intensité de cette liaison détermine la dispersion du nuage des colonnes de $t$ autour de son barycentre.

Ainsi, alors que dans les tableaux de variables ce sont les redondances entre variables qui entraînent un déséquilibre entre les groupes, dans un tableau de contingence, il ne peut y avoir de phénomènes de redondance. Le principe d'équivalence distributionnelle montre bien qu'en aucun cas la ressemblance entre deux colonnes ne perturbe les résultats.

L'opportunité de l'AFM par rapport à l'AFC doit être décidée en référence à cet aspect. Si l'on considère que chaque sous-tableau doit influer dans l'analyse d'autant plus qu'il s'écarte de l'indépendance, l'AFC s'impose : c'est le cas général lorsque chaque sous-tableau correspond à un seul tableau de contingence. Si, en revanche, on souhaite équilibrer l'influence *a priori* des sous-tableaux, on utilisera l'AFM. Nous donnons ci-après quelques exemples pour lesquels l'AFM est adaptée.

1. Chaque sous-tableau est lui-même une juxtaposition de tableaux de contingence, auquel cas des redondances peuvent apparaître. Exemple : on reprend le tableau des vins (*cf.* Figure 7.1 page 150) en conservant les mêmes individus et les mêmes groupes mais en remplaçant chaque colonne $k$ par le tableau de contingence contenant la répartition des 36 juges selon les 5 valeurs de la variable $k$.
2. Comparaison de grilles d'analyse. Exemple : chaque sous-tableau comporte en $k_{ij}$ le nombre de personnes habitant dans la ville $i$ et appartenant à la CSP $j$. Les sous-tableaux diffèrent entre eux par le niveau de détail des CSP.
3. Comparaison des deux tours d'une élection. Chaque sous-tableau concerne l'un des tours et comporte en $k_{ij}$ le nombre de voix obtenues dans le bureau de vote $i$ par le candidat $j$. L'AFM est ici choisie pour l'équilibre entre les deux tours qu'induit la pondération.

➤ Remarques

Les Tableaux Disjonctifs Complets peuvent techniquement être traités soit par l'AFM soit par une analyse par sous-tableaux qui dérive de l'AFC. Comme il s'agit fondamentalement de tableaux de variables, la première solution est préférable. Notons d'ailleurs

que les commentaires concernant les différences entre les marges des sous-tableaux (qui induisent la part inter de l'inertie) ne concernent pas les TDC qui ont une marge constante sur $I$ pour chaque variable.

Au sein d'une AFM, on peut faire intervenir simultanément des groupes de variables de type *fréquence*, *quantitatif* et *qualitatif*. Les groupes de type *fréquence* doivent évidemment avoir la même marge sur $I$ et cette marge impose les poids des individus pour toute l'analyse, même si les groupes de type *fréquence* sont introduits en illustratifs.

### 10.4.8 Bilan sur l'analyse de tableaux juxtaposés

Comme pour l'AFC de la somme de $T$ tableaux (étudiée dans la section 10.3), examinons la qualité des réponses apportées par l'AFC des tableaux juxtaposés aux questions posées par la comparaison de tableaux binaires. Mais d'abord deux remarques sont nécessaires.

➤ Non-symétrie des lignes et des colonnes

Dans tous les aspects de la comparaison, la non-symétrie fondamentale du rôle des lignes et des colonnes dans la juxtaposition des tableaux apparaît. Nous avons juxtaposé les deux tableaux de notre exemple en prenant comme dimension commune les catégories d'emplois. Il est possible de les juxtaposer aussi suivant les niveaux de diplôme (*cf.* **Figure 10.15**). En inversant, dans la juxtaposition, le rôle des emplois et des diplômes, on pose un problème différent. Nous avons comparé les profils d'emplois de chacun des deux sexes, à niveau de diplôme égal ; dans la juxtaposition suivant les diplômes, on compare les profils de diplômes de chacun des deux sexes à catégorie d'emplois égale.

Pour souligner la différence, indiquons seulement que, dans notre exemple, l'inertie des nuages dans le second cas est beaucoup plus faible que dans le premier. Plus précisément, l'inertie intra-emploi est quatre fois plus faible que l'inertie intra-diplôme. Autrement dit, si la répartition des emplois à diplôme égal varie beaucoup d'un sexe à l'autre, celle des diplômes à emploi égal est bien moins différenciée.

➤ Caractère mixte de l'analyse

Cette analyse tient compte à la fois de la dispersion inter-tableaux et de la dispersion intra-tableau. Elle peut aboutir à des facteurs « mixtes » qui traduisent à la fois les deux dispersions (cas du deuxième facteur de notre exemple). Leur interprétation peut être alors assez complexe.

➤ Étude de la structure commune et des écarts à cette structure : un peu

Tout dépend de leur importance relative : s'il existe des tendances communes très fortes par rapport aux écarts, l'AFC des tableaux juxtaposés aboutit à peu près au

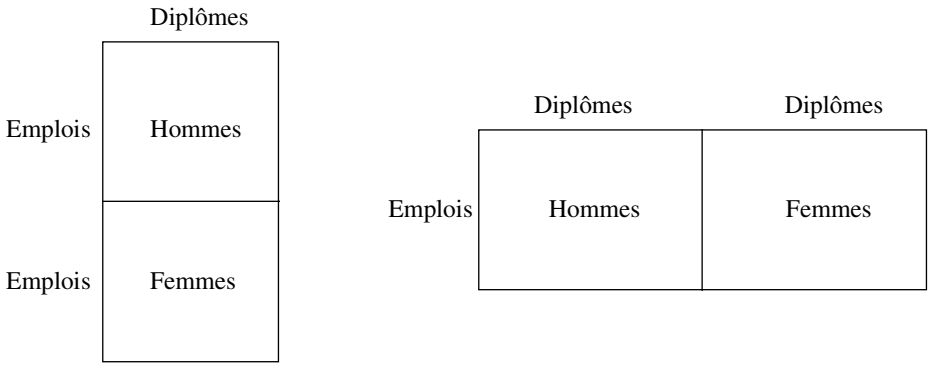



**Figure 10.15** Les deux juxtapositions des tableaux.

même résultat que l'AFC de leur somme et la structure commune est analysée (cas du tableau croisant, par canton, les causes de mortalités $I$ avec les classes d'âge $J$ : la différence entre les causes de mortalité d'une classe d'âge à l'autre est beaucoup plus importante que d'un canton à l'autre pour une même classe d'âge). Si, au contraire, ce sont les écarts qui prédominent, cette AFC représente bien les différences et mal la structure commune. Par son caractère mixte, cette analyse n'est pas la mieux adaptée, ni à l'étude de la structure commune, ni à celle des écarts.

➤ Comparaison des profils des lignes et des colonnes : un peu

Comme dans l'AFC de la somme des tableaux, les profils des colonnes de tous les tableaux sont représentés sur les mêmes graphiques, ce qui permet de les comparer. Les écarts entre les profils des colonnes homologues interviennent maintenant dans la détermination des axes ; ils seront donc *a priori* plus visibles dans cette analyse que dans l'AFC de la somme. Cependant, si ces écarts sont faibles relativement à ceux des différentes colonnes d'un même tableau (comme dans l'exemple des causes de mortalité), ils sont difficiles à détecter. Les profils des lignes des différents tableaux ne sont pas comparés.

➤ Mesure de l'importance relative des différences : colonnes oui, lignes non

Les indices de contribution à l'inertie inter et intra donnent une mesure très satisfaisante de l'importance des écarts des colonnes homologues aux colonnes moyennes. Pour l'ensemble des lignes, il n'y a rien de semblable.

➤ Comparaison des facteurs des analyses séparées : oui

Les facteurs du nuage des colonnes du tableau juxtaposé forment un référentiel commun bien adapté à l'ensemble des facteurs colonnes de tous les tableaux. La projection de ces facteurs sur ce référentiel permet de les comparer efficacement.

## 10.5 TROISIÈME ANALYSE : ANALYSE INTRA

Le terme analyse intra a été introduit par Brigitte Escofier lorsqu'elle a proposé la méthode exposée ci-après. Par la suite, Pierre Cazes a introduit le terme AFC interne pour une méthode plus générale que celle proposée par Brigitte Escofier. Nous conservons le terme *intra* qui peut être commodément précisé (*e.g.* analyse intra-diplômes).

### 10.5.1 Problématique et principes de l'analyse intra

#### *a) Différences entre les profils d'emplois des hommes et des femmes à diplôme égal*

Dans la décomposition de l'inertie du nuage des colonnes du tableau juxtaposé, l'ensemble de ces différences forme l'inertie intra-diplôme. Aucune des deux premières analyses ne permet l'analyse systématique de ces différences. En effet, dans la première (AFC de la somme des tableaux), seule la dispersion inter-diplômes intervient dans le calcul des axes sur lesquels sont projetés les profils d'emplois des diplômés des deux sexes. Dans la deuxième (AFC des tableaux juxtaposés), les dispersions inter et intra interviennent conjointement et l'une peut masquer l'autre. Pour analyser les différences, il faut une analyse dans laquelle seule la dispersion intra intervient.

Géométriquement, la solution est simple. Pour étudier les différences entre les profils d'emplois des hommes et des femmes à diplôme égal, il suffit de considérer le nuage obtenu en recentrant à l'origine tous les sous-nuages de deux points définis par un diplôme (*cf.* **Figure 10.17**). Dans ce nouveau nuage, le point *bachelier-homme*, par exemple, représente la différence entre le profil d'emplois des hommes et celui de tous les bacheliers (hommes et femmes cumulés).

Plus généralement, pour étudier les différences entre les profils des colonnes homologues de $T$ tableaux, nous proposons d'analyser un nuage dérivé du nuage construit dans l'AFC du tableau juxtaposé, en recentrant à l'origine tous les sous-nuages composés des $T$ colonnes homologues. Toute la dispersion inter du nuage initial étant ainsi supprimée, il ne reste que la part intra qui peut être analysée quelle que soit son importance relative. Ce qui résout le problème, par exemple, de la comparaison des causes de mortalité dans les différents cantons à classe d'âge égale.

#### *b) Comparaison des emplois à travers la différence de répartition entre hommes et femmes à diplôme fixé*

On peut aussi chercher à faire une typologie des emplois à travers la différence de recrutement suivant les deux sexes, ceci indépendamment du diplôme possédé. Autrement dit, dans la typologie cherchée, deux emplois sont proches si, pour certains niveaux de diplôme, ils ont tous deux un pourcentage trop (ou pas assez) élevé d'hommes.

Cette question est la duale de la précédente. En effet (*cf.* section b page 242), l'inertie du nuage des lignes (emplois) construit dans l'AFC des tableaux juxtaposés se décompose, comme celle du nuage des colonnes, en une part inter-diplômes et une part intra-diplômes. Pour un emploi donné, caractérisé par les pourcentages des diplômés des deux sexes, le carré de sa distance à un autre emploi est donné par la différence entre leurs profils de diplômes les deux sexes étant cumulés (part inter) et la différence entre les pourcentages d'hommes et de femmes à diplôme fixé (part intra). De même, dans l'exemple du tableau croisant des entreprises ($I$), des catégories d'emplois ($J$) et des années ($T$), les distances entre entreprises induites par la variable croisée $J \times T$ se décomposent en une part inter (induite par $J$, toutes années cumulées) et une part intra (évolution de la répartition des emplois). Pour analyser les évolutions, il faut une analyse dans laquelle ne subsiste que la part intra.

La solution géométrique consiste à construire un nuage dans lequel les distances sont ces distances intra.

#### *c) Principe de l'analyse intra*

L'AFC du tableau juxtaposé ne permet pas d'analyser isolément la dispersion intra. Pour cela, nous généralisons l'AFC et étudions l'écart entre le tableau juxtaposé et un tableau modèle qui n'est pas, comme en AFC classique, le modèle d'indépendance.

Les résultats peuvent être obtenus en utilisant un programme classique d'AFC appliqué à un tableau transformé. On obtient les projections, sur leurs axes d'inertie, d'un nuage de colonnes et d'un nuage de lignes dans lesquelles ne subsiste que l'inertie intra-diplôme ; ces projections sont liées par des formules de transition. Ce principe est suffisant pour comprendre les commentaires de l'exemple de la section 10.5.5. Les sections intermédiaires permettent, au lecteur qui le souhaite, de trouver des précisions techniques sur cette méthode ainsi qu'une ouverture sur une généralisation de l'AFC.

### 10.5.2 Généralisation de l'AFC

L'AFC classique analyse l'écart entre un tableau de fréquence et un tableau modèle correspondant à l'hypothèse d'indépendance. Dans ce modèle, qui n'est autre que le produit des marges, les colonnes (resp. les lignes) ont toutes le même profil. Ce profil modèle est confondu avec le barycentre du nuage. Dans les deux nuages de profils, la référence au modèle d'indépendance se traduit par le centrage du nuage : chaque ligne (resp. colonne) est représentée par la différence entre son profil et le profil moyen.

L'AFC se généralise à un modèle différent du modèle d'indépendance. On suppose que les deux marges du tableau modèle sont égales à celles du tableau étudié (le modèle de l'analyse intra que nous introduisons dans la section suivante vérifie cette condition). Dans cette généralisation de l'AFC, on analyse deux nuages (de lignes d'une part et de colonnes d'autre part) reliés par les relations de dualité. Les points

de ces nuages ont pour coordonnées les différences entre les profils des lignes (resp. colonnes) du tableau de données et du tableau modèle. Les métriques et les poids sont identiques à ceux de l'AFC.

Techniquement, il est possible d'obtenir les résultats de la généralisation de l'AFC en appliquant un programme classique d'AFC aux données préalablement transformées : comme le programme d'AFC se réfère au modèle *produit des marges*, il faut « introduire » le nouveau modèle **et** « supprimer » le modèle ancien. Le tableau traité s'écrit alors :

$$\text{données} - \text{modèle} + \text{produit des marges}$$

Notons $f_{ij}$ le terme général du tableau de données et $m_{ij}$ celui du modèle (avec $f_{i.} = m_{i.}$ et $f_{.j} = m_{.j}$ ; le tableau analysé a pour terme général :

$$f_{ij} - m_{ij} + f_{i.}f_{.j}$$

Lorsque le modèle n'est autre que le produit des marges, on obtient l'AFC classique. Ce tableau peut comporter des termes négatifs mais ses deux marges, étant égales aux marges communes des données et du modèle, sont positives et les programmes d'AFC peuvent s'appliquer. Une ligne $i$ (resp. une colonne $j$), dans le **nuage centré** défini dans l'AFC de ce tableau (obtenu en prenant comme origine le barycentre $f_{.j}$ - resp. $f_{i.}$ -), représente bien la différence entre son profil dans les données et dans le modèle puisque son terme général vaut :

$$\frac{f_{ij}}{f_{i.}} - \frac{m_{ij}}{m_{i.}} + f_{.j} - f_{.j}$$

Les marges du tableau analysé étant égales à celles du modèle, les poids et les métriques sont identiques à ceux de l'AFC du tableau $f_{ij}$.

Contrairement à l'AFC classique, la formule de transition comprend des termes négatifs.

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \left( \frac{f_{ij} - m_{ij}}{f_{i.}} \right) G_s(j)$$

Dans la représentation superposée des lignes et des colonnes, une ligne $i$ est du côté des colonnes auxquelles elle s'associe plus dans les données que dans le modèle et à l'opposé de celles auxquelles elle s'associe moins que dans le modèle. En effet, dans le premier cas, le coefficient de la formule de transition est positif tandis que dans le second cas il est négatif. Le même raisonnement vaut pour les colonnes.

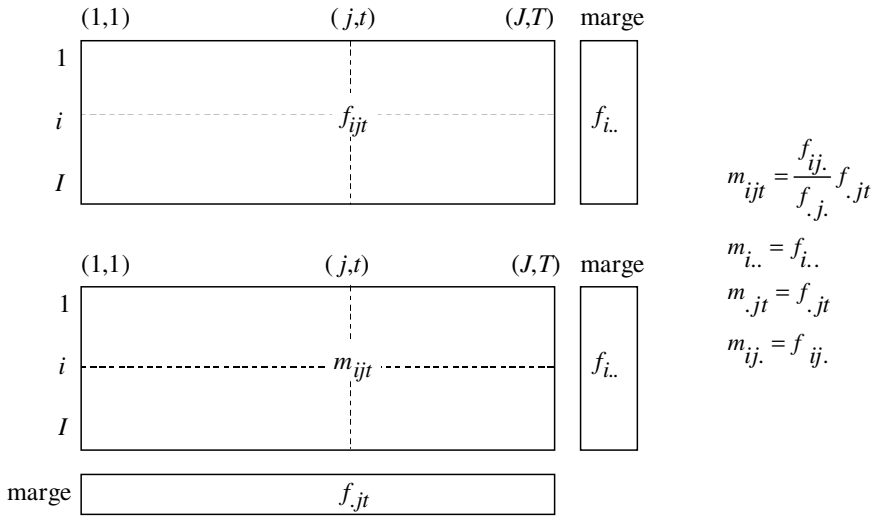



**Figure 10.16** Le tableau juxtaposé et son modèle dans l'analyse intra.

Il est possible de généraliser l'AFC à un modèle dont les marges sont différentes de celles du tableau des données. Dans ce cas, un programme spécial est nécessaire et l'interprétation est plus délicate. Nous n'exposerons pas la technique générale[3].

### 10.5.3 Modèle de l'analyse intra

Le tableau modèle, de mêmes dimensions $I$ et $JT$ que le tableau juxtaposé, est noté $m_{ijt}$. Il est construit pour que son écart avec les données traduise exactement la dispersion intra : il doit donc exprimer la dispersion inter.

Dans le nuage des colonnes du modèle, les profils des $T$ colonnes homologues sont confondus avec le profil moyen $f_{ij.}/f_{.j.}$(pour un diplôme donné, les profils d'emplois des hommes et des femmes sont confondus avec le profil d'emplois de la population entière). Pour que les caractéristiques générales de l'analyse (métriques et poids) soient conservées, il faut que les marges $m_{i..}$ et $m_{.jt}$ soient égales à celles du tableau $f_{ijt}$. Ces contraintes déterminent entièrement le modèle : ainsi, dans le modèle, la colonne $(j,t)$ est obtenue en multipliant le profil moyen $f_{ij.}/f_{.j.}$ par $f_{.jt}$.

$$m_{ijt} = \frac{f_{ij.}}{f_{.j.}} f_{.jt}$$

3. Analyse factorielle en référence à un modèle, B. Escofier, *Revue de Statistique Appliquée*, 1984, vol.XXXII, numéro 4 p. 25.

Dans l'exemple étudié, la colonne *Bac-hommes* du modèle est obtenue en multipliant le profil d'emplois moyen des bacheliers hommes et femmes par l'effectif total des bacheliers-hommes.

Il est facile de vérifier les égalités entre les marges des données et du modèle :

$$m_{ij.} = f_{ij.} \text{ et } m_{.jt} = f_{.jt}$$

ainsi que l'égalité des profils des colonnes homologues :

$$\frac{m_{ijt}}{m_{.jt}} = \frac{f_{ij.}}{f_{.j.}} = \frac{m_{ij.}}{m_{.j.}}$$

Géométriquement, dans l'espace $R^I$, quand on passe du nuage associé au tableau $f_{ijt}$ au nuage du tableau modèle, on ne fait que déplacer les profils des colonnes homologues à leur barycentre sans modifier ni la métrique ni les poids. Il ne reste donc que la part inter de la dispersion sur le nuage des colonnes.

Appliquons au tableau modèle le principe d'équivalence distributionnelle (selon lequel on ne modifie pas les distances entre les lignes d'un tableau lorsque l'on regroupe des colonnes proportionnelles ; *cf.* section 3.4 page 68). Dans ce tableau, les colonnes indicées par le même $j$ sont proportionnelles entre elles, puisque toutes proportionnelles au profil moyen $f_{ij.}/f_{.j.}$. On ne modifie donc pas les distances entre lignes du tableau modèle en regroupant les colonnes indicées par le même $j$. Or ce regroupement conduit au tableau somme de terme général $f_{ij.}$ (=$m_{ij.}$) dont le profil de la colonne $j$ est au barycentre des colonnes $\{(j,t); t = 1, T\}$ du tableau juxtaposé. Ainsi la distance entre lignes induite par le tableau modèle coïncide avec la part inter-$J$ de celle induite par le tableau juxtaposé.

Sous forme probabiliste, ce modèle exprime l'indépendance entre $I$ et $T$ pour la sous-population définie par $j$ :

$$\frac{m_{ijt}}{m_{.j.}} = \frac{m_{ij.}}{m_{.j.}} \frac{m_{.jt}}{m_{.j.}}$$

Cette relation, étant vraie pour tout $j$, $i$, et $t$, définit l'indépendance conditionnelle de $I$ et $T$ par rapport à $J$.

### 10.5.4 Interprétation des formules de l'analyse intra

Dans l'analyse intra, le tableau étudié a pour terme général :

$$r_{ijt} = f_{ijt} - \frac{f_{ij.} f_{.jt}}{f_{.j.}} + f_{i..} f_{.jt}$$

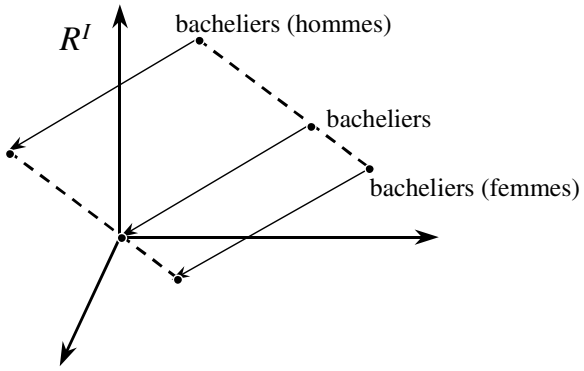



**Figure 10.17** Dans l'analyse intra, les sous–nuages de colonnes homologues sont translatés pour faire coïncider leur barycentre avec l'origine.

➤ Profils des colonnes

La $i^{\text{e}}$ coordonnée du profil de la colonne $(j,t)$ dans l'espace $R^I$, du fait de l'égalité des marges $r_{.jt}$ et $f_{.jt}$, vaut :

$$\frac{r_{ijt}}{r_{.jt}} = \frac{f_{ijt}}{f_{.jt}} - \frac{f_{ij.}}{f_{.j.}} + f_{i..}$$

En prenant comme origine le barycentre du nuage, cette coordonnée s'écrit :

$$\frac{r_{ijt}}{r_{.jt}} - f_{i..} = \frac{f_{ijt}}{f_{.jt}} - \frac{f_{ij.}}{f_{.j.}}$$

Le nuage des colonnes de l'AFC de $r_{ijt}$ se déduit donc de celui considéré dans l'AFC de $f_{ijt}$ en translatant chaque sous-nuage de colonnes $\{(j,t), t = 1, T\}$ défini par un même $j$ pour faire coïncider son barycentre avec l'origine (*cf.* **Figure 10.17**).

➤ Profils des lignes

Les profils des lignes sont, comme ceux des colonnes, obtenus par différence entre les profils du tableau étudié et ceux du modèle. Ce qui donne, pour l'analyse intra, en prenant comme origine le barycentre :

$$\frac{r_{ijt}}{r_{i..}} - f_{.jt} = \frac{f_{ijt}}{f_{i..}} - \frac{f_{ij.}f_{.jt}}{f_{i..}f_{.j.}} = \frac{1}{f_{i..}}\left(f_{ijt} - \frac{f_{ij.}f_{.jt}}{f_{.j.}}\right)$$

On peut vérifier que, dans le carré de la distance entre deux lignes, la part déterminée par les variations inter est supprimée :

$$d^2(i,l) = \sum_{jt}\left(\left(\frac{f_{ijt}}{f_{i..}} - \frac{f_{ljt}}{f_{l..}}\right) - \frac{f_{.jt}}{f_{.j.}}\left(\frac{f_{ij.}}{f_{i..}} - \frac{f_{lj.}}{f_{l..}}\right)\right)^2 \frac{1}{f_{.jt}}$$

$$d^2(i,l) = \sum_{jt} \left( \frac{f_{ijt}}{f_{i..}} - \frac{f_{ljt}}{f_{l..}} \right)^2 \frac{1}{f_{.jt}} - \sum_{j} \left( \frac{f_{ij.}}{f_{i..}} - \frac{f_{lj.}}{f_{l..}} \right)^2 \frac{1}{f_{.j.}}$$

Dans la seconde écriture, le premier terme correspond à la distance (entre les profils $i$ et $l$) dans l'AFC du tableau juxtaposé. Le second correspond à la distance dans l'AFC du tableau *somme*. Confrontée aux équations de la section b page 242, cette équation exprime à nouveau le rôle exclusif des différences intra-diplôme dans le calcul des distances dans cette variante de l'AFC. Autrement dit, la distance entre deux emplois ne dépend pas de la répartition des diplômes, mais seulement des différences entre les pourcentages d'hommes et de femmes pour chaque niveau de diplôme.

➤ Formules de transition

Dans notre exemple, le modèle traduit l'hypothèse suivante : pour chaque diplôme $j$, il y a indépendance entre l'emploi et le sexe. Les emplois loin de l'origine dans l'AFC de $r_{ijt}$ sont donc ceux qui, pour certains diplômes au moins, n'attirent pas de la même façon les deux sexes. Par exemple, sur un axe, un emploi est situé du même côté qu'un diplôme-homme si, parmi les titulaires de ce diplôme, cet emploi attire « trop » les hommes.

### 10.5.5 Commentaires sur le plan issu de l'analyse intra-diplômes

Le plan des deux premiers facteurs (*cf.* **Figure 10.18**) est très différent du plan obtenu dans l'AFC de la somme ou dans celle du tableau juxtaposé. La double structure d'ordre qui déterminait essentiellement ces plans est une liaison inter-diplômes éliminée dans l'analyse intra-diplômes.

➤ Inertie

L'inertie totale du nuage des lignes et du nuage des colonnes est l'inertie intra-diplôme. Cette inertie est assez importante : nous avons vu (*cf.* **Tableau 10.6**) qu'elle représente presque la moitié de l'inertie inter-diplômes.

➤ Premier facteur

Le premier facteur extrait 61 % de l'inertie. Le diplôme qui contribue le plus à ce facteur est le CAP/BEP dont la contribution à l'inertie est de 29 % pour le CAP/BEP-Hommes et de 35 % pour le CAP/BEP-Femmes (donc 64 % en tout). Les deux points représentant le même diplôme sont opposés sur le graphique, puisque le barycentre de chaque sous-nuage (ici de deux points) est situé à l'origine. La prépondérance du CAP/BEP ne nous étonne pas puisque nous avons vu (*cf.* **Tableau 10.6**) que le sous-nuage qu'il définit a une inertie intra très supérieure aux sous-nuages définis par les autres diplômes : c'est le diplôme pour lequel la répartition des emplois est la plus différente entre les hommes et les femmes. Cette différence est entièrement expliquée par le premier facteur puisque la qualité de représentation de chacun des deux points

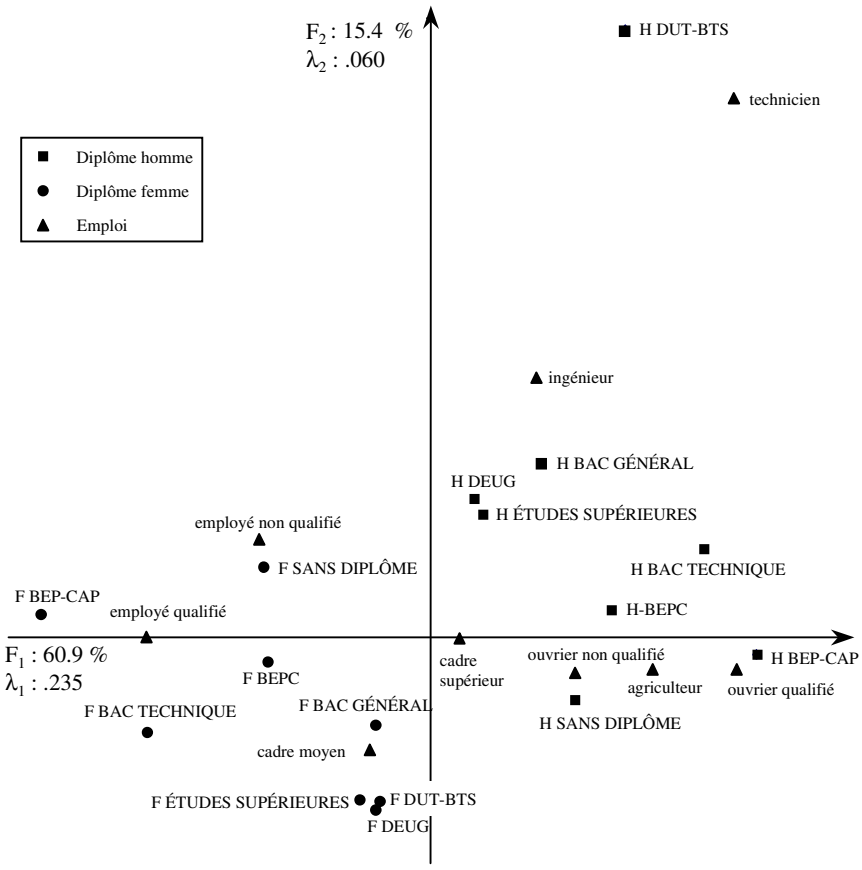



**Figure 10.18** Plan des deux premiers facteurs de l'analyse intra-diplôme.

vaut 0.92 (elles sont forcément identiques puisque les deux points sont alignés avec O). Ce facteur explique bien aussi la différence entre les profils d'emplois du BEPC et des Bacs techniques qui sont aussi des diplômes de niveau moyen. On peut remarquer que tous les diplômes-hommes sont situés à droite tandis que tous les diplômes-femmes sont à gauche. Le premier facteur traduit donc une tendance générale de l'écart entre les profils d'emplois des hommes et des femmes, valable pour chaque diplôme et particulièrement marquée pour les CAP/BEP et les Bacs Techniques.

La projection des emplois permet d'expliquer cette différence. Du côté des femmes, on trouve les employés qualifiés (dont la contribution à l'axe est de 0.44) et du côté des hommes, les techniciens, les ouvriers qualifiés et d'une façon générale les emplois techniques : à diplôme égal, les hommes occupent plutôt des emplois techniques. Cette différence entre les profils d'emplois des deux sexes apparaît en partie au niveau du

quatrième facteur de l'analyse de la somme (analyse inter) et au niveau du deuxième facteur de l'analyse du tableau juxtaposé. Elle est prépondérante dans cette analyse intra qui ne tient compte que de ces écarts.

➤ Deuxième facteur

Le deuxième facteur représente 15 % de l'inertie. Il est déterminé essentiellement par l'emploi de technicien qui attire les hommes et repousse les femmes titulaires d'un DUT/BTS/Santé beaucoup plus qu'il ne le ferait dans l'hypothèse d'indépendance conditionnelle. Contrairement au premier facteur, les diplômes des hommes et des femmes ne sont pas ici systématiquement opposés.

### 10.5.6 Bilan

➤ Facilité de l'interprétation

On pourrait craindre que l'interprétation de l'analyse intra soit complexe car elle utilise des notions de conditionnement qui aboutissent à des concepts difficiles. Or, l'expérience montre que l'interprétation de ses résultats ne pose aucun problème particulier à un utilisateur ayant une pratique de l'AFC. Elle s'appuie, comme elle, sur les formules de transition et sur les indices classiques d'aide à l'interprétation : la qualité de représentation et la contribution à l'inertie des lignes et des colonnes. La différence essentielle avec l'AFC classique, la situation modèle à laquelle on se réfère, s'intègre sans difficulté.

➤ Analyse cumulée des sous-tableaux

On peut aussi voir l'analyse intra comme une analyse cumulée de tous les sous-tableaux. Prenons l'exemple des causes de mortalité dans les différents cantons suivant chaque classe d'âge. Il est assez naturel d'étudier séparément chacune des $T$ classes d'âge qui définissent $T$ nuages de cantons et $T$ nuages de causes de mortalité. Dans l'analyse de chacun des sous-tableaux, les nuages des cantons sont centrés alors que, dans l'analyse du tableau juxtaposé [*décès*×(*cantons*×*classes d'âge*)], les sous-nuages définis par une classe d'âge ne le sont pas. Dans l'analyse intra, le nuage *classes d'âge*×*cantons* est l'union de ces $T$ nuages centrés ; et dualement, le nuage des causes de mortalité est un nuage dans lequel les distances au carré sont les sommes des carrés des distances définies par chaque classe d'âge. Ainsi, une différence de mortalité entre deux cantons pour une certaine cause $j$, peu significative lorsque l'on étudie les classes d'âge séparément (du fait des faibles effectifs), peut devenir notable dans l'analyse intra si elle apparaît dans l'ensemble des classes d'âge.

➤ Classification intra

Il est possible d'appliquer un programme classique de classification aux facteurs obtenus par une analyse intra. On obtient ainsi des partitions des lignes ou des colonnes

suivant les proximités définies dans l'analyse intra ; ce résultat est quelquefois le principal objectif d'une telle analyse.

➤ Compléments d'aide à l'interprétation

Pour faciliter l'interprétation de cette analyse, on peut introduire, comme éléments supplémentaires, le tableau somme $r_{ij.}$. Dans l'AFC du tableau $r_{ijt}$, comme dans toute AFC, le profil de la somme de plusieurs colonnes est situé à leur barycentre. On obtient ainsi la projection des hommes et des femmes (ou des cantons dans l'exemple des causes de mortalité). Dans ce deuxième exemple où le nombre de barycentres est assez important, il est même possible d'appliquer au tableau $r_{ijt}$ une analyse par sous-tableaux en suivant la partition des colonnes induite par les cantons. Les résultats de cette analyse sont identiques à ceux de l'analyse intra mais l'on dispose, en plus, de toutes les aides à l'interprétation concernant les cantons.

➤ Nouvelle analyse dérivée de l'analyse intra

Dans le paragraphe ci-dessus, il est proposé d'utiliser la structure croisée $(J, T)$ des colonnes et d'introduire en éléments supplémentaires les colonnes du tableau de dimensions $I$ et $J$ obtenu en sommant sur $T$. On peut faire une autre analyse en prenant ce tableau somme en actif. Ceci présente de l'intérêt pour de grands tableaux, comme celui de l'exemple des causes de mortalité, qui ont des colonnes d'effectif assez faible : on cumule ainsi les effets des classes d'âge sans travailler directement sur ces colonnes. Il s'agit en quelque sorte d'une analyse inter de l'analyse intra. C'est aussi une analyse du tableau croisant *cantons* et *causes de mortalité* par rapport à un modèle ; ce modèle dérive de l'hypothèse d'indépendance conditionnelle par classe d'âge en moyennant sur l'ensemble des classes d'âge.

➤ Principaux aspects de l'analyse intra

En ce qui concerne les réponses apportées aux questions posées par la comparaison des tableaux, le bilan est rapide car nous en avons souligné les différents aspects tout le long de ce texte. L'analyse intra permet d'étudier :

1. la liaison entre deux variables en conditionnant par une troisième ;
2. les différences entre les profils des colonnes homologues de tous les tableaux ;
3. les différences entre les « évolutions d'un tableau à l'autre » des profils des lignes.

## 10.6 CONCLUSION

Il est assez étonnant de voir le nombre de traitements différents, et par-là même de regards différents, que l'on peut porter sur ces deux petits tableaux.

Cela illustre bien à la fois la complexité du problème des tableaux ternaires et la nécessité d'une réflexion préalable aux traitements qui précise les objectifs poursuivis. Cette réflexion est bien entendu indispensable dans toute analyse, mais les tableaux ternaires, de par leur complexité, exigent une formalisation particulièrement rigoureuse des problèmes. En effet, en permutant le rôle des trois variables, le nombre de traitements possibles devient considérable. La liste suivante, non exhaustive, en donne une idée.

1. **Trois analyses des trois marges binaires** avec les tranches binaires de même dimension que la marge en éléments supplémentaires (en lignes et en colonnes).
2. **Trois analyses de variables croisées** ($I$ avec $JT$, etc.) complétées chacune par deux séries d'indices correspondant à une décomposition du tableau suivant l'une ou l'autre des deux variables croisées.
3. **Six analyses intra** puisque dans l'analyse intra les trois variables sont traitées différemment : l'une est traitée séparément (l'ensemble $I$ des lignes dans notre exemple) et l'on peut alors conditionner par rapport à l'une ou l'autre des deux autres variables. De ces analyses intra dérivent aussi six analyses inter de la dispersion intra.
4. **L'analyse du tableau de Burt** obtenu en juxtaposant les tableaux croisant les variables deux à deux. Cette analyse est la seule qui fait jouer exactement le même rôle aux trois variables. Elle présente peu d'intérêt pour trois variables seulement (nous ne la citons que pour mémoire) pour deux raisons : d'une part, elle ne traite qu'une très faible part de l'information puisqu'elle ne retient du tableau ternaire que les trois marges binaires ; d'autre part, comme nous l'avons déjà dit au début de ce chapitre, lorsque l'on présente les données sous forme de tableau ternaire ou de suite de tableaux binaires, les problèmes ne se posent jamais en termes symétriques en $I$, $J$, $T$. Si l'on reprend l'exemple commenté, il est clair que l'analyse du tableau de Burt, qui n'est autre que l'analyse conjointe des liaisons binaires entre Emplois et Diplômes, entre Emplois et Sexes et entre Diplômes et Sexes ne présente guère d'intérêt.

Pour clore ce chapitre, évoquons le caractère spécifique de la dimension temporelle. Les tableaux ternaires sont souvent définis par une suite de tableaux binaires indicés par le temps. Le problème posé au statisticien s'exprime alors très souvent sous la forme du souhait de « faire entrer la notion de temps dans l'analyse ». Aucune technique ne permet d'intégrer réellement cette notion ; on considère tout au plus l'ordre induit par cette dimension. Cet ordre n'intervient même pas dans les calculs fondamentaux des techniques présentées. Il est possible cependant de le faire apparaître au moment de l'interprétation des résultats. Pour le matérialiser sur les graphiques de projection des nuages, on relie dans l'ordre les points qui représentent le même objet.

Chapitre **11**

# Interprétation des résultats d'une analyse factorielle

## 11.1 PROLÉGOMÈNES

Depuis toujours, ou presque, le terme « interpréter » est utilisé à propos de l'étude de résultats statistiques. L'Analyse Factorielle n'échappe pas à cette règle : les plans factoriels ne sont pas étudiés, ils sont interprétés. Ce terme recouvre plusieurs sens et, avant de proposer une démarche d'interprétation, il est utile au préalable d'en délimiter le champ. Pour cela, nous postulons que ce n'est pas par hasard si le terme « interpréter » a été d'abord choisi puis surtout consacré par l'usage. Nous avons donc utilisé un bon dictionnaire (le *Robert* !) pour explorer le champ sémantique de ce terme si employé. À l'issue de cette exploration, nous indiquons dans quelle mesure on peut proposer une démarche générale d'interprétation.

### 11.1.1 Interpréter, c'est d'abord rendre clair

Les données initiales sont nombreuses mais chacune est claire. (En principe, l'expérimentateur sait ce qu'il mesure : le fait que la vache numéro 77 pèse 623 Kg est une valeur claire, sinon sûre). L'analyse factorielle restitue des résultats (essentiellement, mais pas seulement, des coordonnées) moins nombreux mais **peu clairs en termes des données initiales**. Chaque coordonnée factorielle dépend *a priori* des valeurs de l'ensemble des colonnes pour l'ensemble des lignes : elle n'est pas immédiatement compréhensible et demande à être traduite en termes de données initiales (d'où la nécessité d'un interprète). Cette traduction des résultats factoriels en termes de données initiales est le premier aspect de l'interprétation. En ce sens, l'analyse factorielle est utilisée comme un outil d'exploration du tableau. Plutôt que de lire le tableau

directement, on le lit en traduisant les résultats factoriels. L'intérêt de l'analyse factorielle est alors de sélectionner, par ordre d'importance décroissante, les structures les plus marquantes du tableau. Cette étape est une description des faits statistiques indiscutables.

### 11.1.2 Interpréter, c'est aussi donner un sens

Donner un sens à un objet, c'est **l'intégrer dans un contexte**. Les données initiales sont claires mais, prises individuellement, n'ont pas de sens, pratiquement par définition. Le fait que l'individu 23 ait la valeur 3 pour la variable 11 est peu chargé de sens. De façon un peu simpliste, plus le contexte dans lequel on situe une information est grand et plus celle-ci est chargée de sens.

Le **premier contexte** dans lequel il convient de situer une valeur d'un tableau est le tableau lui-même. C'est d'ailleurs dans cet esprit que l'on peut présenter des transformations préalables aux analyses comme le centrage et la réduction des variables quantitatives, ou la transformation en profils d'un tableau de contingence. Il en est de même des résultats de l'analyse d'un tableau. Le grand pouvoir suggestif des plans factoriels provient de la visualisation simultanée de l'ensemble des lignes et des colonnes d'un tableau. Chaque élément actif, ligne ou colonne, est d'abord situé parmi l'ensemble des lignes et des colonnes actives ce qui constitue un premier contexte.

**Le deuxième contexte** est constitué par les éléments supplémentaires. Remarquons que c'est l'une des raisons d'être fondamentales des éléments supplémentaires, qui permettent par exemple de disposer :

- d'individus qui ne font pas partie de l'étude mais servent de points de repère ;
- de variables qui ne font pas partie du champ strict de l'analyse mais d'un champ voisin.

En étudiant ces éléments supplémentaires, on situe les résultats de l'analyse des éléments actifs dans un champ plus large et, ce faisant, on les charge de sens.

**Le troisième contexte** est extérieur aux données analysées. Il comprend l'expérience générale de l'analyste et ses connaissances sur le phénomène étudié. Seul ce contexte, qui est à l'échelle humaine, permet de donner véritablement un sens aux faits statistiques. Il est difficilement formalisable. On peut ici encore profiter de la possibilité d'éléments supplémentaires : il est utile de reprendre une analyse en ajoutant en supplémentaire des éléments suggérés par un premier passage. Cette introduction itérative d'éléments supplémentaires n'est rien d'autre qu'une traduction technique du mode de pensée associatif.

### 11.1.3 Interpréter, c'est enfin jouer de façon personnelle

Les termes d'art et d'artisanat sont souvent employés à propos de l'Analyse des Données. Même si ces termes sont très galvaudés, il est rare de les voir employés avec autant de régularité dans un domaine scientifique, en particulier par certains de ceux qui se réclament de ce domaine (les connotations d'artiste et d'artisan ne sont pas toutes positives). En tout cas, un examen même superficiel de quelques applications d'analyse factorielle offre des éléments qui ne sont pas sans rappeler l'interprétation d'une pièce de musique.

Le caractère personnel d'une interprétation réside surtout dans le mode de présentation des résultats. Cette présentation peut se limiter à quelques phrases qui résument les principales tendances observées dans les données. En particulier, les noms que l'on donne aux facteurs (par exemple « puissance du vin ») facilitent beaucoup ce type de synthèse. Elle peut contenir ou non des graphiques représentant les plans factoriels et leur commentaire. Elle peut contenir ou non des tableaux issus plus ou moins directement des données. Elle peut aussi décrire les données en indiquant et expliquant les regroupements de lignes ou de colonnes sur les différents graphiques.

Dans tous les cas, cette présentation ne peut expliciter l'intégralité de la richesse des données. On est ainsi conduit à choisir les faits les plus saillants, les plus intéressants. Ce choix, dans lequel l'analyste s'implique, peut différer d'un analyste à l'autre. Cela alarme quelquefois les débutants qui éprouvent des difficultés à séparer ce qui est automatique (qualifié aussi d'objectif) et ce qui est personnel (qualifié aussi de subjectif) dans une interprétation.

- Est automatique le tri des faits statistiques présents dans un tableau par importance décroissante. L'importance est ici mesurée par un critère statistique fondé sur le concept d'inertie.
- Est personnelle, la réévaluation de ces faits à la lumière des connaissances de l'analyste sur le problème étudié qui sont extérieures au tableau de données. Il en résulte un nouveau poids des informations, déterminant dans la présentation des résultats.

Par exemple, le regroupement de lignes et/ou de colonnes sur un graphique présente un aspect objectif (la proximité des points sur le plan) et un aspect subjectif (un nuage de points réalise généralement un continuum que l'on scinde en un ensemble de groupes de points dont les frontières ne s'imposent pas). Lorsque plusieurs possibilités sont également raisonnables du point de vue des proximités, on regroupe plutôt des éléments qui ont un caractère commun, souvent extérieur aux données traitées, mais cependant connu et considéré comme important (voire explicatif) par celui qui dépouille les résultats. On obtient ainsi des groupes, homogènes à la fois du point de vue des variables actives et d'autres critères jugés importants, présentant un fort pouvoir évocateur.

Une interprétation est aussi personnalisée du fait de certains choix à caractère plus ou moins technique. On peut jouer, par exemple, sur le ressort de la dualité : dans une ACP par exemple, il peut être plus clair de parler des principales dimensions de variabilité (on privilégie alors les variables) ou de tendances représentées par des classes d'individus que l'on décrit. Une autre alternative importante est : faut-il commenter les axes ou les plans ? On est souvent tenté d'orienter un commentaire de plan selon d'autres directions – pas forcément orthogonales – que les axes factoriels (cas d'une bissectrice dans un plan issu de l'enquête *Ouest-France* du chapitre 6 page 127).

Les résultats issus d'une Analyse Factorielle posent le problème de la démarche d'interprétation, c'est-à-dire de l'ordre chronologique dans lequel ces différents résultats doivent être examinés. Dans les sections suivantes, nous proposons une démarche d'interprétation pour chacune des méthodes factorielles étudiées dans cet ouvrage.

- La première présentation se réfère à l'ACP : elle est la plus détaillée en ce sens qu'elle introduit les aspects généraux communs à toutes les méthodes.
- Les autres présentations s'appuient sur ce premier schéma, en développant uniquement les points sur lesquels la démarche d'interprétation diffère, entre la méthode examinée et l'ACP.
- Enfin, en guise de conclusion, une dernière section récapitule quelques types de facteurs auxquels peut conduire l'interprétation.

## 11.2 INTERPRÉTATION D'UNE ACP

Dans cette section, l'essentiel de la présentation se réfère à l'ACP normée : nous réservons un paragraphe aux spécificités de l'ACP non normée. Hormis celui-ci, tous les paragraphes sont classés selon un ordre chronologique de dépouillement des résultats qui constitue une démarche générale d'interprétation.

Dans cet ordre chronologique, deux phases principales ont été distinguées.

- Un bilan sur les inerties associées aux différents facteurs, qui ne se préoccupe pas de la signification des facteurs, mais se fonde seulement sur des indices numériques.
- L'interprétation proprement dite des facteurs, difficilement formalisable, qui donne une large place aux connaissances sur le problème étudié extérieures au tableau de données.

### 11.2.1 Étude de l'inertie des facteurs

La première phase de l'analyse permet d'étudier les grands traits de la forme des nuages et l'importance globale des liaisons entre variables.

### *a) Valeurs propres*

Rappelons que la première valeur propre est toujours comprise entre 1 et le nombre de variables $K$. Elle vaut 1 lorsque les variables sont toutes non corrélées deux à deux. Elle est égale à $K$ lorsqu'il existe une liaison linéaire parfaite entre toutes les variables. Dans le cas limite d'une première valeur propre proche de 1, on est conduit à deux attitudes différentes selon l'objectif de l'analyse :

- considérer l'ensemble des dimensions si l'on cherche un résumé des données ;
- ne considérer aucune dimension si l'on s'intéresse aux liaisons entre variables.

Plus la valeur propre est grande, plus elle résume de variables et plus le facteur risque d'être intéressant en terme de synthèse. La situation est claire pour la première valeur propre puisque l'on connaît ses valeurs extrêmes. Pour les valeurs propres suivantes, la valeur 1 reste un point de repère : une composante principale est une variable synthétique, et une valeur propre associée inférieure à 1 indique que cette variable synthétise moins de données qu'une variable isolée. Il convient donc de redoubler de prudence dans l'interprétation d'un facteur associé à une valeur propre proche ou inférieure à 1. La valeur 1 ne peut toutefois être utilisée comme seuil absolu : l'expérience fournit, à l'occasion, des facteurs clairement interprétables dont l'importance très faible relativement aux autres conduit à une valeur propre inférieure à 1.

Enfin, il est quelquefois utile de considérer le nombre de valeurs propres « pratiquement nulles », ce qui permet de calculer la dimension réelle des données analysées.

**Le diagramme des valeurs propres**, appelé souvent abusivement histogramme, est utilisé surtout pour étudier l'allure de la décroissance de ces valeurs. Le principe de lecture de ce diagramme est le suivant : si deux facteurs sont associés à des valeurs propres presque égales, ils représentent la même part de variabilité et il n'y a pas lieu *a priori* de retenir l'un et non l'autre dans l'interprétation. Réciproquement, une forte décroissance entre deux valeurs propres successives incite à retenir dans l'interprétation les facteurs précédant cette décroissance.

Dans la pratique, on observe souvent le phénomène suivant : les $S$ premières valeurs propres présentent une décroissance assez irrégulière ; puis, au delà du rang $S$, la décroissance est lente et régulière. Cette allure indique que les $S$ premiers facteurs correspondent chacun à des irrégularités dans la forme du nuage de points étudié qui demandent à être interprétées et suggère que les facteurs suivants ne représentent que l'inévitable bruit qui accompagne toute observation de nature statistique.

Cas extrême, une décroissance lente et régulière dès la première valeur propre traduit un nuage à peu près « sphérique » et donc des données peu structurées dont les facteurs sont peu synthétiques. Un diagramme de ce type présage un intérêt limité des facteurs.

### *b) Pourcentages d'inertie extraits par les facteurs*

Le pourcentage d'inertie extrait par un facteur est le rapport entre l'inertie associée au facteur (*i.e.* la valeur propre) et l'inertie totale du nuage étudié ; il mesure l'importance relative du facteur dans le tableau. Il est souvent utilisé sous la forme cumulée qui indique le pourcentage d'inertie extrait par les $S$ premiers facteurs.

Il ne faut pas oublier de juger ces pourcentages en fonction de la taille du tableau : 10 % est une valeur faible si le tableau comporte 10 variables (elle est égale à la moyenne et correspond à la valeur propre 1) ; c'est une valeur forte dans le cas de 100 variables.

### *c) Quel nombre de facteurs retenir ?*

À propos des valeurs propres et des pourcentages d'inertie, on a évoqué à plusieurs reprises les pronostics que suggèrent ces indicateurs quant à l'intérêt des facteurs. Poursuivant cette démarche, certains ont demandé à ces indicateurs plus que des pronostics, à savoir une règle de décision quant au nombre de facteurs à retenir dans l'interprétation. Pour cela, on se réfère à une situation de parfaite indépendance des variables qui se traduit par une isotropie des nuages étudiés ; on examine ensuite si l'importance absolue (jugée à partir des valeurs propres) ou relative (jugée à partir des pourcentages d'inertie) des facteurs effectivement obtenus peut être considérée comme grande en regard de la situation de référence. Cette démarche doit être évitée pour un faisceau de raisons convergentes.

- La situation de référence ne correspond à aucune situation concrète ; le tableau que l'on étudie est toujours choisi (plus ou moins soigneusement certes) mais jamais tiré au hasard.
- Toute règle concernant des facteurs de rang supérieur à 1 doit tenir compte du ou des facteurs de rang précédent, ce qui rend le problème pratiquement inextricable.
- L'importance d'un facteur n'est absolument pas un gage de son intérêt. Situation paradoxale (en apparence) bien connue : les facteurs de rang 3 et 4 présentent souvent de l'intérêt précisément parce qu'ils apportent des informations difficiles à voir sur le tableau des données.
- Les critères fondés sur l'inertie ne permettent pas de préjuger de l'intérêt des facteurs, lui-même dépendant d'éléments extérieurs aux données (objectifs de l'analyse, degré de connaissance sur le problème étudié, etc.).

À l'écart de cette fausse route et exploitant la dernière remarque, la règle suivante est tout à fait recommandable : on retient dans l'interprétation d'une analyse les facteurs que l'on sait clairement interpréter. En effet :

- il serait dommage de rejeter avec des critères statistiques un facteur que l'on sait interpréter ;
- il serait délicat de mettre en avant un facteur que l'on ne sait pas interpréter.

### 11.2.2 Interprétation des facteurs

Les facteurs sont appréhendés dans l'ordre décroissant de leurs valeurs propres. Ils peuvent être étudiés séparément ou deux par deux à l'aide des plans factoriels. Il faut constamment garder à l'esprit que le facteur d'ordre $s$ ($s > 1$) traduit les tendances « résiduelles » non prises en compte par les facteurs précédents.

L'ordre proposé pour dépouiller les résultats correspond à une phase de découverte. L'approfondissement d'une interprétation donne toujours lieu à des va-et-vient entre les différents résultats, trop liés aux données et à l'analyste pour être formalisés. En particulier, du fait de la dualité, on est souvent conduit à consulter alternativement les résultats concernant les individus et les variables.

#### *a) Contributions des individus*

L'intérêt d'un facteur dépend en grande partie du nombre d'individus qu'il concerne. On réalise une première approche de ce nombre en consultant la liste des contributions des individus aux facteurs pour repérer si un seul individu ou un très petit nombre d'individus ont une contribution très supérieure à la moyenne. On peut calculer aussi le nombre minimum d'individus totalisant, à eux tous, un pourcentage d'inertie projetée fixé à l'avance (par exemple 50 %). Cet indicateur évalue le degré de généralité d'un facteur au sens du nombre d'individus participant à ce facteur.

Le premier stade de l'interprétation d'un facteur qui apparemment ne concerne que très peu d'individus est en général simple : on identifie rapidement ces individus et leur particularisme. La signification de ce particularisme est plus ou moins immédiate ; elle peut remettre en cause le champ de l'analyse, à savoir l'ensemble des individus étudiés.

Envisageons le cas extrême d'un facteur induit par un seul individu. Deux cas peuvent être distingués.

– Si ce facteur est l'un des premiers, l'individu concerné est nécessairement très différent des autres. Un tel cas particulier est d'une part facilement mis en évidence sans l'analyse et, d'autre part, gêne l'étude du reste de la population. Il faut alors envisager de refaire une analyse en supprimant cet individu des éléments actifs, ce qui modifie le champ de l'étude. Cette nouvelle analyse peut ne différer que de très peu de la première. En effet, on peut montrer que si l'inertie sur l'axe $s$ de l'individu supprimé est inférieure à la différence entre $\lambda_s$ et $\lambda_{s+1}$ , les facteurs de la nouvelle analyse sont très corrélés à ceux de l'ancienne ; l'individu est certes très différent des autres, mais comme cette différence s'inscrit dans une tendance générale il ne perturbe pas les résultats.

– Si l'on observe un tel facteur après quelques facteurs généraux prenant en compte beaucoup d'individus, l'analyse n'est pas nécessairement remise en cause : il est naturel, après avoir extrait des tendances générales, que des phénomènes ponctuels apparaissent.

**Attention** : il ne peut être question d'exclure des individus d'une analyse en se fondant uniquement sur des critères d'inertie car cette exclusion implique une modification des objectifs. Un exemple fictif illustrera cette situation. Supposons que l'étude porte sur les 120 exploitations agricoles orientées vers l'élevage laitier d'une région et que le premier axe mette en évidence le caractère exceptionnel de l'exploitation 27. Renseignements pris, on s'aperçoit que cette exploitation est rattachée à une Ecole d'Agronomie bien connue, alors que les autres sont de structure familiale classique. Exclure cette exploitation revient à modifier le thème de l'étude qui devient l'étude des exploitations **familiales** orientées vers l'élevage laitier.

Remarquons enfin qu'en ACP normée, ce problème d'éléments exceptionnels ne concerne que les individus. En effet, les variables possèdent chacune la même inertie.

### *b) Coordonnées des variables actives*

Il est naturel de commencer l'examen détaillé des graphiques par ce que l'on connaît le mieux. Généralement, les variables sont moins nombreuses et plus chargées de sens que les individus.

Par ailleurs, il est logique de privilégier, au moins dans un premier temps, les éléments actifs : l'interprétation d'un facteur doit se fonder d'abord sur les données qui ont participé directement à sa construction.

Rappelons que, en ACP normée, les variables ayant le même poids et étant équidistantes de l'origine, le carré de leur coordonnée sur un axe se confond avec leur qualité de représentation et est proportionnelle à leur contribution. Aussi, on limite généralement l'étude des variables à celle de leurs coordonnées. À ce niveau, l'interprétation s'appuie essentiellement sur la règle suivante : la coordonnée de la variable $k$ le long de l'axe factoriel $s$ est le coefficient de corrélation entre cette variable $k$ et le facteur $s$.

#### ➤ Interprétation axe par axe

On recense les variables actives les plus liées à chaque axe. Deux situations typiques peuvent se produire.

– Toutes les variables très liées au facteur sont situées d'un même côté de l'axe (cas de l'exemple des vins du chapitre 7 page 149). Le facteur apparaît alors comme une synthèse entre ces variables. L'effet taille (cité section 1.6) dans lequel toutes les variables sont situées d'un même côté de l'axe peut être rattaché à cette situation typique.

– Les variables très liées au facteur présentent une coordonnée positive pour les unes et négative pour les autres. Il faut alors rechercher un dénominateur commun qui, à la fois, relie les variables situées du même côté et oppose les variables situées de part et d'autre de l'origine. Par exemple, supposons que les variables soient des notes dans différentes matières : un facteur peut traduire l'opposition entre matières scientifiques et matières littéraires. Cette phase permet déjà d'obtenir la signification générale de certains axes.

➤ Interprétation par plan

Comparativement à l'étape précédente, le plan factoriel apporte le pouvoir synthétique du graphique, plus suggestif qu'une liste de coordonnées, et la prise en compte simultanée de deux dimensions qui donne une image plus fidèle des données et peut aussi suggérer d'interpréter d'autres directions que les axes factoriels. Il est utile de représenter en plus des points (variables) :

– le cercle de rayon 1, ou cercle des corrélations, car la proximité d'un point au cercle permet de juger aisément de la qualité de représentation des variables ;
– les vecteurs joignant l'origine aux points variables afin de visualiser les angles qui mesurent la liaison entre variables.

Le plan factoriel permet une approche des données qui laisse de côté, au moins en apparence, les facteurs eux-mêmes. Cette approche consiste en un bilan des liaisons entre variables. Les angles entre variables étant déformés par la projection, on limite ce bilan aux variables bien représentées (c'est-à-dire dont l'image est proche du cercle de corrélation). Il est ainsi possible de regrouper visuellement (ce qui est d'autant plus précieux que les variables sont nombreuses) des variables liées entre elles et d'esquisser ainsi une typologie des variables.

La construction des plans factoriels implique la détermination des facteurs que l'on va croiser. Pour cela, on s'appuie sur deux éléments.

– L'inertie associée aux facteurs. On croise de préférence des facteurs d'importance comparable car, dans le cas de deux facteurs associés à des valeurs propres égales, c'est le plan formé par ces deux facteurs qui est stable et non les facteurs eux-mêmes. On est ainsi conduit à construire la suite de plans qui croisent les facteurs 1 et 2, les facteurs 2 et 3, etc.
– La signification du facteur. On peut vouloir focaliser son attention sur certaines variables et donc sur les plans qui en fournissent une bonne représentation.

### *c) Coordonnées des variables supplémentaires*

Le rôle des variables supplémentaires est d'élargir le contexte d'interprétation. On recense, parmi ces variables, celles qui sont très liées aux facteurs : cela permet éventuellement d'expliquer certains facteurs ou d'affiner les interprétations déjà proposées

et/ou peut suggérer de réexaminer un facteur délaissé sur la seule vue des variables actives. Ce dernier point, qui donne une certaine prééminence aux variables supplémentaires, est important. L'existence de variables supplémentaires très liées à un facteur, en tant que validation *a posteriori*, fournit une forte présomption selon laquelle ce facteur est chargé de sens.

### *d) Coordonnées et aides à l'interprétation des individus actifs*

Plutôt que l'étude des coordonnées, fastidieuse si les individus sont nombreux, on examine le plan pour trois raisons essentielles.

- Étudier l'allure générale de la répartition de l'ensemble des individus. Toute plage de très faible densité ou de très forte concentration doit être décelée.
- Aider le choix d'individus types qui permettent de concrétiser les dimensions de variabilité. Dans le choix d'individus types, il est bon de consulter les qualités de représentation pour sélectionner de préférence des individus qui ne sont caractéristiques que du ou des facteurs étudiés et sont donc moyens pour les autres facteurs. Par l'intermédiaire de ces individus, il est commode de relier les facteurs aux données initiales.
- Faire apparaître une typologie des individus, en délimitant des domaines connexes communément appelés « patatoïdes ». Par rapport à un résultat de classification, ces typologies présentent deux caractéristiques. La première est de se fonder sur un plan, c'est-à-dire seulement deux axes (il est ainsi possible d'obtenir plusieurs typologies différentes, correspondant chacune à un plan donc à un aspect des données) : ceci limite leur valeur statistique au sens du rapport *inertie inter* / *inertie totale* mais leur forte adéquation à un plan est un avantage si ce dernier est prépondérant dans les interprétations. La deuxième est qu'elles peuvent tenir compte d'informations extérieures aux variables actives en favorisant le regroupement d'individus possédant des caractères communs. Sans perdre nécessairement beaucoup de valeur statistique du point de vue des variables actives, on facilite ainsi grandement l'interprétation des classes.

Il est souvent nécessaire de regarder la répartition des individus appartenant à une même sous-population. On peut identifier sur les graphiques les individus par leur modalité pour une variable qualitative (dans l'exemple des vins du chapitre 7 page 149, ceux-ci sont représentés par un signe indiquant leur origine). Cette pratique est une façon très fine de faire intervenir dans une ACP des variables qualitatives en tant qu'éléments supplémentaires.

On peut aussi représenter les barycentres de ces populations en introduisant en lignes supplémentaires les moyennes des individus appartenant à la même sous-population. Certains logiciels permettent même de représenter les axes d'inertie des projections des sous-nuages ce qui permet de voir l'allure générale du sous-nuage sur le plan. Cela est

particulièrement intéressant dans le cas où les individus sont nombreux et où la seule information que l'on possède sur eux est constituée par les données. Cette situation est typiquement celle des enquêtes. L'ensemble des individus ne présente alors de l'intérêt que dans la mesure où il permet d'accéder à une population encore plus vaste. Il est clair que de telles analyses se situent dans une perspective inférentielle : le fait que cette inférence soit formalisée de façon assez lâche n'implique pas qu'elle soit sans valeur pratique (certains esprits facétieux disent même : au contraire !).

#### *e) Coordonnées et aides à l'interprétation des individus supplémentaires*

C'est un peu par principe que l'étude des individus supplémentaires ne vient qu'après celle des individus actifs. Cet ordre s'applique bien aux individus mis en supplémentaires parce qu'ils s'écartent des autres. En revanche, il s'applique moins bien lorsqu'il s'agit d'un individu supplémentaire servant de point de repère ou représentant le centre de gravité d'une classe. Ces derniers individus supplémentaires, finalement plus chargés de sens que les actifs, sont généralement moins nombreux et peuvent intervenir, dans le dépouillement, juste après l'examen de la répartition des individus actifs.

### 11.2.3 Cas de l'ACP non normée

Cette analyse peut être considérée comme une ACP normée dans laquelle on affecte à chaque variable un poids égal à sa variance. Le fait d'analyser un ensemble de variables pondérées ne modifie pas les grandes lignes de l'interprétation mais influe sensiblement sur quelques résultats.

➤ Valeurs propres

L'inertie de chaque variable ne vaut pas systématiquement 1. Les valeurs propres ne sont donc pas comparables d'une analyse à l'autre. Le seuil de 1 n'a plus de signification. On s'appuie plutôt sur les pourcentages d'inertie pour apprécier l'importance d'un facteur.

➤ Stabilité et degré de généralité d'un facteur

Les variables étant munies de poids, le premier axe factoriel peut parfaitement être dû à une seule variable. Il s'ensuit que l'on examinera en premier lieu non seulement les contributions des individus mais aussi celles des variables pour détecter d'éventuels éléments prépondérants.

➤ Coordonnées des variables

C'est seulement si l'on a effectivement réalisé une ACP normée de variables pondérées, et non une ACP non normée, que les coordonnées des variables actives s'interprètent encore comme des coefficients de corrélation. Le carré de cette coordonnée

mesure alors la qualité de représentation mais n'est plus proportionnel à la contribution. Finalement, en ACP non normée, les deux représentations des variables (par leurs corrélations et par leurs covariances) sont utiles.

## 11.3 INTERPRÉTATION D'UNE AFC

En AFC, les lignes et les colonnes sont des objets de même nature (des modalités de variables qualitatives) qui jouent des rôles symétriques, analogues dans une certaine mesure à celui des individus en ACP. Il s'ensuit que la démarche dans l'interprétation d'une AFC est voisine dans ses grandes lignes de celle d'une ACP mais en diffère sur certains points. Ce paragraphe reprend globalement le plan utilisé pour l'ACP mais ne détaille que les aspects sur lesquels les deux méthodes diffèrent. L'essentiel de ces remarques concerne l'application de l'AFC à un tableau de contingence. Un paragraphe final envisage d'autres cas.

### 11.3.1 Valeurs propres

Les valeurs propres sont inférieures ou égales à 1, valeur atteinte lorsqu'un axe rend compte de façon parfaite d'une association entre une partition des lignes d'une part et une partition des colonnes d'autre part (*cf.* Figure 3.9 page 78). Ainsi, un facteur associé à une valeur propre voisine de 1 exprime une forte liaison entre les lignes et les colonnes qu'il sera toujours facile de traduire en termes de données initiales. En revanche, une valeur propre faible (pour fixer les idées, indiquons l'ordre de grandeur de 0.1) correspond à une liaison faible : le facteur associé devra être interprété avec précaution en s'appuyant sur les données.

Au nombre d'individus près, la somme des valeurs propres est égale à l'indice $\chi^2$ mesurant la liaison entre deux variables qualitatives. En AFC, on s'intéresse peu à cette valeur qui est un indice global et ne permet guère de préjuger de l'intérêt des facteurs. Cette somme n'étant pas constante, le pourcentage d'inertie extrait par un facteur ne se déduit pas de la valeur propre et du nombre de colonnes. Les valeurs propres et les pourcentages d'inertie sont des informations indépendantes qu'il est utile de consulter pour juger numériquement de l'intérêt d'un facteur.

### 11.3.2 Contributions des lignes et des colonnes

Comme en ACP, il importe de s'assurer qu'un nombre suffisant d'éléments contribue aux premiers facteurs. La démarche est la même qu'en ACP à la différence près qu'elle s'applique aux lignes et aux colonnes.

Dans l'AFC d'un tableau de contingence, la mise en évidence de facteurs dus à un très petit nombre d'éléments est plus embarrassante qu'en ACP : en effet le recours à la mise en supplémentaire d'une ligne ou d'une colonne est délicat en AFC puisqu'il

conduit à étudier la liaison entre deux variables en ne considérant qu'un sous-ensemble de modalités. Si l'on opte pour cette solution, il faut préciser avec soin la modification du champ de l'étude qu'elle implique. Il existe, en AFC, la possibilité de regrouper des lignes et/ou des colonnes. Cette possibilité n'est toutefois pas très efficace pour contourner le problème de facteurs dus à un très petit nombre d'éléments car, dans ce cas, elle conduit à regrouper des modalités de profils différents, ce qui rend difficile l'interprétation des modalités ainsi obtenues.

### 11.3.3 Coordonnées des éléments actifs

La tactique, présentée à propos de l'ACP, qui consiste à étudier d'abord les axes au vu des listes de coordonnées puis des plans s'applique ici. Naturellement, il n'y a aucune raison, en AFC, pour toujours commencer l'interprétation par l'étude des lignes ou des colonnes. Néanmoins, il semble y avoir quelque avantage dans l'attitude systématique qui consiste à interpréter un axe d'abord en fonction d'un ensemble puis de l'autre, les associations entre lignes et colonnes n'étant exploitées que dans un second temps.

Dans le cas général, en AFC, les éléments ont des poids différents. Aussi, la coordonnée d'un point, sa qualité de représentation et sa contribution à l'inertie constituent des informations différentes. Pour interpréter un facteur, on s'appuie de façon privilégiée sur les éléments types qui présentent :

- une forte contribution ; leur importance provient de ce que leur suppression de l'ensemble des éléments actifs risque d'entraîner la disparition du facteur ;
- une coordonnée extrême jointe à une forte qualité de représentation ; ces éléments sont les plus commodes pour qualifier un facteur : ils sont très différents du profil moyen (leur coordonnée est extrême) et cette différence est presque entièrement traduite par le facteur (ils ont une bonne qualité de représentation) ;
- une coordonnée extrême jointe à une qualité de représentation moyenne ; ils présentent à un fort niveau les caractéristiques associées au facteur, ce qui leur donne une grande valeur. Mais ces caractéristiques s'additionnent à d'autres, ce qui les rend plus difficiles à mettre clairement en évidence.

### 11.3.4 Cas de modalités ordonnées ou partitionnées

Fréquemment, il existe une structure *a priori* sur l'un ou les deux ensembles mis en correspondance. Ainsi, les modalités de la variable « niveau de diplôme » peuvent être *a priori* ordonnées selon le nombre d'années d'études nécessaires, ou partitionnées selon le critère *enseignement technique* / *enseignement général*. L'analyse de ce type de données comporte toujours la recherche des facteurs mettant en évidence de telles structures.

### 11.3.5 Cas dans lesquels le tableau analysé n'est pas un tableau de contingence

L'AFC peut être employée avec profit dans l'analyse de différents types de tableaux. Le cas du Tableau Disjonctif Complet est suffisamment important (de par le nombre de ses applications) et spécifique pour mériter un paragraphe particulier (*cf.* section 11.4). Un cas fréquent est celui où le tableau analysé résulte de la juxtaposition de tableaux de contingence.

Dans l'ensemble, les règles d'interprétation précédentes demeurent inchangées. Les valeurs propres restent comprises entre 0 et 1 et la valeur de 1 correspond toujours à une association parfaite entre une partition des lignes et une partition des colonnes en deux classes. Toutefois, la somme des valeurs propres ne s'interprète plus comme un $\chi^2$.

## 11.4 INTERPRÉTATION D'UNE ACM

Fondamentalement, comme l'ACP, l'ACM s'applique à un tableau croisant des individus et des variables (c'est la nature des variables qui change d'une technique à l'autre), mais les calculs auxquels elle conduit consistent en une AFC sur tableau disjonctif complet. Dès lors, il faut s'attendre à ce que la démarche d'interprétation d'une ACM s'apparente à la fois de celle de l'ACP et à celle de l'AFC.

Dans ce qui suit, nous notons $I$ le nombre d'individus, $J$ le nombre de variables et $K$ le nombre total de modalités.

### 11.4.1 Inertie de facteurs

➤ Valeurs propres

La somme des valeurs propres est égale à $(K/J)1$, rapport entre le nombre de modalités et le nombre de variables, le tout diminué de 1. Comme en ACP, et à la différence de l'AFC simple, elle ne dépend pas de la structure des données.

En pratique, on observe que les valeurs propres sont faiblement et régulièrement décroissantes : l'allure générale de « l'histogramme » des valeurs propres est rarement suggestive en ACM.

La valeur propre associée à un facteur est égale à la moyenne des rapports de corrélation entre le facteur et chaque variable (*cf.* section 4.3.6 page 96). Elle vaut 1 si tous les rapports de corrélation sont égaux à 1 donc si **pour chaque variable** tous les individus présentant la même modalité sont situés au même point. Cette situation constitue un extrême dont on est toujours très loin en pratique : il s'ensuit que les valeurs propres sont souvent très faibles en ACM.

➤ Pourcentages d'inertie

Une variable à $r$ modalités est représentée par un sous-espace de dimension $r - 1$ (*cf.* section 4.3.5 page 95). Lorsqu'un facteur est très lié à cette variable (c'est-à-dire si le rapport de corrélation entre la variable et le facteur vaut 1), le pourcentage d'inertie extrait de cette variable est $100/(r-1)$. Il en résulte que, lorsque les variables possèdent un grand nombre de modalités, même les pourcentages d'inertie associés aux premiers facteurs sont, du fait de la nature du tableau, très faibles.

➤ Bilan sur les valeurs propres et les pourcentages d'inertie

La représentation des modalités, en ACM, peut indifféremment être obtenue par une AFC sur le Tableau Disjonctif Complet ou sur le tableau de Burt. Or, d'une analyse à l'autre, le même facteur n'est pas associé à la même valeur propre. Cette remarque, ainsi que les considérations précédentes, expliquent que les valeurs propres et les pourcentages d'inertie ont peu d'influence sur l'interprétation d'une ACM.

### 11.4.2 Contributions des individus et des modalités

Pour identifier d'éventuels éléments prépondérants, l'étude des axes d'une ACM commence par l'étude des contributions des individus.

Comme en ACP, les variables ne peuvent être « aberrantes », mais il est possible en ACM que le ou les premiers facteurs soient dus à un petit nombre de modalités. Cela peut se produire s'il existe des modalités de faible effectif partagées par les mêmes individus puisque le carré de la distance d'une modalité au centre de gravité est inversement proportionnel à son effectif (*cf.* section 4.3.3 page 92). Lorsque l'examen des contributions des modalités indique qu'un petit nombre de modalités est largement prépondérant, les individus qui présentent cette ou ces modalités possèdent généralement aussi une contribution très grande. Aussi, en ACM, lorsque l'on cherche à éliminer un facteur s'appuyant sur un trop petit nombre d'éléments, il faut examiner simultanément la mise en supplémentaire de lignes et la suppression ou le regroupement de modalités.

### 11.4.3 Contributions des variables

En sommant pour le facteur de rang $s$ les contributions des modalités d'une même variable, on obtient la contribution de la variable à ce facteur. Cette contribution est égale, au coefficient $J\lambda_s$ près ($\lambda_s$ : inertie associée au facteur du rang $s$ ; $J$ : nombre de variables), au rapport de corrélation entre la variable et le facteur. Il en résulte que :

– en ordonnant les variables par contribution décroissante, on peut sélectionner les variables les plus liées à un facteur, c'est-à-dire celles sur lesquelles l'interprétation pourra s'appuyer de façon privilégiée ;

– il peut être intéressant de réaliser des graphiques dans lesquels les variables ont pour coordonnée sur l'axe $s$ leur contribution au facteur de rang $s$ (*cf.* figure 4.6 page 98). Ce graphique facilite la sélection précédemment citée et fournit une visualisation des proximités entre variables.

L'interprétation de l'inertie projetée des variables en tant que rapport de corrélation fait qu'il est intéressant de calculer cette quantité aussi pour les variables supplémentaires.

### 11.4.4 Coordonnées des modalités et des individus

L'étude des coordonnées des modalités précède presque toujours celle des individus. La démarche qui consiste à étudier d'abord pour chaque axe (au vu des listes de coordonnées) les éléments actifs puis les supplémentaires, puis les plans, est semblable à celle (décrite en détail à propos de l'ACP) des autres analyses factorielles.

Le cas des modalités ordonnées est fréquent dans la pratique de l'ACM. On commence toujours, dans l'étude des coordonnées, par repérer les facteurs sur lesquels les modalités des variables ordonnées se trouvent dans leur ordre naturel (sur les graphiques, on relie ces modalités dans leur ordre naturel).

La qualité de représentation des modalités est elle-même un indicateur peu pertinent. En effet, les modalités d'une même variable étant orthogonales, elles ne peuvent être simultanément bien représentées sur un axe. En outre, une modalité est généralement perçue comme le centre de gravité des individus qui la possèdent (*cf.* la propriété barycentrique en ACM) : or la qualité de représentation d'une modalité est différente de celle du centre de gravité correspondant (*cf.* section 4.3.4 page 94).

La démarche dans l'étude des individus est la même qu'en ACP, les individus actifs étant toujours très nombreux en ACM.

## 11.5 INTERPRÉTATION D'UNE AFM

L'AFM fait intervenir trois types d'objets : les individus, les variables et les groupes de variables. Les règles d'interprétation concernant les individus et les variables sont globalement les mêmes qu'en ACP et ACM. À ce niveau et par rapport à ces méthodes, le fait d'avoir pris en compte la structure en groupes infléchit les résultats mais ne modifie pas leur nature fondamentale. En revanche, l'AFM fournit des résultats spécifiques de la structure en groupes qui possèdent leurs règles d'interprétation propres. La présente section précise ces règles et indique quelle place accorder à l'examen de ces résultats dans une démarche d'interprétation.

### 11.5.1 Résultats de l'analyse séparée de chaque groupe

On regarde le diagramme des valeurs propres de chaque groupe, essentiellement pour évaluer le nombre de dimensions qui interviendront de manière significative dans l'analyse globale. La surpondération de l'AFM fait que seule la forme de ce diagramme importe : l'importance ultérieure d'une valeur propre est déterminée par son rapport avec la première de ces valeurs. Ces diagrammes permettent aussi de comparer la forme générale des nuages définis par chaque groupe, sans tenir compte des éléments qui le composent.

Ainsi, comme la pondération des variables dans l'analyse globale respecte la structure des groupes, on détecte à ce niveau les groupes presque unidimensionnels qui ne peuvent influencer plusieurs facteurs et les groupes fortement multidimensionnels qui influencent plusieurs facteurs.

### 11.5.2 Valeurs propres de l'analyse globale

Les valeurs propres peuvent être considérées comme des indices de liaison entre le facteur associé et l'ensemble des groupes dans la mesure où la valeur maximum possible – le nombre $J$ de groupes actifs – n'est atteinte que lorsqu'un facteur de l'analyse globale est confondu avec le premier facteur de l'analyse séparée de chaque groupe. Le parallèle avec l'ACM doit être fait : en ACM, le maximum est atteint lorsque toutes les partitions définies par les variables qualitatives sont totalement respectées (*i.e.* les individus d'une classe définie par chaque modalité ont la même coordonnée sur le facteur). Ce parallèle est d'autant plus licite que l'AFM se confond avec l'ACM dans le cas où chaque groupe comporte une seule variable qualitative.

Une attention toute particulière sera accordée à la première valeur propre. Si elle est proche de $J$, le premier facteur est à la fois commun à l'ensemble des groupes et représente une direction d'inertie importante dans chacun d'eux. Si elle est faible, on ne peut rien en dire (sinon que l'on n'est pas dans le cas précédent). Les valeurs propres suivantes ne peuvent être interprétées de la même façon puisque leur valeur maximum dépend de la structure de chacun des groupes : du fait de la pondération, si chacun des groupes présente un facteur prépondérant, la seconde valeur propre de l'analyse globale est nécessairement faible, même si elle correspond au deuxième facteur de chaque groupe. En revanche, le diagramme des valeurs propres et des pourcentages d'inertie se lit comme dans les autres méthodes factorielles.

### 11.5.3 Relations entre les facteurs de l'analyse globale et les groupes

L'étude de ces relations est la première étape du dépouillement de l'analyse globale. Il est en effet préférable d'avoir d'abord une idée de la structure générale des données avant de s'intéresser à des aspects plus précis mais plus parcellaires. L'expérience a d'ailleurs montré qu'il est plus efficace, si l'on souhaite réaliser des analyses séparées

complètes de chaque groupe, de les faire après l'AFM. Nous proposons d'étudier les indices concernant les liens entre les groupes et les facteurs dans l'ordre suivant.

➤ Corrélations entre les facteurs communs et leurs représentants dans les groupes

Lorsque les corrélations entre un facteur de l'ensemble des groupes et ses représentants dans tous les groupes sont proches de 1, il s'agit d'un facteur commun aux groupes (*cf.* section 8.3.5 page 187). Comme les groupes que l'on étudie simultanément sont généralement liés entre eux (ce qui est conforme à l'intuition de l'analyste qui les étudie simultanément), il y a au moins un facteur pour lequel plusieurs de ces corrélations sont assez élevées.

On dit qu'un facteur est **commun** aux groupes pour lesquels ces corrélations sont fortes (c'est-à-dire que la tendance qu'il traduit apparaît dans ces groupes) et qu'un facteur n'existe pas dans les groupes pour lesquels ces corrélations sont faibles.

Il peut arriver qu'un seul groupe ait une corrélation importante avec un facteur donné. Le facteur est alors une dimension **spécifique** du groupe.

Il peut arriver aussi qu'un groupe n'ait de corrélations élevées qu'avec des facteurs qui lui sont spécifiques. On en déduit alors l'absence de liaisons linéaires entre ce groupe et les autres. Il est généralement judicieux de recommencer alors l'analyse en supprimant (des groupes actifs tout au moins) ce groupe.

Pour décider si une corrélation est faible ou élevée, il n'y a pas de limite bien définie. Cela dépend du nombre d'individus et du nombre de groupes. Lorsqu'elles ne sont très proches ni de 1 ni de 0, on raisonne -comme toujours d'ailleurs- en termes de comparaison. On regarde pour un facteur donné si les corrélations associées à chaque groupe sont, ou non, du même ordre de grandeur ; on ordonne les groupes par corrélation décroissante. Inversement, pour un groupe donné, on examine et on ordonne les corrélations associées aux différents facteurs pour repérer les facteurs proches de directions de dispersion de ce groupe. On regarde aussi de quels autres groupes ces facteurs sont proches.

➤ Rapport [inertie inter / inertie totale]

Cet indice concerne l'ensemble des groupes. Proche de 1, il confirme le caractère « commun » d'un facteur, ce que des corrélations élevées ont déjà pu faire pressentir. Dans ce cas, les points représentant le même individu à travers les différents groupes sont globalement proches. La représentation superposée est alors utilisable pour un tel facteur.

➤ Coordonnées et aides à l'interprétation des groupes

Rappelons que la coordonnée du groupe $j$ le long de l'axe de rang $s$ s'interprète aussi comme la contribution absolue des variables du groupe $j$ au facteur $s$, c'est-à-dire en tant que mesure de liaison entre le groupe $j$ et le facteur $s$.

Le premier intérêt de ces coordonnées est de fournir une mesure de **l'importance** de la direction associée à un facteur donné dans les nuages $N_K^j$ des variables associés à chacun des groupes $j$. Naturellement, cette coordonnée du groupe $j$ le long de l'axe $s$ n'est intéressante que lorsque les coefficients de corrélation ont permis de conclure que le facteur étudié est une direction de dispersion qui apparaît dans le groupe $j$.

Il est fréquent que la coordonnée de chaque groupe le long du premier axe soit proche de 1 : les premiers facteurs de chaque groupe sont alors assez proches entre eux et le premier facteur global en est un compromis. Les valeurs des coordonnées suivantes sont à juger en référence aux diagrammes des valeurs propres des analyses séparées. Un groupe quasiment unidimensionnel ne peut avoir plusieurs coordonnées proches de 1 !

En tant que contribution, ces coordonnées s'utilisent tout à fait comme les contributions des individus ou des variables. On repère les groupes qui ont déterminé le plus les facteurs. On s'appuie sur eux au moment de l'interprétation des facteurs. Du fait de la pondération, la contribution des groupes au premier facteur est généralement assez équilibrée. Si elle ne l'est pas, on cherche à expliquer cette anomalie. Pour les facteurs suivants, toutes les situations peuvent se présenter.

C'est finalement en tant que coordonnées que ces valeurs sont le moins utilisées. On consulte certes les graphiques représentant les groupes, surtout quand ces derniers sont nombreux, mais plus en tant qu'illustration des interprétations précédentes qu'en tant que projection. La raison en est que la proximité entre deux points est une approche extrêmement pauvre de la ressemblance entre deux groupes, ce qui d'ailleurs se retrouve dans des qualités de représentation presque toujours très faibles.

➤ Coordonnées et aides à l'interprétation des axes des analyses séparées

Les coordonnées des axes des analyses séparées ne sont autres que les corrélations entre les facteurs des analyses séparées et ceux de l'analyse globale. Elles permettent de relier l'analyse globale aux analyses séparées en répondant aux questions suivantes : le facteur global d'ordre $s$ est-il proche d'un des facteurs de chaque groupe ? Sur quels facteurs globaux les premiers facteurs des groupes sont-ils bien représentés ?

➤ Conclusion

À ce niveau, on peut décider de continuer le dépouillement des résultats ou de refaire une analyse en modifiant le nombre de groupes actifs et/ou la répartition des variables dans les groupes, etc. Une telle décision peut intervenir lorsque l'on a trouvé soit un groupe indépendant des autres, soit une anomalie dans la structure des groupes, soit plusieurs ensembles de groupes assez distincts entre eux pour justifier des études séparées, soit des facteurs trop peu communs. L'attitude vis-à-vis des groupes est analogue à celle que l'on peut avoir vis-à-vis des variables en ACP ou en ACM. On peut les considérer en quelque sorte comme des **« supervariables »**.

### 11.5.4 Projections des variables et du nuage moyen des individus

Les projections, aides à l'interprétation et graphiques s'interprètent globalement comme en ACP ou en ACM. Notons cependant que, pour les variables qualitatives, les coordonnées des indicatrices sont les corrélations avec les facteurs et non pas, comme en ACM, les centres de gravité des classes. On consulte donc plutôt les coordonnées et aides à l'interprétation de ces centres de gravité qui apparaissent, dans les programmes, comme des individus supplémentaires (que la variable qualitative soit active ou supplémentaire). Rappelons que la somme des contributions des modalités d'une même variable (en tant que centres de gravité) à un facteur est égale au rapport de corrélation entre la variable et le facteur (*cf.* section b page 201).

Sur les graphiques des variables, on s'intéresse d'abord aux groupes les plus liés aux facteurs (au sens de la contribution) ; puis, à l'intérieur de chaque groupe, on cherche les variables les plus liées aux facteurs. L'interprétation se fait presque toujours à deux niveaux : on décèle une tendance dans un groupe puis on précise à travers quelles variables du groupe elle s'exprime.

### 11.5.5 Représentations superposées des individus et des modalités

Ces représentations n'ont d'intérêt que pour les facteurs communs à plusieurs groupes.

Ayant tout d'abord constaté la proximité globale des points représentant un même individu, on examine les individus qui s'écartent de ce schéma général et présentent des images différentes selon les groupes qui les décrivent. Dans ce type d'interprétation, on raisonne avec les groupes comme classiquement avec les variables. Ainsi, dans l'exemple des vins (section 7.1.5 page 156), on relève que tel vin est plus puissant du point de vue de l'olfaction au repos que de la gustation. À ce niveau, il est nécessaire de se référer fréquemment aux données initiales.

Lorsque plusieurs individus présentent le même type d'écart entre leurs représentations au travers des différents groupes, on recherche leur point commun (quelquefois, ce point commun est une zone du plan factoriel). S'il existe, ce point commun mérite toujours l'attention.

En tant que centres de gravité, les modalités participent à cette représentation superposée. Elles sont particulièrement précieuses lorsque les individus sont nombreux, voire même rendent inutile la représentation superposée des individus dans des données de type enquête.

### 11.5.6 Cas où tous les groupes comprennent les mêmes variables

Dans ce cas, on peut réaliser deux ACP (ou deux ACM) en juxtaposant soit les variables, soit les individus. L'AFM contient simultanément des résultats analogues à ceux de ces deux analyses. Notons toutefois que dans l'ACP juxtaposant les mêmes

individus caractérisés par chacun des groupes, les variables sont centrées sur cet ensemble répété d'individus, alors qu'en AFM elles le sont sur chaque groupe (comme dans l'ACP juxtaposant les variables).

## 11.6 QUELQUES TYPES DE FACTEURS

Il est communément admis que l'habileté dans l'interprétation des résultats d'Analyses des Données dépend beaucoup de l'expérience. Grossièrement, cette expérience est constituée d'un ensemble de cas auxquels l'analyste se réfère plus ou moins explicitement. Sans prétendre remplacer l'expérience, il est utile, à propos de l'interprétation, d'évoquer quelques situations typiques rencontrées en analyse factorielle.

Nous décrivons ci-après sept types de facteurs. Indiquons d'emblée qu'il ne s'agit pas d'une partition, un même facteur pouvant être présenté de plusieurs façons, mais de situations typiques auxquelles on peut se référer dans bon nombre de cas concrets. Pour chaque type, nous évoquons un exemple précis, très schématique, mais inspiré d'une analyse réelle. Enfin, nous abordons le problème, crucial dans la présentation des résultats d'une analyse, de l'attribution d'un nom à un facteur.

### 11.6.1 Facteur dû à quelques éléments aberrants

Le terme d'aberrant est discutable et discuté mais il est en passe d'être consacré par l'usage. Un élément est aberrant si, possédant quelques particularités remarquables, il se trouve très éloigné des autres. Il possède de ce fait une inertie importante qui peut influencer de façon prépondérante l'un des premiers axes.

Nous avons décrit quelques critères pour détecter de telles situations ainsi que la conduite à tenir le cas échéant (*cf.* section a). Les éléments aberrants peuvent être des individus en ACP ou en AFM, des lignes ou des colonnes en AFC, des individus ou des modalités en ACM ou en AFM.

**Exemple** : on a réalisé une ACM sur des données d'enquêtes. Un petit nombre d'individus n'a pas fourni de réponse à la plupart des questions. À eux seuls, ils engendrent l'un des premiers axes factoriels. Deux conduites sont possibles.

– Restreindre le champ de l'étude aux individus qui ont suffisamment rempli le questionnaire. On recommence l'analyse en neutralisant les individus possédant trop de non-réponses, c'est-à-dire en les mettant en supplémentaire voire en les éliminant. Lors de cette opération, on doit vérifier que les modalités concernées par cette élimination, c'est-à-dire principalement les non-réponses, conservent des effectifs suffisamment importants (l'ACM est sensible aux modalités d'effectif très faible).
– Conserver l'analyse si, d'une part, l'objectif de l'étude comprend la façon de répondre à un questionnaire et si, d'autre part, on peut considérer que les quelques

éléments aberrants n'ont fait que mettre en évidence un facteur non-réponse qui serait peut-être passé inaperçu sans eux, car alors affecté d'un rang élevé.

### 11.6.2 Facteur d'opposition

Lors de l'interprétation d'un facteur, on s'intéresse de manière privilégiée aux éléments possédant les coordonnées les plus extrêmes. Si l'opposition qui en résulte est claire, elle résume l'interprétation.

**Exemple** : les individus sont des stations sur lesquelles on a mesuré l'abondance de différentes plantes. Un facteur peut être interprété comme opposant les « prairies » aux « bois ». Cette opposition concerne aussi bien les stations (stations situées en zone de prairies, stations situées en zones boisées) que les plantes (plantes typiques des prairies, plantes typiques des bois).

### 11.6.3 Facteur mettant en évidence un groupe

Ce type de facteur s'apparente aux deux précédents. Il met en évidence un groupe d'éléments particuliers trop important pour être qualifié d'aberrant. Ce groupe s'oppose à l'ensemble des autres éléments qui occupent la zone centrale ; il en résulte une dissymétrie due aux effectifs différents des deux groupes.

**Exemple** : on a effectué différentes mesures biométriques sur des vaches. Un facteur met en évidence le caractère très particulier des animaux de la race charolaise.

### 11.6.4 Facteur associé à une partition

On est tenté de résumer un facteur par une partition lorsque la répartition des éléments le long de l'axe présente des discontinuités. On cherche alors à identifier les classes d'éléments ; si l'on trouve une identification claire, on privilégie cette partition dans la description du facteur.

**Exemple** : on effectue plusieurs mesures sur un ensemble de fromages issus de plusieurs procédés de fabrication. Un facteur peut séparer nettement les fromages selon le procédé dont ils sont issus : cette séparation constitue un élément essentiel de la description du facteur.

Une situation remarquable de facteurs associés à une partition se rencontre en AFC lorsqu'une valeur propre est égale à 1. Il est alors possible de partitionner les lignes et les colonnes du tableau en classes telles que, à l'intérieur d'une classe, les éléments possèdent exactement la même coordonnée sur le facteur associé (*cf.* Figure 3.9 page 78).

### 11.6.5 Facteur d'échelle

Il s'agit d'un facteur facilement et efficacement résumé par une variable quantitative, ou qualitative à modalités ordonnées, à laquelle il est très lié.

**Exemple** : des individus sont repérés par la possession de différents matériels d'équipement ménager. On peut trouver, selon un axe, les individus rangés selon le nombre d'appareils qu'ils possèdent : un tel axe peut être résumé par le terme « niveau d'équipement », qui évoque bien une échelle.

Cette situation se rencontre lorsque l'existence d'un facteur est supposé *a priori* et que l'Analyse Factorielle a précisément pour but de l'expliciter. Une situation typique est celle dans laquelle on cherche à mesurer l'intensité d'un phénomène unique au travers de plusieurs variables, généralement quantitatives ou qualitatives ordonnées. On souhaite obtenir une variable quantitative unique, compromis entre les variables initiales, afin de l'utiliser comme indicateur d'intensité du phénomène étudié.

**Exemple** : on veut construire un indice de précocité pour différentes variétés de maïs. Pour cela, on dispose d'un ensemble de mesures s'appuyant sur la date d'apparition de différents organes (feuilles, etc.). Chaque mesure constitue un indice de précocité ; ces indices sont très liés entre eux mais ne sont pas identiques. On réalise alors une ACP dont le premier facteur réalise un compromis entre les différentes variables initiales et s'interprète comme un indicateur synthétique de précocité.

### 11.6.6 L'effet taille en ACP

Il est classique, en ACP (*cf.* figure 1.7 page 17), d'observer que les coefficients de corrélation des variables actives avec un facteur (généralement le premier) sont tous positifs (cette situation se présente lorsque toutes les variables sont corrélées positivement entre elles, ce qui peut se lire directement sur la matrice des corrélations). Le vocable « effet taille » fait référence à des données biométriques, de type mensurations, sur des individus : avec ce type de données, le facteur précité classe les individus depuis ceux qui présentent les plus faibles valeurs pour l'ensemble des variables (c'est-à-dire les petits) jusqu'à ceux qui présentent les plus fortes valeurs pour l'ensemble des variables (c'est-à-dire les grands).

### 11.6.7 L'effet Guttman en AFC ou ACM

En AFC, lorsqu'un facteur d'échelle est très fort, il influence plusieurs axes selon la propriété suivante : le facteur de rang $s$ est une fonction polynôme de degré $s$ du premier. Ce phénomène se détecte facilement à partir du plan factoriel (1,2) sur lequel le nuage des lignes et des colonnes présente l'allure d'une parabole (*cf.* section 10.3.2 page 231). Le fait d'identifier un effet Guttman ne modifie pas sensiblement l'interprétation des deux premiers axes d'une AFC (le premier axe est un facteur d'échelle, le second un facteur d'opposition entre les situations extrêmes et les situations moyennes).

En revanche, cela conduit à négliger les facteurs suivants qui sont des fonctions polynômes du premier. L'effet Guttman est plus ou moins net selon l'intensité, dans les données, du phénomène qu'il met en évidence. Si le premier plan factoriel fait apparaître un nuage de points dont la forme parabolique est floue, il est possible que l'influence du premier facteur ne se fasse sentir que sur quelques axes seulement : il est alors possible de trouver des facteurs de rang moyennement élevé (*e.g.* 3, 4 ou 5) s'interprétant indépendamment du premier (*cf.* section b page 234). Il est donc prudent de s'assurer d'un effet Guttman au delà des deux premiers facteurs.

En ACM, on observe ce phénomène surtout lorsque les modalités de chacune des variables sont ordonnées *a priori*. Par exemple, si des variables qualitatives proviennent du recodage d'un ensemble de variables quantitatives dont l'ACP produit comme premier facteur un effet taille, l'ACM de ces variables qualitatives conduit presque automatiquement à un effet Guttman, le premier axe supportant la même interprétation globale dans les deux analyses. En ce sens, l'ACM est susceptible de mettre en évidence un effet taille ; cet effet se traduit par plusieurs facteurs (au maximum $r - 1$ si les variables possèdent chacune $r$ modalités), alors qu'il se traduit par un facteur unique en ACP.

### 11.6.8 Le nom d'un facteur

L'interprétation d'une analyse factorielle est une opération complexe en ce sens qu'elle met en jeu un grand nombre d'éléments d'origines variées. Elle comprend en particulier une étape qui consiste à donner un nom aux facteurs. Cette étape n'est pas forcément difficile et n'intervient pas nécessairement en fin de processus. Elle est néanmoins la plus voyante car son aspect synthétique fait qu'elle est privilégiée dans la présentation des résultats d'une analyse dont elle est souvent l'élément le plus mémorisable. L'objectif est ici d'illustrer, autour de l'affectation d'un nom à un facteur, d'une part la notion de type de facteurs et sa relativité, et d'autre part les différents contextes de l'interprétation. Nous reprenons ci-après, de façon simplifiée, l'interprétation du deuxième facteur d'une ACM réalisée à partir de l'enquête *Ouest-France* (*cf.* section 6.4 page 132).

Le premier contexte, celui des variables actives met en évidence une opposition entre les modalités *lecture* et les modalités *non-lecture* des rubriques d'information générale. L'interprétation se situe d'abord dans ce premier contexte dans lequel le premier facteur peut être résumé par l'opposition *lecture*/*non-lecture* de ces rubriques.

Le deuxième contexte, celui des variables supplémentaires, fait apparaître une liaison entre ce facteur et les CSP. Schématiquement, les agriculteurs et les ouvriers lisent peu ces rubriques, les étudiants et cadres supérieurs les lisent beaucoup, les cadres moyens occupant une position intermédiaire entre ces deux extrêmes. Une deuxième façon de présenter ce facteur, focalisée sur cette partition, est « facteur social ».

Le troisième contexte, celui de nos connaissances générales sur la société, indique que le facteur place les CSP selon un statut social croissant. Le facteur apparaît alors plutôt comme un facteur d'échelle : le statut social.

Cet exemple illustre, de façon schématique mais cependant réaliste, quelques problèmes généraux apparaissant lors d'une interprétation. La notion de type de facteur envisagé précédemment fournit des points de repère commodes et non une grille contraignante : le même facteur est appréhendé d'abord en termes d'opposition, puis en termes de partition et enfin en termes d'échelle. L'intervention successive des différents contextes fait apparaître une coupure ; le troisième contexte introduit les connotations les plus synthétiques et des éléments de validation (on eût réexaminé d'un œil suspicieux le facteur précédent s'il avait opposé d'une part agriculteurs et cadres supérieurs à, d'autre part, étudiants et cadres moyens).

Chapitre **12**

# Fiches techniques

## 12.1 FICHE 1 : MOYENNE ET BARYCENTRE, VARIANCE ET INERTIE

### 12.1.1 Cas d'une variable

Une variable $x$ définie sur un ensemble $I$ d'individus se représente par un nuage de points sur un axe. L'individu $i$ est représenté par le point d'abscisse égale à la valeur $x_i$ prise par la variable $x$ pour l'individu $i$.

#### *a) Moyenne et barycentre*

Si l'importance des individus est la même pour tous, la **moyenne** de la variable $x$, notée $\bar{x}$, est égale à :

$$\bar{x} = \frac{1}{I} \sum_i x_i$$

Plus généralement, si l'individu a un poids $p_i$ (par exemple si les individus représentent des populations d'effectifs inégaux), la moyenne $\bar{x}$ s'écrit :

$$\bar{x} = \sum_i p_i x_i / \sum_i p_i$$

Souvent les poids sont tels que $\sum_i p_i$=1 ce qui allège l'écriture : $\bar{x} = \sum_i p_i x_i$

**Sur l'axe de représentation du nuage, le point d'abscisse $\bar{x}$ est le barycentre des points $x_i$ muni des poids $p_i$. Ce barycentre est la traduction géométrique de la notion statistique de moyenne.**

En retirant à chaque $x_i$ la moyenne $\bar{x}$, on obtient une variable centrée. En passant de $x$ à $x - \bar{x}$ on effectue une translation du nuage sur l'axe (ou une translation de l'origine de l'axe) qui fait coïncider son barycentre avec l'origine.

### *b) Variance et inertie*

Si l'importance des individus est la même pour tous, la **variance** d'une variable $x$, notée $s_x^2$, est égale à :

$$s_x^2 = \frac{1}{I} \sum_i (x_i - \bar{x})^2$$

Si l'individu $i$ a un poids $p_i$ elle s'écrit :

$$s_x^2 = \sum_i p_i (x_i - \bar{x})^2 / \sum_i p_i$$

Lorsque les poids sont tels que $\sum_i p_i = 1$ on a : $s_x^2 = \sum_i p_i (x_i - \bar{x})^2$.

La variance mesure la dispersion des valeurs autour de la moyenne. Le fait de considérer les carrés des écarts et non les valeurs absolues des écarts facilite les calculs et permet des décompositions suivant le théorème de Pythagore et celui de Huygens rappelé plus loin. L'**écart-type** $s_x$ est la racine carrée de la variance.

**La notion statistique de variance correspond à la notion mécanique d'inertie d'un nuage de points par rapport à son barycentre.**

En effet, l'inertie d'un point $i$ de poids $p_i$ par rapport à un point A de coordonnée $x_a$ est, par définition, le produit du poids de $i$ par le carré de sa distance à A soit : $p_i(x_i - x_a)^2$.

L'inertie d'un nuage de points est la somme des inerties des points du nuage. L'inertie d'un nuage de points représenté sur un axe, par rapport au point G d'abscisse $\bar{x}$, est égale à $\sum_i p_i (x_i - \bar{x})^2$ ; on retrouve la variance lorsque $\sum_i p_i = 1$.

Quand on divise chaque valeur $x_i - \bar{x}$ de la variable centrée par son écart-type $s_x$, on obtient une variable de variance 1 appelée **variable centrée-réduite**.

**La transformation géométrique qui permet de passer de $x - \bar{x}$ à $(x - \bar{x})/s_x$ est une homothétie de centre G et de rapport égal à $1/s_x$.**

### *c) Théorème de Huygens*

La forme la plus simple du théorème de Huygens est la relation entre l'inertie d'un nuage par rapport à un point quelconque Z d'abscisse $z$ et son inertie par rapport à G. La première est égale à la seconde augmentée de l'inertie, par rapport à Z, de G affecté du poids total du nuage :

$$\sum_i p_i (x_i - z)^2 = \sum_i p_i (x_i - \bar{x})^2 + \left( \sum_i p_i \right) (\bar{x} - z)^2$$

En appliquant cette relation à $J$ sous-nuages, on obtient la forme décrite ci-après sous laquelle le théorème de Huygens est rencontré le plus souvent en statistique.

L'inertie d'un nuage de points dans lequel on distingue $J$ sous-nuages est la somme des inerties de ces sous-nuages par rapport à leur barycentre (inertie intra) augmentée de l'inertie du nuage des $J$ barycentres chacun affecté du poids total du sous-nuage qu'il représente (inertie inter). Ceci s'écrit, en notant $I_j$ le $j^{\text{e}}$ sous-nuage, $\bar{x}_j$ son barycentre et $p_j$ son poids ( $p_j = \sum_i p_i$ pour $i \in I_j$) (*cf.* illustration **Figure 12.1**) :

$$\sum_{i \in I} p_i (x_i - z)^2 = \sum_j \sum_{i \in I_j} p_i (x_i - \bar{x}_j)^2 + \sum_j p_j (\bar{x}_j - z)^2$$

C'est la forme « mécanique » de la décomposition classique de la variance :

$$\text{variance totale} = \text{variance inter} + \sum_j p_i \text{ variance intra } I_j$$

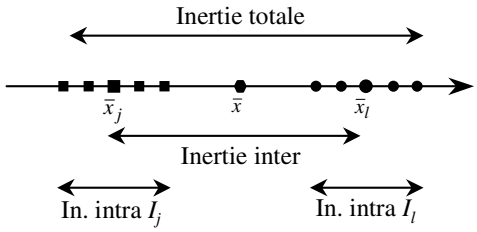



**Figure 12.1** Théorème de Huygens pour une variable. 8 points sont répartis en 2 sous-nuages de 4 points : $I_j$ (carrés) et $I_l$ (disques).

## 12.1.2 Cas de deux variables

Ces propriétés se généralisent à un tableau de données comportant 2 variables $x$ et $y$. L'ensemble des valeurs des 2 variables se représente par un nuage dans un plan rapporté à deux axes orthogonaux correspondant respectivement aux deux variables. Un individu $i$ est représenté par un point dont les 2 coordonnées sont ses valeurs $x_i$ et $y_i$.

### *a) Centrage et réduction*

Le point $G$ de coordonnées $(\bar{x}, \bar{y})$ est le barycentre des points du nuage munis des poids $p_i$. Quand on retire à chaque valeur $x_i$ la moyenne $\bar{x}$ et à chaque valeur $y_i$ la moyenne $\bar{y}$, on obtient un tableau centré. La transformation géométrique qui permet

de passer du nuage associé au tableau initial au nuage associé au tableau centré est une translation qui fait coïncider l'origine O et le barycentre G.

Quand on divise les valeurs $x_i - \bar{x}$ par $s_x$ et les valeurs $y_i - \bar{y}$ par $s_y$ , on obtient un tableau centré-réduit. La transformation géométrique qui permet de passer du nuage centré au nuage centré-réduit est la composition de deux homothéties de centre G (la première, de rapport $1/s_x$ dans la direction de $x$, la seconde, de rapport $1/s_y$ dans la direction de $y$). Une autre façon de voir cette transformation est de considérer que l'on adopte $s_x$ et $s_y$ comme unités de mesure (*cf.* **Figure 12.2**).

Un nuage centré-réduit possède, en projection sur chaque axe, une inertie égale à 1.

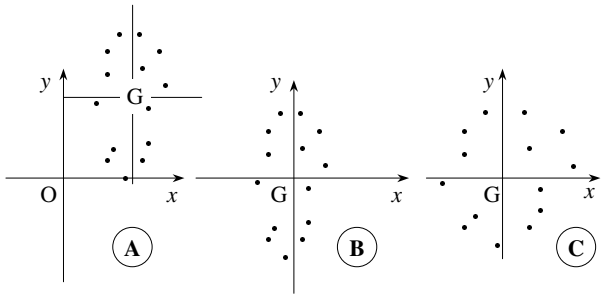



**Figure 12.2** Nuage associé aux mêmes 2 variables $x$ et $y$ brutes (A), centrées (B) et centrées-réduites (C).

### *b) Théorème de Huygens*

Le carré de la distance d'un point $i$ à l'origine vaut : $\|Oi\|^2 = x_i^2 + y_i^2$. On en déduit que :

$$\text{inertie de } i = p_i \, \|Oi\|^2 = p_i x_i^2 + p_i y_i^2$$

D'où, pour le nuage des points $i$ :

$$\text{inertie totale } = \sum_i p_i \, \|Oi\|^2 = \sum_i p_i x_i^2 + \sum_i p_i y_i^2$$

L'inertie du nuage se décompose donc suivant les deux axes : elle est la somme des inerties de ses deux projections suivant les deux directions orthogonales. Si les variables sont centrées, elle est donc égale à la somme des variances des deux variables. Si les variables sont centrées-réduites, l'inertie du nuage vaut 1 dans chaque direction et vaut donc 2 dans le plan.

Le théorème de Huygens se généralise sans difficulté au cas de deux variables puisque l'inertie d'un nuage se décompose sur chaque axe suivant le théorème de Pythagore (*cf.* **Figure 12.3**).

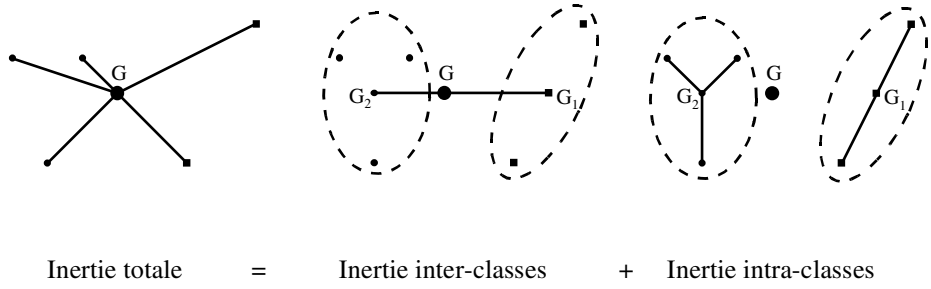



**Figure 12.3** Théorème de Huygens dans le plan.

### 12.1.3 Cas d'un nombre quelconque *K* de variables

Ces propriétés se généralisent à un nombre quelconque $K$ de variables. À chaque individu $i$, on associe le point de $R^K$ dont la $k^{\text{e}}$ coordonnée est la valeur de la $k^{\text{e}}$ variable pour $i$. Le point dont la $k^{\text{e}}$ coordonnée est la moyenne de la variable $k$ (quel que soit $k$) est le barycentre G du nuage. Centrer le tableau de données, c'est-à-dire retrancher à chaque variable sa moyenne, consiste encore à prendre le point G comme origine des axes.

Réduire le tableau de données, c'est appliquer $K$ homothéties successives dans les directions des axes pour avoir une inertie égale à 1 dans chacune de ces $K$ directions.

Le théorème de Huygens se généralise sans difficulté au cas de plusieurs variables.

## 12.2 FICHE 2 : REPRÉSENTATION DES VARIABLES DANS $R^I$

### 12.2.1 Espace et métrique

Une variable $x$ définie sur un ensemble $I$ d'individus est représentée par un vecteur de $R^I$ dont les $I$ composantes sont égales aux valeurs $x_i$ prises par la variable $x$ pour l'individu $i$.

$$x = (x_1, \ldots, x_i, \ldots, x_I)$$

Sur l'espace $R^I$ est défini un **produit scalaire** (*cf.* Fiche 3). Si les poids des individus sont tous égaux, le produit scalaire entre deux vecteurs $x$ et $y$ s'écrit :

$$\langle x, y \rangle = \frac{1}{I} \sum_i x_i y_i$$

Plus généralement, si les individus ont des poids $p_i$ tels que $\sum_i p_i = 1$ :

$$\langle x, y \rangle = \sum_i p_i x_i y_i$$

Soit $u$ le vecteur colinéaire à la première bissectrice dont toutes les composantes sont égales à 1. Ce vecteur a pour norme 1 :

$$\begin{aligned} u &= (u_1, \ldots, u_i, \ldots, u_I) = (1, \ldots, 1, \ldots, 1) \\ \|u\|^2 &= \langle u, u \rangle = \sum_i p_i u_i^2 = \sum_i p_i = 1 \end{aligned}$$

### 12.2.2 Centrage

La **moyenne** $\bar{x}$ d'une variable $x$ est égale à la coordonnée de la projection de $x$ sur $u$ :

$$\begin{aligned} \bar{x} &= \sum_i p_i x_i = \sum_i p_i x_i u_i = \langle x, u \rangle \\ \bar{x}u &= \text{projection orthogonale de } x \text{ sur } u \end{aligned}$$

Une variable **centrée** est représentée par un vecteur orthogonal à $u$ car :

$$\sum_i p_i x_i = 0 \text{ équivaut à } \langle x, u \rangle = 0$$

**Centrer une variable** c'est considérer, au lieu de $x$, la variable centrée de composantes $x_i - \bar{x}$. Cette variable centrée est représentée par le vecteur $x - \bar{x}u$ :

$$\begin{aligned} x - \bar{x}u &= (x_1 - \bar{x}, \ldots, x_i - \bar{x}, \ldots, x_I - \bar{x}) \\ &= x - [\text{projection orthogonale de } x \text{ sur } u] \end{aligned}$$

Le vecteur $x - \bar{x}u$ (orthogonal à $u$) est la projection de $x$ sur l'hyperplan orthogonal à $u$. Centrer $x$ revient donc à considérer sa projection sur l'hyperplan orthogonal à $u$ (*cf.* **Figure 12.4**).

### 12.2.3 Réduction

La **variance** d'une variable $x$ est égale au carré de la norme du vecteur représentant la variable centrée ; son écart-type $s_x$ est égal à la norme de ce vecteur :

$$\text{variance de } x = \sum_i p_i (x_i - \bar{x})^2 = \|x - \bar{x}u\|^2 = s_x^2$$

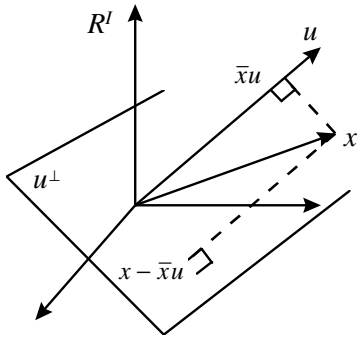



**Figure 12.4** Interprétation géométrique du centrage dans $R^I$. $u$ : vecteur unitaire constant. La projection de $x$ sur $u$ est le vecteur constant dont chaque coordonnée est égale à $\bar{x}$. La variable centrée $x - \bar{x}u$ est la projection de $x$ sur l'hyperplan $u^\perp$ orthogonal à $u$.

Une variable **centrée-réduite** est représentée par un vecteur de norme 1 orthogonal à $u$.

**Centrer et réduire** une variable c'est considérer, au lieu de $x$, la variable centrée et réduite de composantes $(x_i - \bar{x})/s_x$.

Réduire une variable centrée consiste à la diviser par son écart type ; le vecteur représentant la variable est alors divisé par sa norme.

### 12.2.4 Coefficient de corrélation

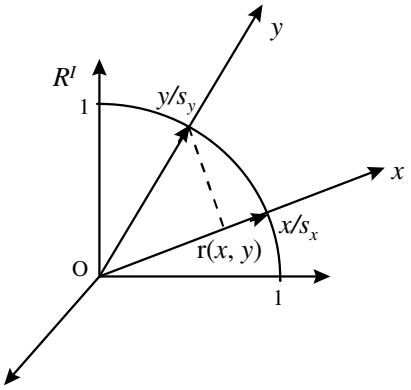
Le **coefficient de corrélation** entre deux variables $x$ et $y$, noté $\mathrm{r}(x, y)$, est égal au cosinus de l'angle entre les vecteurs représentant les variables centrées, c'est-à-dire au produit scalaire entre les vecteurs représentant les variables centrées-réduites :

$$\begin{aligned} \text{corrélation}(x, y) &= \mathrm{r}(x, y) = \frac{\sum_i p_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{\langle x - \bar{x}u, y - \bar{y}u \rangle}{\|x - \bar{x}u\| \|y - \bar{y}u\|} \\ &= \frac{\langle x, y \rangle}{\|x\| \|y\|} \quad \text{si } x \text{ et } y \text{ sont centrées} \\ &= \langle x, y \rangle \quad \text{si } x \text{ et } y \text{ sont centrées et réduites} \end{aligned}$$

Plus la corrélation entre les variables est élevée, plus l'angle entre les vecteurs est faible. Si la corrélation entre $x$ et $y$ est nulle, les vecteurs sont orthogonaux ; si elle est égale à 1 ou -1, les vecteurs sont colinéaires.

## 12.3 FICHE 3 : DISTANCE, NORME ET PRODUIT SCALAIRE

Cette fiche précise les notions de distance, norme et produit scalaire ainsi que les relations entre ces structures. Nous donnons d'abord une vision générale de l'ensemble des

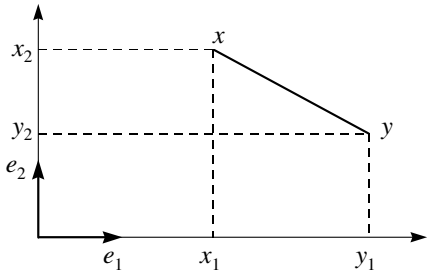



$x$ et $y$ sont deux variables centrées. Leurs normes sont égales à leurs écarts-types $s_x$ et $s_y$. Le produit scalaire entre les deux variables normées est égal à leur coefficient de corrélation $r(x, y)$.

**Figure 12.5** Réduction de variables centrées et coefficient de corrélation dans $R^I$.

termes techniques. Les définitions mathématiques généralisent les notions habituelles du plan (et de l'espace $R^3$) auxquelles nous nous référons systématiquement.

### 12.3.1 Espace vectoriel et espace euclidien




$\|e_1\| = 1 \quad \|e_2\| = 1 \quad \langle e_1, e_2 \rangle = 0$
$x = x_1 e_1 + x_2 e_2$
Le carré de la distance entre les points $x$ et $y$ est :

$$\begin{aligned} d^2(x, y) &= (x_1 - y_1)^2 + (x_2 - y_2)^2 \\ &= \|x - y\|^2 \\ &= \langle x - y, x - y \rangle \end{aligned}$$

**Figure 12.6** Distance, norme et produit scalaire dans le plan. Les vecteurs $x$ et $y$ se décomposent sur la base $e_1$, $e_2$.

La notion la plus générale est la notion de distance qui peut être définie sur un ensemble quelconque. Sur un espace vectoriel, une distance peut dériver d'une norme, on parle alors d'espace normé. Une norme peut elle-même dériver d'un produit scalaire. Une norme qui dérive d'un produit scalaire est une norme euclidienne et la distance qui en découle est une distance euclidienne. On appelle espace euclidien un espace

vectoriel réel de dimension finie sur lequel est défini un produit scalaire. Dans la suite, nous parlerons uniquement de l'espace $R^n$, seul espace utilisé en analyse factorielle ($n$ désigne la dimension de l'espace). Nous parlons aussi de métrique euclidienne pour désigner la structure définie sur $R^n$ par un produit scalaire.

### 12.3.2 Distance

Une distance sur un ensemble $E$ est une application du produit de $E$ par lui-même dans $R^+$ : à tout couple de points $(x, y)$ est associé un nombre positif, la distance entre $x$ et $y$ notée $\mathrm{d}(x, y)$.

Cette application vérifie certaines propriétés quels que soient $x$ et $y$ appartenant à $E$ :

$$\begin{aligned} \mathrm{d}(x, y) &= 0 \text{ si et seulement si } x = y \\ \mathrm{d}(x, y) &= \mathrm{d}(y, x) \\ \mathrm{d}(x, y) &\leqslant \mathrm{d}(x, z) + \mathrm{d}(z, y) \quad \text{(inégalité triangulaire)} \end{aligned}$$

La distance usuelle (ou canonique) de $R^2$ s'écrit, en notant $x_i$ et $y_i$ les coordonnées des points $x$ et $y$ sur la base usuelle (ou canonique) :

$$\mathrm{d}^2(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2$$

Plus généralement la distance usuelle de $R^n$ s'écrit :

$$d^2(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$$

### 12.3.3 Norme

Une norme sur $R^n$ est une application de $R^n$ dans $R^+$ : à tout vecteur $x$ est associé un nombre positif, la norme de $x$, notée $\|x\|$.

Cette application vérifie certaines propriétés (O = origine des axes) :

$$\begin{aligned} \|x\| &= 0 \text{ si et seulement si } x = \mathrm{O} \\ \|ax\| &= |a|\,\|x\| \text{ pour tout } x \text{ de } R^n \text{ et tout } a \text{ de } R \\ \|x + y\| &\leqslant \|x\| + \|y\| \text{ pour tous } x \text{ et } y \text{ de } R^n \end{aligned}$$

Toute norme induit une distance par la relation : $\mathrm{d}(x, y) = \|x - y\|$. Ainsi, lorsqu'une distance dérive d'une norme :

1. la distance d'un point à l'origine O des axes est la norme (ou longueur) du vecteur qui le joint à O ;
2. la distance entre deux points $x$ et $y$ est la longueur du vecteur qui joint ces deux points (*cf.* **Figure 12.3**).

Une distance qui dérive d'une norme a des propriétés spécifiques.

La distance usuelle de $R^2$ dérive de la norme : $\|x\|^2 = x_1^2 + x_2^2$

Plus généralement, la distance usuelle de $R^n$ dérive de la norme :

$$\|x\|^2 = \sum_{i=1}^{n} (x_i)^2$$

### 12.3.4 Produit scalaire

Un produit scalaire sur un espace vectoriel $E$ est une application du produit de $E$ par lui-même dans $R$ : à tout couple de vecteurs $(x, y)$ est associé un nombre, le produit scalaire entre $x$ et $y$, noté $\langle x, y\rangle$. Cette application vérifie certaines propriétés. Ainsi, quels que soient les nombres $a$ et $b$ et les vecteurs $x$, $y$ et $z$, on a :

$$\begin{aligned}
\langle x, x\rangle &= 0 \text{ si et seulement si } x = 0 \\
\langle x, y\rangle &= \langle y, x\rangle \text{(symétrie)} \\
\langle ax + by, z\rangle &= a\langle x, z\rangle + b\langle y, z\rangle \quad \text{(bilinéarité)} \\
\langle z, ax + by\rangle &= a\langle z, x\rangle + b\langle z, y\rangle \quad \text{(bilinéarité)}
\end{aligned}$$

Un produit scalaire induit une norme par la relation :

$$\|x\|^2 = \langle x, x\rangle$$

Le produit scalaire usuel (ou canonique) de $R^2$ s'écrit, en notant $x_i$ et $y_i$ les composantes des vecteurs $x$ et $y$ sur la base usuelle (ou canonique) :

$$\langle x, y\rangle = x_1 y_1 + x_2 y_2$$

Plus généralement le produit scalaire usuel de $R^n$ s'écrit :

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

La norme usuelle (et donc la distance usuelle) de $R^n$ dérive de ce produit scalaire.

### 12.3.5 Angles

Un produit scalaire induit, en plus de la notion de norme, la notion d'angle. L'angle $\theta$ entre deux vecteurs $x$ et $y$ est défini par son cosinus qui, par définition, est égal au produit scalaire de ces deux vecteurs divisé par le produit de leurs normes :

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|\,\|y\|}$$

Le cosinus est compris entre -1 et 1. S'il vaut 1, les vecteurs sont colinéaires et de même sens ; s'il vaut -1, ils sont colinéaires de sens opposé.

### 12.3.6 Orthogonalité, théorème de Pythagore et projection orthogonale

Du produit scalaire on déduit une notion d'orthogonalité : deux vecteurs $x$ et $y$ sont orthogonaux si et seulement si leur produit scalaire est nul. Un vecteur $x$ est orthogonal à un sous-espace s'il est orthogonal à tous les vecteurs de ce sous-espace.

Le théorème de Pythagore s'applique à tout produit scalaire : si un vecteur $z$ est la somme de deux vecteurs orthogonaux $x$ et $y$, le carré de sa longueur est la somme des carrés des longueurs de $x$ et de $y$.

On définit la projection orthogonale d'un vecteur $x$ sur un axe. Si $u$ est un vecteur unitaire de cet axe, la projection $y$ de $x$ est le vecteur colinéaire à l'axe de coordonnée $\langle x, u\rangle$ et le vecteur $x - y$ est orthogonal à l'axe.

$$\text{projection de } x \text{ sur } u = \langle x, u\rangle u$$

On définit aussi la projection orthogonale sur un sous-espace $E$ : $y$ est la projection de $x$ sur $E$ si le vecteur $x - y$ est orthogonal à $E$.

### 12.3.7 Expression matricielle

Soit la base canonique de $R^n$ et notons $m_{ij}$ le produit scalaire entre les vecteurs $e_i$ et $e_j$ de la base. Du fait de la bilinéarité, le produit scalaire entre $x$ et $y$ s'écrit :

$$\langle x, y\rangle = \langle \sum_i x_i e_i, \sum_j x_j e_j\rangle = \sum_i \sum_j x_i y_j \langle e_i, e_j\rangle = \sum_i \sum_j m_{ij} x_i y_j$$

soit, matriciellement, en notant $M$ la matrice de terme général $m_{ij}$ et $x'$ le transposé de $x$ :

$$\langle x, y\rangle = x'My = y'Mx$$

La norme et la distance induites par le produit scalaire s'écrivent :

$$\|x\|^2 = \langle x, x\rangle = x'Mx$$

$$\mathrm{d}^2(x, y) = \|x - y\|^2 = (x - y)'M(x - y)$$

La structure de la matrice $M$ est souvent utilisée pour qualifier une métrique. Ainsi, on parle de métrique diagonale si $M$ est diagonale. De même, la distance euclidienne usuelle étant associée à la matrice identité, on la nomme souvent « métrique identité ».

### 12.3.8 Produit scalaire et base orthonormée

Une base d'un espace euclidien (ou d'un sous-espace) est orthogonale pour le produit scalaire si les vecteurs de cette base sont orthogonaux deux à deux. Si de plus ces vecteurs ont pour longueur 1, cette base est orthonormée.

Dans une base orthonormée (pour le produit scalaire considéré), le produit scalaire s'exprime sous la forme canonique.

Si la base est orthogonale pour le produit scalaire considéré, les termes non diagonaux de la matrice $M$ sont nuls ; cette matrice est diagonale et le produit scalaire se réduit à :

$$\langle x, y\rangle = \sum_i m_{ii} x_i y_i$$

C'est le cas des métriques utilisées en analyse factorielle. La différence entre ces métriques "diagonales" et la métrique habituelle est que chaque vecteur de base a un "poids", qui s'exprime en particulier dans la distance :

$$\mathrm{d}^2(x, y) = \sum_i m_{ii}(x_i - y_i)^2$$

Si le produit scalaire n'est pas le produit scalaire usuel, la distance induite ne correspond pas à la vision habituelle. Pour obtenir une représentation des distances directement perceptible à l'oeil, il faut se ramener au produit usuel. Pour cela il suffit d'exprimer et de représenter les points dans une base orthonormée pour le produit scalaire considéré. C'est ce qui est fait en analyse factorielle.

Dans le cas d'une métrique diagonale, une base orthonormée se déduit de la base canonique en divisant les vecteurs de base par leur norme, ce qui revient à multiplier les coordonnées correspondantes par cette norme.

Précisons cela en prenant l'exemple de $R^2$ muni de la métrique diagonale valant 4 pour le premier vecteur de base et 1/9 pour le second. Pour travailler avec la métrique habituelle, il suffit de faire la transformation qui à tout point $x$ de coordonnées $(x_1, x_2)$ associe le point de coordonnées $(2x_1, x_2/3)$. Le poids de la première coordonnée étant supérieur à 1, cette coordonnée est dilatée tandis que la seconde est contractée.

Soit {A, B, C, D} un nuage de 4 points représentés dans $R^2$ muni de la base $\{u_1, u_2\}$ et de la métrique diagonale $\{4, 1/9\}$ (*cf.* **Figure 12.7.A**). La matrice des distances inter-individuelles est donnée figure 12.7. Dans cet espace, la base $\{e_1, e_2\}$ est orthonormée.

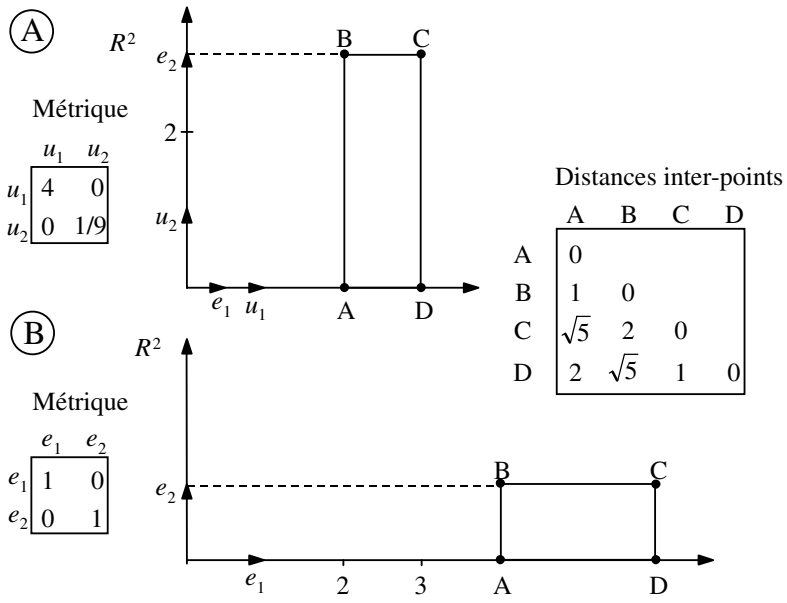



**Figure 12.7** Nuage de 4 points {A, B, C, D} représenté dans le plan $R^2$ muni d'une métrique diagonale (A) et dans une base orthonormée de ce même plan (B).

La **Figure 12.7.B** représente ce même ensemble de points dans la base $\{e_1, e_2\}$. Dans cette base, la distance euclidienne usuelle conduit à la même matrice des distances inter-individuelles que précédemment.

Cette transformation permet donc d'analyser avec les règles usuelles un nuage de points évoluant initialement dans un espace muni d'une métrique autre que l'identité. C'est bien ce qui est fait par exemple en AFC.

# Index systématique

d'éléments (lignes ou colonnes d'un tableau de données) qui regroupent dans une même classe des éléments qui se ressemblent ; les méthodes les plus utilisées s'appuient sur les mêmes représentations géométriques que les analyses factorielles (nuages définis comme en ACP, AFC, ACM ou AFM suivant le type de tableau). Les algorithmes tendent à obtenir des partitions qui maximisent le quotient de l'inertie inter-classe par l'inertie totale (2.3.3). L'optimum global est généralement inaccessible. Les méthodes de partitionnement construisent directement une partition (exemple : **agrégation autour des centres mobiles**) ; les méthodes hiérarchiques (exemple **CAH**) construisent un arbre hiérarchique, outil commode pour raisonner le choix d'une partition.

**codage** = traduction numérique d'un ensemble de données en vue d'un traitement statistique particulier ; codage d'une variable numérique en variable qualitative (4.5) ; codage condensé de variables qualitatives (4.1.2) ; codage disjonctif complet (4.1.3).

**composante principale** = **facteur** sur les individus en ACP (1.6).

**contribution à l'inertie** : contribution d'un élément à l'inertie d'un nuage (1.9.1.3) = [inertie de cet élément / inertie du nuage] ; contribution d'un élément à l'inertie d'un axe ou d'un facteur = contribution de cet élément à l'inertie du nuage projeté sur l'axe ; les points dont la contribution à l'inertie est la plus importante peuvent déterminer le facteur : en ACP (11.2.2.1 et 1.9.2), en AFC (11.3.2) ; contribution des modalités d'une variable qualitative : en ACM (4.3.6 et 11.4.3), en AFM (8.6.2 et 11.5.4) ; contribution d'un groupe de variables (7.1.3, 8.3.2 et 9.2.2.2) ; contribution d'un sous-tableau (10.4.3) ; on parle quelquefois, abusivement, de la contribution d'un élément supplémentaire en lui appliquant le même rapport d'inertie ; en AFM, contribution d'une composante principale d'un groupe (8.3.3 et 9.2.4) ; contribution d'un facteur à l'inertie d'un nuage = quotient entre l'inertie de ce facteur et l'inertie du nuage = **qualité de représentation** du nuage sur l'axe (1.9) = pourcentage d'inertie extrait par l'axe.

**corrélation** = coefficient de corrélation linéaire. Définition (1.1) ; limite d'interprétation (4.5.1) ; représentation des corrélations par des cosinus (1.4 et 12.2.4) ; en ACP normée, les coordonnées des projections des variables sur les axes sont égales à leur corrélation avec les composantes principales (1.6) ; coefficient de corrélation multiple (8.3.4.2) ; rapport de corrélation (4.3.6).

**dimension, dimensionalité** : la dimension d'un espace est le nombre de vecteurs orthogonaux 2 à 2 que peut contenir cet espace ; la dimension ou dimensionalité d'un nuage est la plus petite dimension d'un espace dans lequel on peut représenter ce nuage ; si l'on souhaite tenir compte du fait que l'inertie d'un nuage est très inégalement répartie selon les dimensions, on peut calculer un indicateur de dimensionalité (8.4.2).

# Bibliographie


[1] BENZECRI J.-P. et coll. (1973) *L'analyse des données*. Tome 1 : La taxinomie. Tome2 : L'analyse des correspondances. Dunod.

[2] BENZECRI J.-P. et F. (1980) *Pratique de l'analyse des données*. Tome 1 : analyse des correspondances, exposé élémentaire. Dunod.

[3] BENZECRI J.-P, BASTIN Ch., BOURGARIT Ch., CAZES P. (1980) *Pratique de l'analyse des données*. Tome 2 : Abrégé théorique, étude de cas modèles. Dunod.

[4] BENZECRI J.-P. et coll. (1984) *Pratique de l'analyse des données*. Tome 3 : Linguistique et Lexicologie. Dunod.

[5] BENZECRI J.-P. (1972) *La place de l'a priori*. in Encyclopédia Universalis.

[6] BENZECRI J.-P.et coll. (1984) *Pratique de l'analyse des données en économie*. Dunod.

[7] BOUROCHE J.-M. et SAPORTA G. (1980) *L'analyse des données*. PUF Collection Que Sais-je ?

[8] CAILLEZ F. et PAGES J.-P. (1976) *Introduction à l'analyse des données*. Smash.

[9] CEHESSAT R. (1981) *Exercices commentés de statistique et d'informatique appliquées*. 2 édition. Dunod.

[10] GOVAERT G. (1989) *Classification automatique des données*. Dunod.

[11] ESCOFIER B. (2003) *Analyse des correspondances*. Presses Universitaires de Rennes.

[12] ESCOFIER B. et PAGES J. (1997) *Initiation aux traitements statistiques : méthodes, méthodologie*. Presses Universitaires de Rennes.

[13] FENELON J.-P. (1982) *Qu'est-ce que l'analyse des données ?* Lefonen.

[14] GERI (1996) *Analyse des données évolutives*. Technip.

[15] GOVAERT G. (2003) *Analyse des données*. Hermès-Lavoisier.

[16] GRANGE D. et LEBART L. (1993) *Traitements statistiques des enquêtes*. Dunod.

[17] HUSSON F. et PAGES J. (2005) *Statistiques générales pour utilisateurs. 1 – Exercices corrigés*. Presses Universitaires de Rennes.

[18] JAMBU M. (1978) *Classification automatique pour l'analyse des données*. tome1 : Méthodes et algorithmes. Dunod.

[19] JAMBU M. et LEBEAUX M.O. (1978) *Classification automatique pour l'analyse des données*. tome 2 : Logiciels. Dunod.

[20] JAMBU M. (1989) *Exploration informatique et statistique des données*. Dunod.

[21] LEBART L., MORINEAU A., FENELON J.-P. (1979) *Traitement des données statistiques*. Dunod.

[22] LEBART L., MORINEAU A., PIRON M. (1998) *Statistique exploratoire multidimensionnelle*. Dunod.

[23] LEBART L. et SALEM A. (1994) *Statistique textuelle*. Dunod.

[24] MOREAU J., DOUDIN P.-A., CAZES P. (2000) *L'Analyse de correspondances et techniques connexes*. Spinger.

[25] PAGES J. (2005) *Statistiques générales pour utilisateurs. 1 - Méthodologie*. Presses Universitaires de Rennes.

[26] SAPORTA G. (1989) *Probabilités, analyse des données et statistique*. Technip.

[27] SCHIFFMAN S., REYNOLDS M., YOUNG F. (1981) *Introduction to multidimensional scaling*. New-york Academic Press.

[28] TENENHAUS M. (2007) *Statistique*. Dunod.

[29] VOLLE M. (1985) *Analyse des données*. Economica.

## LOGICIEL

Toutes les méthodes décrites dans ce livre sont intégrées dans FactoMineR, logiciel libre (en R) d'analyse des données. FactoMineR est développé par le laboratoire de Mathématiques appliquées d'AgroCampus.

SCIENCES SUP

*4e édition*

Brigitte Escofier
Jérôme Pagès

# ANALYSES FACTORIELLES SIMPLES ET MULTIPLES

## Objectifs, méthodes et interprétation

Cet ouvrage est destiné aux étudiants en Masters de mathématiques appliquées, d'économie ou d'économétrie, ainsi qu'aux élèves ingénieurs. Il aborde les méthodes d'analyse des données qui ont démontré leur efficacité dans l'étude des grandes masses complexes d'informations. Ces méthodes sont maintenant appliquées dans tous les domaines où l'on accumule d'importants fichiers de données, et sont largement utilisées hors de leurs champs traditionnels.

Pour cette quatrième édition, le texte a été révisé et augmenté notamment sur deux points qui correspondent à une demande croissante des utilisateurs :

– une présentation de l'analyse factorielle sur données mixtes (AFDM) ;
– une présentation de l'Analyse Factorielle Multiple Hiérarchique (AFMH), prolongement naturel de l'AFM.

Le cours est illustré par de nombreuses études de cas.

BRIGITTE ESCOFIER
a été professeur à l'Université de Rennes et à l'IUT de Vannes. Elle était l'une des fondatrices de l'École française d'analyse des données.

JÉRÔME PAGÈS
est ingénieur agronome, professeur à l'Agrocampus de Rennes.

MATHÉMATIQUES
PHYSIQUE
CHIMIE
SCIENCES DE L'INGÉNIEUR
INFORMATIQUE
SCIENCES DE LA VIE
SCIENCES DE LA TERRE

LICENCE 1 2 3 | MASTER 4 5 | DOCTORAT 6 7 8

ISBN 978-2-10-053809-6

www.dunod.com

DUNOD