

# Statistique Exploratoire Multivariée

Philippe MICHEL

Année universitaire 2012-2013



# Chapitre 1

## Introduction

L'objet de ce document est de présenter les méthodes que l'on regroupe sous le terme générique quelque peu vague d'Analyse des Données et de faciliter leur mise en oeuvre. La démarche suivie vise à permettre au lecteur une pratique de ces différentes techniques, en lui assurant également une présentation des fondements théoriques, de manière à faciliter outre une meilleure compréhension de l'outil, un éventuel approfondissement. Chaque méthode présentée sera illustrée à l'aide d'un exemple réel et non d'un cas d'école fabriqué pour l'occasion. En ce qui concerne les aspects informatiques, sauf mention contraire, le logiciel SPAD a été seul utilisé ici. Notre objet n'est pas dans ce domaine d'émettre un jugement sur les divers produits proposés. Par ailleurs, il ne s'agit pas non plus de présenter les commandes informatiques correspondantes. Enfin, c'est par la pratique et sur ses propres données que l'on peut le mieux appréhender ces différents outils. En effet, seule une connaissance approfondie du domaine étudié permet de sérier un résultat élémentaire, voire trivial, d'une piste de réflexion originale.

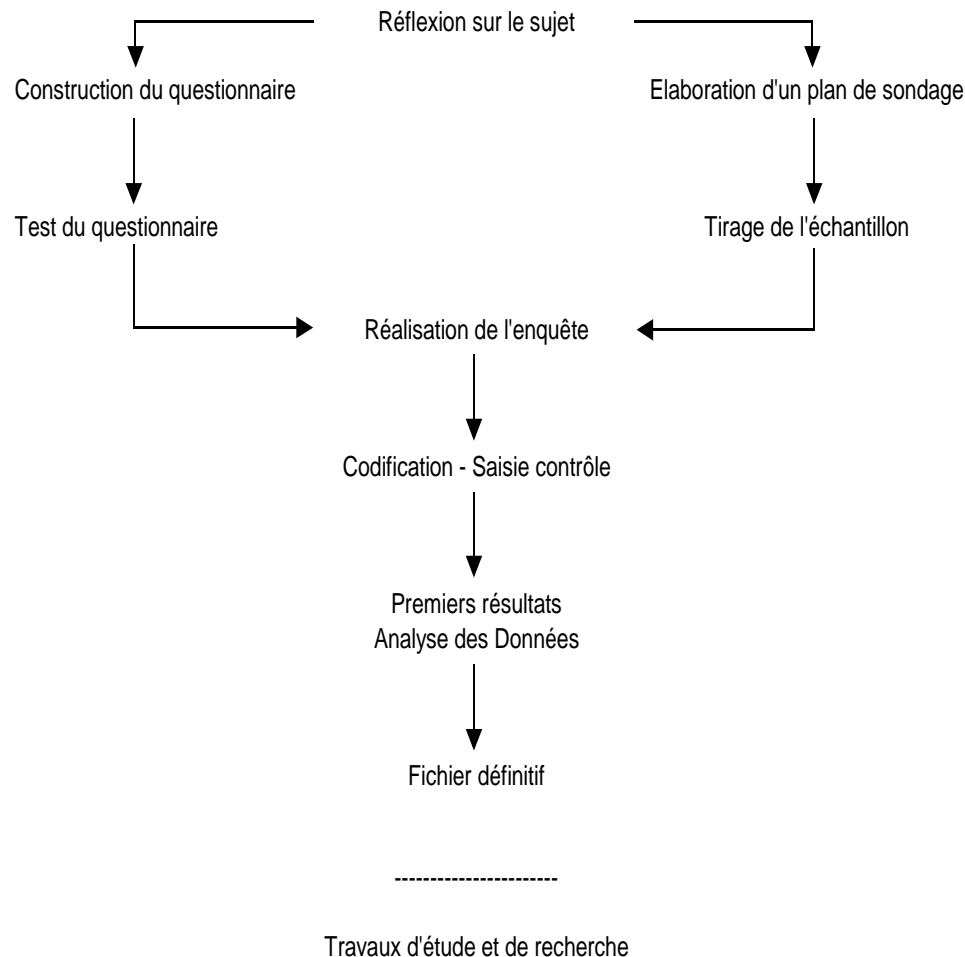
Si le terme générique d'Analyse des Données est quelque peu équivoque, il recouvre toutefois un domaine bien précis. L'Analyse des Données constitue désormais un outil puissant dans le traitement des données dont plus personne, semble-t-il, ne conteste l'utilité. Si la Statistique ne se résume certes pas à l'Analyse des Données, il semble en revanche difficile d'imaginer qu'elle n'en fasse pas partie. Ces méthodes se situent essentiellement dans un contexte exploratoire. La Statistique descriptive usuelle décrit les variables isolément, voire deux par deux, mais ne peut aller au-delà. En ce sens, l'Analyse des Données peut être définie dans un premier temps comme l'extension de la Statistique descriptive au cas multidimensionnel. L'approche s'affirme ainsi avant tout descriptive et vise à permettre une analyse simultanée de l'ensemble des variables. Nous disposons en effet désormais de capacités de

stockage quasi illimitées et d'un accès immédiat à l'information à travers les bases de données. Toutefois disposer d'une masse de données pose le problème des outils susceptibles d'extraire de ces gisements l'information pertinente. Tel est précisément l'objet des méthodes qui seront progressivement présentées.

Avant d'aborder cette présentation précisons le rôle et la place de l'Analyse des Données dans le cadre du traitement d'une enquête par exemple. L'opération commence en règle générale par une définition précise des objectifs puis par l'élaboration progressive du questionnaire ainsi que par la confection éventuelle d'un plan de sondage<sup>1</sup>. Après divers essais la version définitive du questionnaire est mise au point. L'enquête se déroule alors concrètement sur le terrain et les questionnaires sont ensuite récupérés pour être codés et saisis informatiquement (La tendance actuelle semble d'ailleurs évoluer vers le recueil direct des réponses lors de l'entretien, voire par l'administration directe du questionnaire par Internet). Après réalisation de tests qui de plus en plus sont réalisés en temps réel (saisie-contrôle) vient la phase de sortie des premiers résultats. Traditionnellement ceux-ci comportent une batterie d'indicateurs (moyennes, minimum, ...) ainsi que des tris à plat ou des tabulations (tableaux croisant deux variables). Les limites d'une telle démarche sont vite atteintes. L'information disponible n'est pas réellement analysée de manière globale. Les données sont plutôt examinées séquentiellement et les interactions éventuelles entre plusieurs variables ne peuvent pas être détectées. Prenons le cas d'une enquête de dimension modeste comportant 100 questions. Des notions élémentaires d'analyse combinatoire montrent que le nombre de tableaux croisés que l'on peut constituer est de 4950, ce qui constitue une masse considérable tout aussi inexploitable que la première ... selon toutes probabilités. C'est précisément à ce niveau que l'Analyse de Données peut trouver son utilité la plus grande. Elle va permettre de dégrossir les données, de repérer les grandes lignes et surtout de synthétiser et de hiérarchiser l'information. De plus, elle peut également parfois suggérer des lignes de recherche. Elle peut par ailleurs être utilisée vers l'amont en tant qu'outil lié à la production statistique en permettant un repérage d'erreurs ou une détection de résultats atypiques. Vient ensuite la phase de recherche et d'étude proprement dite et en règle générale le travail du statisticien s'arrête là où commence celui du chargé d'étude ou du chercheur. Il est dommage que parfois ce dernier ignore les résultats de la phase précédente qui pourraient lui permettre de travailler plus efficacement. Ainsi en va-t-il notamment de la

---

1. Cette phase peut elle-même impliquer l'utilisation de l'Analyse de Données, par exemple lors de la réalisation d'un plan de sondage stratifié.



détection des observations atypiques (outliers) dans la phase de modélisation où des méthodes diverses et variées sont utilisées là où l'Analyse Factorielle aurait permis très en amont la détection des observations correspondantes.

Le lecteur intéressé par une présentation plus complète pourra se reporter à l'ouvrage de Michel VOLLE <sup>2</sup>[1]

Ainsi l'Analyse des Données constitue un outil à la fois orienté vers l'amont (production) et vers l'aval (exploitation).

---

2. Analyse des Données - Michel VOLLE - Editions Economica - pages 8 et suivantes.

Quelles méthodes recouvre le terme d'Analyse de Données ?

En fait comme nous l'avons souligné ce terme générique quelque peu équivoque recouvre des techniques bien précises, de manière schématique l'Analyse Factorielle et la Classification.

L'objet des méthodes factorielles est toujours de « résumer » au mieux des tableaux rectangulaires de données. Cette définition est certes encore très vague mais elle sera progressivement affinée. Les différentes techniques se distinguent selon la nature des données analysées. Leur démarche consiste à remplacer les variables d'origine, nombreuses, par de nouvelles variables, synthétiques, et à conserver seulement les premières d'entre elles. Il est possible de mettre en évidence une démarche commune à ces techniques qui peut se résumer en quatre étapes :

1. Transformer les données initiales. La traduction mathématique de cette opération consiste à choisir une distance. La nature des données initiales conditionne très largement le choix de cette distance.
2. Déterminer à partir des données représentées dans l'espace formé par les variables initiales un nouveau repère. Ce nouveau repère est constitué séquentiellement, axe par axe, de manière à ce que chaque axe résume les données initiales de manière de moins en moins satisfaisante. L'aptitude de chaque axe à représenter les données d'origine se dégradant au fur et à mesure que l'on observe des axes de rang de plus en plus élevé.
3. A projeter les données d'origine sur chacun des axes ainsi formés.
4. A réduire la dimension d'origine de l'espace dans lequel se situent les données par sélection des premiers axes déterminés au point précédent.

Avant d'aborder plus en détail ces différentes phases et leur formalisation, indiquons que celles-ci sont réalisées automatiquement par les logiciels et que l'objet de l'Analyse de Données se situe ailleurs. Pratiquer l'Analyse de Données consiste face à un problème posé, en premier lieu à choisir une ou plusieurs méthodes selon à la fois la nature des données et l'objectif recherché, puis à interpréter les résultats, c'est à dire essayer de donner aux nouvelles variables, ou variables de synthèse, une signification. Ce travail parfois long et fastidieux demeure lui à l'échelle humaine. En ce sens, l'Analyse de Données ne fait pas « gagner » de temps mais permet de saisir l'essentiel de l'information contenue dans les données et de la hiérarchiser.

Précisons maintenant la démarche suivie :

Dans un premier chapitre nous présenterons de manière formelle la démarche sous-jacente aux diverses méthodes factorielles, moyennant quelques hypothèses simplificatrices et ce, sans référence à une technique particulière. Nous consacrerons ensuite une seconde partie à la méthodologie de l'interprétation. La présentation de la démarche générale effectuée nous aborderons les différentes méthodes ainsi que leurs propriétés. L'aspect pratique guidera notre choix et seules les méthodes les plus utilisées seront abordées. Nous reformulerons enfin la démarche générale en supprimant les hypothèses particulières, à savoir masses égales sur les observations et métrique identité. Nous montrerons de plus que l'on peut toujours se ramener par une transformation des données au cas initial.

Le second chapitre sera consacré à l'une des méthodes factorielles les plus anciennes mais sans conteste l'une des plus utilisées, l'Analyse en Composantes Principales (ACP). Celle-ci traite des tableaux individus-variables lorsque ces dernières sont quantitatives ou continues. L'ACP sur données centrées réduites sera exposée car elle constitue le cas le plus fréquent.

Le troisième chapitre sera dévolu à l'une des techniques factorielles particulièrement prisée en France, l'Analyse des Correspondances (AFC). Sur un cas concret nous verrons l'intérêt d'une telle méthode et pourquoi l'ACP se révèle mal adaptée au problème posé. L'AFC concerne en premier lieu les tableaux croisant une population selon deux critères qualitatifs ou encore tableaux dits de contingence. Cette méthode permettra de présenter une distance bien adaptée à l'étude des profils et connue sous le nom de distance du khi-2. Après avoir présenté les principales propriétés de cette technique nous ferons le lien avec le test, non moins célèbre, du même nom. L'AFC, outre son intérêt propre, offre de nombreuses extensions. L'une des plus intéressantes demeure sans conteste son utilisation sur des données binaires, connue sous le nom d'Analyse des Correspondances Multiples (ACM).

Le chapitre quatre sera dévolu à la présentation de cette méthode comme une simple extension de l'AFC au cas des données mises sous forme disjonctive complète, l'aspect théorique étant reporté à une partie ultérieure. La finalité de ce document étant avant tout d'ordre appliquée, il a paru opportun de présenter en premier lieu les propriétés de l'outil et les problèmes liés à son utilisation plutôt que la justification théorique de la méthode.

La seconde partie du cours est consacrée aux méthodes de classification. Cet ensemble très vaste constitue un champ d'étude à lui seul. L'optique privilégiée est ici l'utilisation des techniques de classification en complément des

méthodes factorielles, ce qui est souvent le cas en pratique.

En premier lieu nous verrons la classification ascendante hiérarchique (CAH), très utilisée sur le plan pratique et qui s'est désormais affranchie des contraintes qui il y a peu de temps encore limitaient son emploi.

Dans un second temps nous présenterons deux méthodes de partitionnement, centres mobiles et nuées dynamiques, dont l'emploi régresse désormais en raison de la possible utilisation des méthodes hiérarchiques sur de grands gisements de données.



# Chapitre 2

## Analyse d'un nuage de points quelconque

La démarche générale de l'Analyse d'un nuage de points est présentée ici et ce, sans référence à une méthode particulière. Une fois posé le modèle général, il suffira ensuite pour chaque méthode de préciser son originalité et ses propriétés spécifiques.

### 2.1 Introduction

Soit un tableau de données rectangulaire. Nous supposons que ce tableau noté  $X$  comporte  $n$  lignes et  $p$  colonnes. Par souci de simplification nous considérerons que les lignes correspondent aux individus et les colonnes aux variables. La césure est certes habituelle mais nous verrons que dans certains cas, l'AFC par exemple, elle est artificielle.  $X$  est donc une matrice  $[n, p]$  de terme générique  $x_{ij}$ , ce terme désignant la valeur prise par la variable  $j$  sur l'individu  $i$ , par exemple la surface agricole (variable  $j$ ) de l'exploitation  $i$ .

### 2.2 Analyse directe

#### 2.2.1 Position générale du problème

Chaque individu peut être représenté par un point dans l'espace à  $p$  directions défini par les variables. Nous ne pouvons représenter graphiquement un tel espace dès que  $p$  devient supérieur à 3 mais nous pouvons conceptuellement imaginer un tel espace, chaque variable définissant un axe, ces derniers étant par ailleurs orthogonaux deux à deux. Cet espace se note  $\mathbf{R}^p$ . L'ensemble

des points ainsi représenté forme ce qu'il est convenu d'appeler le nuage des individus, noté  $N(I)$ . Le nuage des individus représente ainsi les  $n$  lignes (individus) dans l'espace des  $p$  colonnes (variables).

Le problème fondamental des méthodes factorielles est de déterminer un sous-espace de dimension réduite qui soit « compréhensible » par l'oeil sur lequel projeter le nuage. Si l'oeil était capable de voir dans  $\mathbf{R}^p$  alors l'Analyse de Données n'aurait pas d'utilité. La projection étant une opération déformante il importe de pouvoir apprécier la déformation subie, ce qui nous amène à caractériser dans un premier temps la notion de forme d'un nuage de points.

### 2.2.2 La démarche

Expliquons la démarche qui va nous guider dans l'obtention de ce sous-espace de dimension réduite. Nous allons dans un premier temps chercher un axe sur lequel le nuage se déforme le moins en projection. Une fois cet axe déterminé nous chercherons alors un second axe, sur lequel le nuage se déforme le moins en projection, après le premier, tout en étant orthogonal au premier. Il suffira ensuite de réitérer le processus jusqu'à l'obtention de  $p$  axes. Cette démarche consiste en fait à substituer au repère d'origine, constitué par les variables de base, un nouveau repère formé par les axes ainsi construits. La différence entre les deux repères réside dans le fait que dans le second repère les axes ne véhiculent pas la même information selon leur rang. Leur capacité à « résumer » le nuage se détériore au fur et à mesure que l'on observe des axes de rang élevé.

### 2.2.3 Inertie d'un nuage de points

Soit  $G$  le centre de gravité du nuage encore appelé barycentre ou point moyen. Nous considérerons dans un premier temps le nuage centré, c'est à dire que le barycentre coïncide avec l'origine. Si tel n'est pas le cas il suffira alors de centrer les données de manière à translater le barycentre à l'origine. Pour caractériser la forme du nuage nous allons considérer l'ensemble de ses constituants, donc chaque point. Un point intervient dans la forme du nuage à travers deux éléments : son éloignement de l'origine et sa masse (ou poids). Pour caractériser l'éloignement nous prendrons le carré de la distance entre cet élément et l'origine, soit  $d^2(i, O)$ . La distance euclidienne usuelle qui consiste à prendre la somme des carrés des écarts entre les coordonnées, constitue l'exemple le plus simple :

$$d^2(i, i') = (x_{i1} - x_{i'1})^2 + \dots + (x_{ij} - x_{i'j})^2 + \dots + (x_{ip} - x_{i'p})^2$$

La masse associée à un individu peut par exemple être son poids lors d'un sondage. Le poids ou coefficient d'extrapolation est alors l'inverse du taux de sondage. Considérons deux éléments tirés dans un échantillon, l'un possédant un taux de sondage de  $1/10$ , l'autre de  $1/20$ . Dans ce cas le premier élément représente 10 individus de l'univers et aura en conséquence un poids ou une masse de 10 et le second de 20. L'action de ces deux éléments, masse et distance, sera saisie sous la forme de leur produit, soit :

$$m_i d^2(i, O)$$

Définition : On appelle inertie du nuage de points par rapport à l'origine, notée  $I_0$ , la quantité :

$$I_0 = \sum_{i=1}^n m_i d^2(i, O)$$

Cette quantité caractéristique du nuage considéré, fait intervenir tous les éléments du nuage.

### 2.2.4 Recherche d'un axe $\Delta u_1$ conservant au mieux le nuage

Nous avons à la fois fixé notre objectif, déterminer un nouveau repère sur lequel représenter notre nuage et la démarche, à savoir procéder de manière séquentielle. L'indicateur que nous allons utiliser pour déterminer le premier axe est bien entendu l'inertie du nuage projeté. Dans l'espace d'origine  $\mathbf{R}^p$ , l'inertie du nuage est  $I_0$ . Sur  $\Delta u_1$ , l'inertie du nuage projeté est nécessairement inférieure à  $I_0$ , les  $n$  points étant alors alignés. Soit  $I_1$  cette quantité. Nous allons chercher  $\Delta u_1$  tel que l'inertie du nuage projeté  $I_1$  soit maximale.

Remarque : Une démarche alternative consiste à utiliser comme indicateur la somme des distances entre tous les points du nuage pris deux à deux, puis à rechercher l'axe sur lequel cette quantité se conserve le mieux. Cette approche est identique à la précédente en ce sens qu'elle conduit au même résultat. Ce fait n'est guère surprenant dans la mesure où l'on peut montrer que si  $I_0$  désigne l'inertie du nuage centré et  $D^2$  la somme des distances entre points alors  $D^2 = 2nI_0$ . En conséquence la direction de l'espace qui maximise l'une de ces deux quantités maximise également l'autre.

C'est ce type de représentation que l'Analyse Factorielle permet d'obtenir et dont l'interprétation repose sur l'examen attentif d'indicateurs que nous allons présenter.

## 2.3 Formalisation

### 2.3.1 Notations

Soient :

$X[n, p]$  la matrice des données de terme générique  $x_{ij}$

$x_i$  le vecteur  $[p, 1]$  des données relatives à l'individu  $i$ . Le transposé du vecteur  $x_i$  forme la ligne  $i$  de  $X$

$c_{i1}$  la longueur de la projection de  $x_i$  sur l'axe  $\Delta u_1$  de vecteur unitaire  $u_1$ . Cette quantité est donnée par le produit scalaire des vecteurs  $x_i$  et  $u_1$ , noté  $\langle x_i, u_1 \rangle$

Nous supposons dans un souci de simplification les masses égales, soit  $m_i = 1$  pour  $i = [1, n]$

Dans ces conditions, l'inertie du nuage projeté sur  $\Delta u_1$  s'écrit :

$$I_1 = \sum_{i=1}^n m_i c_{i1}^2$$

soit encore, compte tenu de l'hypothèse précédente :

$$I_1 = \sum_{i=1}^n c_{i1}^2$$

Notons  $C_1$  le vecteur  $[n, 1]$  dont le terme générique est  $c_{i1}$ . Alors l'inertie du nuage projeté  $I_1$  peut encore s'écrire sous forme matricielle  $I_1 = {}^t C_1 C_1$  où  ${}^t C_1$  désigne le transposé de  $C_1$ .

### 2.3.2 Position du problème

Chercher  $\Delta u_1$  rendant  $I_1$  maximale revient à résoudre le programme de maximisation suivant :  $\text{Max}_{\{u_1\}} {}^t C_1 C_1$  sous la contrainte  $u_1$  unitaire soit  $\|u_1\| = 1$ .

Or  $C_1 = X u_1$  (vérification immédiate),

d'où  ${}^t C_1 C_1 = {}^t (X u_1) (X u_1) = {}^t u_1 {}^t X X u_1$

Rappel :  ${}^t(AB) = {}^tB{}^tA$

Résolution : Il s'agit d'un programme de maximisation d'une forme quadratique sous contrainte qui va être résolu par la méthode des multiplicateurs de Lagrange.

Formons le lagrangien :  $L(u_1) = {}^tu_1{}^tXXu_1 - \lambda_1({}^tu_1u_1 - 1)$

Les conditions du premier ordre s'écrivent<sup>1</sup> :

$$\frac{\partial L}{\partial u_1} = 0 \Rightarrow 2{}^tXXu_1 - 2\lambda_1u_1 = 0$$

soit :

$${}^tXXu_1 = \lambda_1u_1 \quad (2.1)$$

La matrice  ${}^tXX$  que l'on peut noter  $V$  est carrée et symétrique et appelée matrice d'inertie. D'après la relation (2.1)  $u_1$  apparaît comme vecteur propre de la matrice  ${}^tXX$  associé à la valeur propre  $\lambda_1$ .

Cherchons à préciser  $\lambda_1$ .

Pré-multiplions la relation (2.1) par  ${}^tu_1$ , il vient :  ${}^tu_1{}^tXXu_1 = \lambda_1{}^tu_1u_1$  d'où  $\lambda_1 = {}^tu_1{}^tXXu_1$  car  $u_1$  étant normé par construction  ${}^tu_1u_1 = \|u_1\|^2 = 1$ . Or la quantité  ${}^tu_1{}^tXXu_1$  n'est autre que  $I_1$  l'inertie du nuage projeté sur l'axe  $\Delta u_1$ . En conséquence il suffit de prendre pour  $u_1$  le vecteur propre associé à la plus grande valeur propre de la matrice  $V = {}^tXX$ .

Remarques :

1. La matrice d'inertie  $V = {}^tXX$  étant symétrique et définie positive est, d'après un résultat d'Algèbre, diagonalisable et toutes ses valeurs propres sont positives ou nulles.
2. La relation  $C_1 = Xu_1$  exprime que  $C_1$  est obtenu comme combinaison linéaire des variables d'origine  $X_j$  au moyen du vecteur  $u_1$ .
3.  $u_1$  est appelé axe factoriel.

### 2.3.3 Recherche du second axe

Cherchons maintenant un axe  $\Delta u_2$  tel que l'inertie  $I_2$  du nuage  $N(I)$  projeté sur cet axe soit maximale après  $I_1$ . Nous imposerons de plus à ce premier axe d'être orthogonal au premier.

---

1. Voir la formule de dérivation vectorielle en annexe de ce chapitre

D'où le programme :  $Max_{\{u_2\}} {}^t u_2 {}^t X X u_2$

sous les contraintes :

$${}^t u_2 u_2 = 1 \text{ (} u_2 \text{ normé)}$$

$${}^t u_2 u_1 = 0 \text{ (} u_2 \text{ orthogonal à } u_1 \text{)}$$

Soit le Lagrangien  $L = {}^t u_2 {}^t X X u_2 - \lambda_2 ({}^t u_2 u_2 - 1) - \mu {}^t u_2 u_1$

Les conditions du premier ordre s'écrivent :

$$\frac{\partial L}{\partial u_2} = 0 \Rightarrow 2 {}^t X X u_2 - 2 \lambda_2 u_2 - \mu u_1 = 0$$

Pré-multiplions cette expression par  ${}^t u_1$  il vient :

$$2 {}^t u_1 {}^t X X u_2 - 2 \lambda_2 {}^t u_1 u_2 - \mu {}^t u_1 u_1 = 0$$

soit

$$2 {}^t u_1 {}^t X X u_2 - \mu = 0$$

car  ${}^t u_1 u_1 = 1$  et  ${}^t u_2 u_1 = 0$

Or  ${}^t u_1 {}^t X X u_2 = {}^t u_2 {}^t X X u_1$  car il s'agit d'un scalaire qui est donc égal à son transposé

d'autre part  ${}^t X X u_1 = \lambda_1 u_1$  (voir point précédent)

d'où  ${}^t u_2 {}^t X X u_1 = \lambda_1 {}^t u_2 u_1 = 0$  donc  $\mu = 0$

Dès lors le programme de maximisation se ramène au cas précédent. On en déduit qu'il suffit de prendre pour  $u_2$  le vecteur propre de la matrice d'inertie associé à la seconde plus grande valeur propre de la matrice d'inertie  ${}^t X X$ . L'essentiel d'un programme d'Analyse Factorielle se résume donc sur le plan informatique en une diagonalisation d'une matrice symétrique définie positive puis en un classement par ordre décroissant de ses valeurs propres, les vecteurs propres associés déterminant les axes du nouveau repère.

## 2.4 Propriétés

### 2.4.1 Inertie d'un sous-espace

L'inertie étant additive lorsqu'elle se décompose sur des sous-espaces orthogonaux (la propriété immédiate pour un sous-espace de dimension 2 se géné-

ralise sans difficulté par récurrence au cas  $q > 2$ ) nous pouvons écrire :

$$I_0 = \sum_{k=1}^p I_k = \sum_{k=1}^p \lambda_k$$

où  $\lambda_k$  désigne l'inertie conservée par l'axe de rang  $k$ .

Propriété : Soit par ailleurs  $U$  la matrice  $[p, p]$  dont la colonne  $j$  est formée du vecteur  $u_j$  ( $U$  est la matrice de passage de la base canonique de  $\mathbf{R}^p$  à la base des vecteurs propres).

Les matrices  $\Lambda$  et  $V = {}^t X X$  étant semblables vérifient la relation  $\Lambda = U^{-1} V U$  soit encore  $\Lambda = {}^t U V U$ , la matrice  $U$  étant orthonormale ( $U^{-1} = {}^t U$ ), d'où  $\text{tr} \Lambda = \text{tr}({}^t U V U)$

D'après les propriétés de l'opérateur trace<sup>2</sup> nous pouvons écrire :

$$\text{tr} \Lambda = \text{tr}({}^t U V U) = \text{tr} V$$

Par conséquent l'inertie totale du nuage  $I_0$  est égale à la trace de la matrice d'inertie :

$$I_0 = \text{tr}({}^t X X)$$

### 2.4.2 Recherche du meilleur sous-espace de dimension $q$ pour représenter le nuage

Il est immédiat que le meilleur sous-espace de dimension  $q$  ( $q < p$ ) pour représenter le nuage est obtenu en sélectionnant les  $q$  premiers axes. En particulier le meilleur plan pour représenter les données initiales est celui formé par les deux premiers axes. La démonstration en procédant par l'absurde est immédiate. Supposons en effet que le meilleur plan ne contienne pas  $u_1$ , il en existerait alors un meilleur contenant  $u_1$ . L'indicateur traditionnellement utilisé pour apprécier la capacité d'un axe à représenter les données est le taux d'inertie, c'est à dire le rapport entre  $I_k$ , inertie du nuage projeté sur l'axe de rang  $k$ , et  $I_0$  l'inertie totale :

## 2.5 Analyse dans $\mathbf{R}^n$

### 2.5.1 Introduction

Dans l'analyse directe (dans  $\mathbf{R}^p$ ) le tableau  $X$  à  $n$  lignes et  $p$  colonnes était considéré du point de vue des lignes, chacune constituant un point de l'espace

---

2.  $\text{tr}(ABC) = \text{tr}(BCA)$  lorsque les dimensions des matrices sont compatibles

de dimension  $p$  formé par les colonnes. Cette représentation paraît naturelle dans la mesure où un tel tableau est souvent du type individus-variables, les individus étant alors représentés dans l'espace défini par les variables.

De manière symétrique nous pourrions décider de représenter les variables dans l'espace défini par les individus ( $\mathbf{R}^n$ ). Cette démarche identique à la précédente constitue ce que l'usage a consacré sous le terme d'analyse duale. L'objet de cette partie est de montrer qu'il n'y a pas lieu de réitérer l'ensemble des calculs précédents mais que :

1. Les axes factoriels dans  $\mathbf{R}^n$  se déduisent de ceux obtenus dans l'analyse directe ( $\mathbf{R}^p$ ) et réciproquement
2. Les taux d'inertie sont identiques pour des axes de même rang dans les deux analyses

### 2.5.2 Formalisation

Dans l'analyse directe nous étions conduits à rechercher  $u_1$  tel que  ${}^tXXu_1 = \lambda_1 u_1$ ,  $\lambda_1$  étant la valeur propre maximale de la matrice d'inertie  ${}^tXX$ . Par raison de symétrie l'analyse duale va nous conduire à chercher  $v_1$  ( $\in \mathbf{R}^n$ ) tel que :

$$X^tXv_1 = \mu_1 v_1 \quad (2.2)$$

$v_1$  étant vecteur propre de la matrice  $X^tX$  associé à la valeur propre  $\mu_1$ .  
Prémultiplions cette dernière relation par  ${}^tX$  il vient :

$${}^tXX({}^tXv_1) = \mu_1({}^tXv_1)$$

On en déduit que  ${}^tXv_1$  est vecteur propre de  ${}^tXX$  associé à la valeur propre  $\mu_1$ .

La plus grande valeur propre de  ${}^tXX$  étant  $\lambda_1$  il est immédiat que  $\mu_1 \leq \lambda_1$ . En procédant de même avec la relation  ${}^tXXu_1 = \lambda_1 u_1$  on en déduirait que  $Xu_1$  est vecteur propre de  $X^tX$  associé à la valeur propre  $\lambda_1$ , d'où  $\lambda_1 \leq \mu_1$ , d'où finalement :  $\mu_1 = \lambda_1$ .

Les valeurs propres, donc les taux d'inertie issus des deux analyses, sont les mêmes pour des axes de rang homologue. En supposant sans perte de généralité  $p < n$ , alors dans  $\mathbf{R}^n$  nous obtenons  $p$  valeurs propres  $\lambda_k$  positives ou nulles et  $n - p$  valeurs propres nulles.

Cherchons maintenant une relation entre les vecteurs  $u_k$  et  $v_k$ .

${}^tXv_1$  est vecteur propre de  ${}^tXX$  associé à la valeur propre  $\lambda_1$ . Ce vecteur n'est pas unitaire. En effet  $\|{}^tXv_1\|^2 = {}^t v_1 X^t X v_1 = I_1 = \mu_1$ , l'espace étant muni de la métrique euclidienne canonique  $M = I$ . Par conséquent le vecteur

$$\frac{1}{\sqrt{\lambda_1}} {}^tXv_1$$



est normé.

On en déduit les deux relations suivantes qui peuvent être étendues à des axes de rang quelconque :

$$u_k = \frac{1}{\sqrt{\lambda_k}} {}^t X v_k \quad (2.3)$$

et

$$v_k = \frac{1}{\sqrt{\lambda_k}} X u_k \quad (2.4)$$

Ces relations fondamentales, appelées formules de transition, montrent que les axes dans un espace peuvent se déduire de ceux obtenus dans l'autre espace. Sur le plan informatique les formules précédentes impliquent l'économie d'une analyse. Une fois la matrice  ${}^t X X$  diagonalisée il n'est pas nécessaire de réitérer l'opération pour la matrice  $X {}^t X$ .

Ces relations permettent en effet d'obtenir directement les coordonnées des variables. En effet soit  $D_k$  le vecteur  $[p, 1]$  des coordonnées des variables sur l'axe  $k$ , alors

$$D_k = {}^t X v_k$$

d'après (2.4) il vient :

$$D_k = \frac{1}{\sqrt{\lambda_k}} {}^t X X u_k$$

comme par ailleurs :

$$C_k = X u_k$$

on en déduit :

$$D_k = \frac{1}{\sqrt{\lambda_k}} {}^t X C_k \quad (2.5)$$

de manière symétrique :

$$C_k = \frac{1}{\sqrt{\lambda_k}} X D_k \quad (2.6)$$

Les formules de transition s'étendent donc aux coordonnées des éléments.

Ces relations fondamentales, appelées formules de transition, montrent que les axes dans un espace peuvent se déduire de ceux obtenus dans l'autre espace.

## 2.6 Reconstitution des données initiales

A partir de la relation :

$$v_k = \frac{1}{\sqrt{\lambda_k}} X u_k$$

multiplions à droite les deux membres par  ${}^t u_k$ , il vient :

$$v_k^t u_k = \sqrt{\lambda_k} X u_k^t u_k$$

Sommons sur  $k$  ces deux quantités :

$$X \sum_{k=1}^p u_k^t u_k = \sum_{k=1}^p \sqrt{\lambda_k} v_k^t u_k$$

Les vecteurs  $(u_k)$  formant une base orthonormale alors :

$$\sum_{k=1}^p u_k^t u_k = I$$

Donc :

$$X = \sum_{k=1}^p \sqrt{\lambda_k} v_k^t u_k \quad (2.7)$$

Cette dernière relation montre que le tableau  $X$  des données initiales peut être reconstruit à partir des axes factoriels dans  $\mathbf{R}^p$  (les vecteurs  $u_k$ ) et dans  $\mathbf{R}^n$  (les vecteurs  $v_k$ ). Il s'agit donc bien d'une formule de reconstitution des données initiales.

Lors de l'analyse nous allons sélectionner le nombre d'axes et en retenir un nombre  $q(< p)$ . Dans ce cas appelons  $X^*$  le tableau reconstitué à partir des  $q$  premiers axes, soit :

$$X^* = \sum_{k=1}^{q < p} \sqrt{\lambda_k} v_k^t u_k$$

Sélectionner les  $q$  premiers axes revient à substituer  $X^*$  à  $X$ .

La qualité de cette approximation peut être appréciée en comparant les traces des matrices  ${}^t X^* X^*$  et  ${}^t X X$ , donc les inerties respectives :

$$\sum_{k=1}^q I_k$$

et

$$\sum_{k=1}^p I_k$$

Le rapport de ces deux quantités fournit un indicateur de l'approximation opérée.

Observons que si l'on prend les  $p$  axes factoriels, l'inertie totale  $I_0$  du nuage est conservée. Dans ce cas le nuage est représenté dans un espace de même

dimension et seul un changement de base a été effectué. Les méthodes factorielles consistent à accepter de perdre de l'inertie pour pouvoir interpréter les données. Soit à « consentir une perte d'information pour obtenir un gain en signification » (VOLLE).

Remarque :

La démarche qui vient d'être présentée se fonde sur la décomposition en valeurs singulières d'une matrice rectangulaire encore appelée décomposition de ECKART-YOUNG. Certaines applications se fondent directement sur la reconstitution des données à partir d'un nombre réduit d'axes.

## 2.7 Méthodologie de l'interprétation

### 2.7.1 Sélection du nombre d'axes à analyser

Nous avons vu que l'essentiel des méthodes factorielles consiste à substituer aux variables d'origine de nouvelles variables. Interpréter les résultats d'une analyse consiste précisément à donner une signification à ces nouveaux axes. Dans l'interprétation des résultats, la première étape concerne en général le nombre d'axes à retenir. Cette question simple n'appelle pas hélas de réponse simple. Il convient de souligner que l'examen des taux d'inertie, pour nécessaire qu'il soit, n'en est pas pour autant suffisant :

1. D'une part selon la méthode utilisée les taux, y compris sur les premiers axes, peuvent par construction être faibles. Tel est le cas en Analyse des Correspondances Multiples où la mise sous forme disjonctive complète des données, en substituant les modalités aux variables, conduit à de tels résultats.
2. En second lieu des taux très élevés n'impliquent pas pour autant que l'axe présente un quelconque intérêt. Il se peut que le phénomène mis en évidence soit trivial et de plus masque un élément intéressant qui apparaîtra sur les axes suivants. A cet égard, si les termes d'inertie et d'information sont souvent employés comme synonymes, ils ne possèdent pas pour autant la même signification.
3. Il faut également tenir compte de la dimension du tableau dans l'appréciation des résultats. Un taux d'inertie de 10 % sur un axe ne possède pas la même signification selon que le tableau comporte 20 ou 100 variables.

Nous allons maintenant examiner quelques critères.

### Critères théoriques

Certains auteurs ont cherché à construire des tests statistiques afin de déterminer le caractère significatif ou non d'un axe. La plupart de ces tests reposent sur l'hypothèse de normalité et de plus été sont obtenus dans un cadre asymptotique. Tel est le cas en Analyse en Composantes Principales, où un logiciel comme SPAD fournit pour chaque valeur propre un intervalle de confiance au seuil de 95 % (Intervalle Laplacien d'Anderson<sup>3</sup>). Ces intervalles fournissent une indication sur la stabilité de la valeur propre correspondante. Le recouvrement de deux valeurs propres consécutives suggère alors l'instabilité des axes correspondants, qu'il vaut mieux dans ces conditions ne pas retenir dans l'interprétation. Nous verrons sur un exemple que ce test se révèle en pratique peu applicable, surtout lorsque le nombre d'observations est faible. Il peut en effet conduire à rejeter la totalité des axes alors que les premiers axes sont parfaitement interprétables.

En Analyse Factorielle des Correspondances la loi suivie par les valeurs propres a été approchée (loi de Wishart). Par ailleurs des simulations ont permis la construction d'abaques. Ces courbes fournissent selon la taille du tableau traité la valeur maximale que peut atteindre la première valeur propre dans l'hypothèse d'indépendance des lignes et des colonnes. Cette démarche n'est valable qu'en AFC<sup>4</sup>. Pour cette dernière méthode il existe par ailleurs un test spécifique. Ces différentes approches seront vues lors de la présentation de la méthode correspondante.

### Critères empiriques

Le meilleur critère qui puisse être utilisé en pratique dans la sélection des axes est empirique et consiste à examiner non pas la signification statistique d'un axe mais son interprétabilité. Peut-on donner une signification claire à l'axe ? L'hypothèse implicite est que l'axe cessera d'être interprétable avant de cesser d'être statistiquement significatif. Signalons également parmi les critères empiriques celui du « coude ». Ce test consiste à repérer l'évolution des taux d'inertie. On observe en général des sauts sur les premiers axes puis une décroissance régulière à partir d'un certain rang. Les données traitées contiennent dans ce cas des phénomènes structurels, qui expliquent précisément les sauts, et du « bruit » qui implique l'allure de la courbe à partir d'un certain rang. En ce sens les méthodes factorielles peuvent être vues comme un outil permettant de séparer l'information du bruit.

---

3. Voir Statistique Exploratoire Multidimensionnelle - Lebart, Morineau et Piron - p 375

4. Voir Techniques de la Description Statistique - P 223 et suivantes

HISTOGRAMME DES 8 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	16.6794	43.89	43.89	*****
2	6.1735	16.25	60.14	*****
3	5.0191	13.21	73.35	*****
4	3.0110	7.92	81.27	***
5	2.1222	5.58	86.86	**
6	1.2627	3.32	90.18	*
7	.0084	.10	100.00	*
8	.0000	.00	100.00	*

Sur cet exemple on observe nettement la décroissance régulière des taux d'inertie avec le quatrième axe. Il est fort probable que seuls les trois premiers présentent un éventuel intérêt.

Le scree-test de Catell fournit une autre méthode de sélection des axes. Il consiste à repérer, s'il existe, un point d'inflexion dans la courbe de décroissance des valeurs propres. On commence par calculer les différences premières entre valeurs propres, soit  $\lambda_{k-1} - \lambda_k = \mu_k$ , puis les différences secondes  $\mu_{k-1} - \mu_k$ . On retient ensuite les axes pour lesquels les différences secondes sont toutes de même signe.

Appliquons ce test à l'exemple précédent :

Axe	Valeur propre $\lambda_k$	$\lambda_{k-1} - \lambda_k = \mu_k$	$\mu_{k-1} - \mu_k$
1	16.6794		
2	6.1735	10.5059	
3	5.0191	1.1544	9.3515
4	3.0110	2.0081	-0.8537

L'application de ce test conduit également ici à retenir les trois premiers axes. Le lecteur intéressé par les critères pratiques de sélection des axes selon chaque méthode pourra également se reporter à l'ouvrage de B. ESCOFIER et J. PAGES<sup>5</sup>.

Bien que l'Analyse Factorielle soit fondamentalement une méthode exploratoire, une autre démarche peut consister à examiner les axes non pas de manière séquentielle, mais à rechercher sur quel(s) axe(s) se manifeste un éventuel phénomène que l'on suppose a priori. Le problème posé étant celui de l'interprétation d'un axe nous allons maintenant examiner ce point en détail. Il convient en premier lieu d'examiner les axes séparément et en second lieu d'examiner les nuages, individus et variables, de manière séparée.

5. Analyse Factorielles Simples et Multiples - p225 et suivantes

## 2.7.2 Examen d'un nuage

### Une notion fondamentale : la contribution (absolue)

Le choix du nuage comportant une part d'arbitraire on peut décider d'interpréter l'un ou l'autre en premier lieu selon l'habitude, la démarche étant identique quel que soit le nuage. Rappelons que l'inertie du nuage projeté sur l'axe de rang  $k$  s'écrit :

$$I_k = \sum_{i=1}^n m_i c_{ik}^2$$

où  $m_i$  désigne la masse afférente à l'élément  $i$  et  $c_{ik}$  la longueur de sa projection sur l'axe de rang  $k$ .

Chaque élément participe donc à la formation de la quantité  $I_k$ .

Afin de déterminer le rôle pris par l'individu  $i$  dans l'élaboration de l'axe  $k$ , on examine la part de sa contribution à  $I_k$ . Elle est souvent notée  $CTR_k(i)$  et exprimée en pourcentage ou en millièmes :

$$CTR_k(i) = \frac{m_i c_{ik}^2}{I_k}$$

Cette quantité est appelée contribution de l'individu  $i$  à l'inertie de l'axe  $k$  (ou encore contribution absolue, voire parfois contribution relative).

### De l'importance de l'examen des contributions

Ces quantités sont fondamentales et c'est sur leur examen que repose l'interprétation d'un axe. Les erreurs qui sont le plus souvent commises dans l'interprétation des résultats d'une analyse, bien que celles-ci tendent à s'estomper, proviennent de l'interprétation directe des résultats à partir des seules sorties graphiques. Dans ce cas l'interprétation s'opère implicitement sur le seul examen des éléments éloignés de l'origine donc sur la distance. Or l'inertie fait intervenir deux éléments, masse et distance. L'interprétation « spatiale-naïve » conduit à ne pas tenir compte des masses, ce qui fausse l'interprétation. On objectera que dans le cas où les individus sont assortis de masses égales la démarche précédente est licite. Elle constitue toutefois une très mauvaise habitude dans la mesure où avec certaines méthodes existent des masses implicites (en AFC par exemple) et où il convient d'avoir toujours une idée précise des contributions.

### Le cas des très fortes contributions

Par ailleurs une contribution absolue très forte d'un élément doit attirer l'attention. Une telle situation révèle soit une erreur, soit un élément atypique, et il conviendra d'opérer un retour aux données d'origine à des fins de vérification. Dans le cas où il ne s'agit pas d'une erreur lors du recueil, du codage ou de la saisie des données, une forte contribution traduit l'existence d'un élément atypique. Il convient alors de s'interroger sur le maintien éventuel d'un tel élément dans l'analyse. C'est en ce sens que les méthodes factorielles peuvent être utilisées comme outil lors de la phase préliminaire d'apurement des données. Prenons un exemple sur un cas concret. A l'issue d'une enquête sur le cheptel porcin en Bretagne on désire dresser une typologie des producteurs de la région. Sur les 1300 exploitations analysées l'une se révèle avoir une contribution d'environ 6 % sur le premier axe, soit 78 fois la contribution moyenne. Après examen des données de base il s'avère qu'il ne s'agit pas d'une erreur mais d'une exploitation tout à fait particulière dans la région. L'objet de l'étude étant, entre autres, de dresser une typologie des producteurs de la région, la décision a été prise de retirer cette exploitation de l'analyse en raison de son caractère singulier. Cette unité ne représente en fait qu'elle même.

Compte tenu de la définition de la contribution l'atypie d'un élément ne peut provenir que de sa masse ou de sa distance. Si la distance est en cause cela signifie que l'élément se distingue de l'ensemble par les valeurs prises par les variables. Il faut alors examiner si celles-ci constituent un cas extrême ou traduisent un cas différent par nature. Dans ce second cas s'il s'agit d'une enquête par sondage on pourra alors réfléchir à la constitution d'une strate éventuelle contenant ces cas et qui seront sondés de manière exhaustive.

Dans le cas où l'atypie provient de la masse afférente à l'individu, le diagnostic dépend de la méthode utilisée et de l'origine des données. S'il s'agit d'une enquête par sondage, la masse étant constituée par le coefficient d'extrapolation soit l'inverse du taux de sondage, cela signifie que ce dernier était peut être trop faible dans la strate correspondante.

#### Remarque :

Il n'existe pas de règle indiquant combien d'individus sélectionner. Plusieurs méthodes peuvent être envisagées. En règle générale on commencera par observer la présence d'éventuels éléments atypiques. On peut ensuite décider de retenir les éléments dont la contribution est supérieure à la contribution moyenne par exemple. Toutefois cette dernière démarche peut conduire à retenir trop d'éléments. Une variante consiste à se fixer un seuil minimum de contribution du type  $n$  fois la contribution moyenne. Une autre démarche

consiste à se fixer a priori un pourcentage global, par exemple 70 %, et à sélectionner les éléments jusqu'à obtenir 70 % de l'inertie totale de l'axe. Le seuil de 70 % est parfaitement arbitraire.

### La phase d'interprétation

Une fois les éléments possédant la plus forte contribution sélectionnés on pourra en règle générale les scinder en deux groupes selon le signe de leur coordonnée sur l'axe. L'axe va posséder des caractéristiques propres à ces deux groupes, puisque les éléments qui les composent interviennent fortement dans son élaboration. En procédant de même pour l'autre nuage on dégagera progressivement une interprétation de l'axe correspondant. C'est la règle "toujours transiter par les axes pour passer d'un nuage à l'autre". Afin d'illustrer l'interprétation précédente on pourra effectuer un retour aux données d'origine en sélectionnant les valeurs prises par les individus à forte contribution sur les variables également à forte contribution. C'est en ce sens que l'Analyse Factorielle constitue un outil d'analyse et d'exploration des grands tableaux en permettant d'extraire des données initiales l'information essentielle.

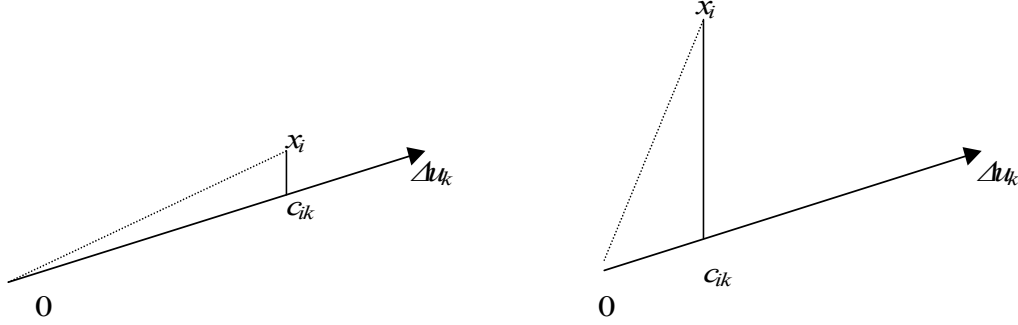
Remarque :

L'examen des axes s'effectue de manière séquentielle, axe par axe. Il appartiendra à l'analyste d'opérer lui-même les appariements sous forme de plan. L'édition systématique de plans du type  $(1 \times 2)$ ,  $(3 \times 4)$ , ... ne possède pas nécessairement un intérêt. Signalons que les représentations graphiques sont en général séparées et font l'objet de 2 graphiques distincts, l'un pour les individus l'autre pour les variables. Nous verrons que dans certains cas une propriété légitime la représentation simultanée des 2 nuages. Toutefois s'il est licite, sous certaines conditions, de comparer les positions prises par 2 éléments d'un même espace, il est en revanche périlleux de faire de même entre éléments d'espaces différents. On appelle plan  $(j \times k)$  le plan formé par l'appariement des axes  $j$  et  $k$ .

## 2.8 Qualité de la représentation

L'opération de projection du nuage sur un axe constituant une opération déformante, on peut chercher à savoir quelle est la proximité entre un élément et un axe de manière à répondre à la question suivante : l'élément est-il proche ou éloigné de sa projection ?





Dans le premier cas, l'élément est proche de sa projection sur l'axe. Il est dit « bien représenté » sur l'axe (de rang  $k$ ). A l'inverse, dans le second cas, l'élément est éloigné de sa projection. Il est dit « mal représenté » sur cet axe.

L'indicateur que l'on songe de manière naturelle à utiliser est l'angle formé entre l'élément et sa projection sur l'axe. L'indicateur utilisé est le carré du cosinus de cet angle. Lorsque l'angle est faible, proche de 0, le cosinus carré est lui proche de 1. En revanche lorsque l'angle est important, proche de  $\frac{\pi}{2}$ , le cosinus carré est lui proche de 0. On obtient ainsi un coefficient qui varie entre 0 et 1 et dont l'interprétation est simple : proche de 0, l'élément est mal représenté sur l'axe, proche de 1, l'élément est bien représenté sur l'axe. Ce coefficient présente une analogie avec le  $R^2$  de la régression.

La qualité de la représentation de l'individu  $i$  sur l'axe  $k$  est donnée par :

$$q_{lt_k}(i) = \cos^2 \theta(i, k) = \frac{(\langle x_i, u_k \rangle)^2}{\|x_i\|^2 \|u_k\|^2} = \frac{c_{ik}^2}{\|x_i\|^2}$$

car  $u_k$  est normé.

La qualité de la représentation de la variable  $X_j$  sur l'axe  $k$  est donnée par :

$$q_{lt_k}(j) = \cos^2 \theta(j, k) = \frac{(\langle X_j, v_k \rangle)^2}{\|X_j\|^2 \|v_k\|^2} = \frac{d_{jk}^2}{\|X_j\|^2}$$

car  $v_k$  est normé.

Le cosinus carré possède en outre une propriété que ne possède pas l'angle. Il

est additif sur des sous-espaces orthogonaux. Ainsi si l'on connaît la qualité de la représentation de l'élément  $i$  sur l'axe  $k$ , soit  $qlt_k(i)$ , et la qualité de la représentation de ce même élément sur l'axe  $k'$  ( $k' \neq k$ ), pour obtenir la qualité de sa représentation sur le plan formé par les axes  $k$  et  $k'$  il suffit de prendre :

$$qlt_{(k*k')}(i) = qlt_k(i) + qlt_{k'}(i)$$

Cette propriété se généralise au cas de sous-espaces de dimension supérieure à 2. Ainsi la qualité de représentation d'un élément sur le sous-espace formé par les  $q$  premiers axes sera :

$$qlt_{(1*...*q)}(i) = \sum_{k=1}^q qlt_k(i)$$

en particulier

$$\sum_{k=1}^p qlt_k(i) = 1$$

L'examen de ces quantités met en évidence des éléments qui ne contribuent pas fortement à l'axe mais qui présentent des caractéristiques propres à l'axe. Ces éléments n'expliquent pas l'axe mais sont expliqués par lui. La qualité de représentation permet également de savoir si la proximité de deux éléments sur un plan est réelle et donc traduit une ressemblance où si elle est fortuite et due au caractère déformant de la projection. Ces quantités sont parfois appelées contributions relatives. Il nous semble plus clair de parler de contribution à l'inertie de l'axe pour les contributions (absolues) et de qualité de représentation pour les secondes. Attention toutefois à ne pas confondre les deux notions :

1. Les contributions à l'inertie permettent de déceler quels éléments interviennent dans l'élaboration d'un axe.
2. La qualité de la représentation permet de déceler sur quel(s) axe(s) un élément est bien ou le mieux représenté.

A l'issue de l'interprétation d'un axe quatre types d'éléments peuvent donc apparaître :

1. Les éléments à forte contribution et à forte qualité de représentation :  
Ces éléments interviennent sur l'axe (forte contribution) et sur cet axe seul (forte qualité).
2. Les éléments à forte contribution et à qualité de représentation faible ou moyenne :  
Ils interviennent sur l'axe (CTR forte) et sur un ou plusieurs autres (qlt moyenne).

3. Les éléments à faible contribution et à qualité de représentation forte : Ils n'interviennent pas sur l'axe (CTR faible) mais sont bien représentés sur l'axe (qlt élevé). Ils "n'expliquent" pas l'axe mais sont "expliqués" par l'axe. On dit aussi qu'ils « illustrent » l'axe.
4. Les éléments à faible contribution et à faible qualité de représentation : Ils présentent des caractéristiques qui ne sont pas présentes sur cet axe.

## 2.9 Elément supplémentaire

Dans une enquête toutes les variables ne possèdent pas le même statut. Dans une enquête d'opinion, les variables telles que la profession, l'âge, le sexe forment ce que l'on appelle la signalétique. En règle générale on souhaite voir comment ces variables illustrent différents comportements. Le problème est que leur inclusion directe dans l'analyse risque d'être prégnante et de masquer les phénomènes qui nous intéressent. Comment les inclure de telle manière qu'elles puissent illustrer certains comportements sans participer de manière active à l'élaboration des axes ? Il suffit simplement de les projeter sur les axes mais sans pour autant les faire participer à leur construction.

Définition : On appelle élément supplémentaire, ou illustratif, un individu ou une variable qui est simplement projeté sur les axes factoriels mais qui ne participe pas à leur élaboration.

Remarque : Les individus supplémentaires peuvent par exemple représenter des éléments caractéristiques de sous groupes composant la population étudiée. Ainsi dans une étude sur les producteurs de lait on pourra souhaiter visualiser les caractéristiques des jeunes exploitants âgés de moins de 35 ans par exemple.

Individu supplémentaire :

Notons  $x_{i+}$  le vecteur  $[p, 1]$  des données relatives à l'individu supplémentaire  $i_+$ . Sa coordonnée sur l'axe  $k$  s'écrit :

$$c_{i+k} = \langle x_{i+}, u_k \rangle = {}^t x_{(i+)} u_k$$

Variable supplémentaire :

Soit  $X_{j+}$  le vecteur  $[n, 1]$  des données relatives à la variable supplémentaire  $j^+$ . Alors sa coordonnée sur l'axe  $k$  s'écrit :

$$d_{j+k} = \langle X_{j+}, v_k \rangle = {}^t X_{(j+)} v_k$$

Il est à noter qu'il n'est pas nécessaire de disposer des données relatives aux variables supplémentaires observées sur les individus supplémentaires :

$$\begin{bmatrix} X & X^+ \\ X_+ & \end{bmatrix}$$

De la même manière que pour les éléments actifs il est possible de donner la qualité de la représentation des éléments supplémentaires. Bien entendu un élément supplémentaire ne possède pas de contribution à l'inertie par définition. Il peut être vu comme un élément dont la masse, donc la contribution, est nulle. Tout se passe en fait comme si une fois l'analyse réalisée on souhaitait voir où certains éléments se situent.

## 2.10 Analyse d'un nuage de points. Cas général

Dans un souci pédagogique et afin de ne pas alourdir l'exposé nous avons supposé l'espace muni de la métrique  $M = I$  et les observations assorties de masses égales ( $m_i = 1$  ou  $m_i = 1/n$ ). Nous allons reformuler l'analyse d'un nuage de points au cas où la métrique ambiante  $M$  est quelconque et où les masses  $m_i$  sont inégales. Nous montrerons dans un second temps que l'on peut toujours se ramener au cas initial  $M = I$ .

### 2.10.1 Notation

Soient :

$X = (x_{ij})$  la matrice  $[n, p]$  des données

$M$  la matrice d'ordre  $p$  définissant la métrique dans  $\mathbf{R}^p$

$P$  la matrice diagonale d'ordre  $n$  de terme générique  $m_i$

### 2.10.2 Définitions

$M$  étant une matrice symétrique définie positive, on appelle produit scalaire des vecteurs  $x$  et  $y$  au sens de la métrique  $M$  ou produit  $M$ -scalaire la quantité notée  $\langle x, y \rangle_M$  égale à  ${}^t x M y$ . Le produit scalaire usuel est pris avec la métrique  $M = I$ .

Deux vecteurs  $x$  et  $y$  seront dits  $M$ -orthogonaux si  $\langle x, y \rangle_M = 0$ .

De la même manière on peut définir une  $M$ -norme par  $\|x\|_M^2 = {}^t x M x$ . Un vecteur  $x$  est dit  $M$ -normé ou unitaire pour la norme  $M$  si  $\|x\|_M = 1$ .

Enfin, la distance notée  $d_M(x, y)$  entre deux éléments induite par  $M$  est définie par :  $d_M(x, y) = {}^t(x - y)M(x - y)$ .

### 2.10.3 Analyse directe

Le problème consiste toujours à chercher un axe  $\Delta u_1$ , de vecteur  $u_1$  unitaire pour la norme induite par  $M$  maximisant l'inertie du nuage projeté. Les projections des  $n$  points sur cet axe sont obtenus en utilisant le produit  $M$ -scalaire. Soit  $C_1$  le vecteur  $[n, 1]$  de leurs coordonnées, alors :

$$C_1 = X M u_1$$

L'inertie  $I_1$  du nuage projeté sur  $\Delta u_1$  s'écrit :

$$I_1 = \sum_{i=1}^n m_i c_{ij}^2$$

soit sous forme matricielle :

$$I_1 = {}^t C_1 P C_1$$

soit encore :

$$I_1 = {}^t u_1 M^t X P X M u_1$$

Le problème de maximisation conduit alors à chercher  $u_1$  rendant la quantité  $I_1$  maximale sous la contrainte  ${}^t u_1 M u_1 = 1$ .

Le lagrangien s'écrit :

$$L(u_1) = {}^t u_1 M^t X P X M u_1 - \lambda_1 ({}^t u_1 M u_1 - 1)$$

Les conditions du premier ordre s'écrivent :

$$M^t X P X M u_1 = \lambda_1 M u_1$$

soit encore,  $M$  étant inversible :

$${}^t X P X M u_1 = \lambda_1 u_1 \tag{2.8}$$

$u_1$  apparaît donc comme vecteur propre de la matrice  ${}^t X P X M$  associé à la valeur propre  $\lambda_1$ . En prémultipliant cette relation par  ${}^t u_1$  on voit qu'il suffit de prendre pour  $\lambda_1$  la plus grande valeur propre de la matrice  ${}^t X P X M$ .

### 2.10.4 Analyse duale

Celle-ci demande un petit effort d'imagination. Dans l'analyse directe la matrice des données est  $X_{[n,p]}$ , la métrique  $M_{[p,p]}$  et la matrice des masses  $P_{[n,n]}$ . Afin de retrouver les relations de transition (relations (2.3) et (2.4)) il suffit dans  $\mathbf{R}^n$  de prendre  ${}^tX_{[p,n]}$  comme matrice des données,  $P$  comme métrique et  $M$  comme matrice des masses.

En effet, d'après (2.8), dans  $\mathbf{R}^p$  nous avons :

$${}^tXPXM u_1 = \lambda_1 u_1$$

De manière symétrique dans  $\mathbf{R}^n$  nous aurons :

$$XM {}^tXP v_1 = \mu_1 v_1 \quad (2.9)$$

Il suffit alors de prémultiplier cette dernière égalité par  ${}^tXP$ , soit :

$${}^tXPXM ({}^tXP v_1) = \mu_1 ({}^tXP v_1)$$

${}^tXP v_1$  apparaît comme vecteur propre de la matrice  ${}^tXPXM$  associé à la valeur propre  $\mu_1$ , d'où  $\mu_1 \leq \lambda_1$ .

En procédant de même de même avec (2.9) on en déduit que  $XM u_1$  est vecteur propre de  $XM {}^tXP$  donc que  $\lambda_1 \leq \mu_1$ , d'où finalement  $\mu_1 = \lambda_1$ .

Sa M-norme vérifie :  $\|{}^tXP v_1\|_M^2 = {}^t v_1 P X M {}^tXP v_1 = \lambda_1$ , d'où :

$$u_1 = \frac{1}{\sqrt{\lambda_1}} {}^tXP v_1$$

de même :

$$v_1 = \frac{1}{\sqrt{\lambda_1}} XM u_1$$

Etendues aux coordonnées des éléments et pour l'axe de rang  $k$ , ces relations s'écrivent :

$$C_k = \frac{1}{\sqrt{\lambda_k}} XM D_k$$

et :

$$D_k = \frac{1}{\sqrt{\lambda_k}} {}^tXP C_k$$

### 2.10.5 Formules de reconstitution des données

D'après les relations de transition entre vecteurs :

$$u_k = \frac{1}{\sqrt{\lambda_k}} {}^t X P v_k$$

et :

$$v_k = \frac{1}{\sqrt{\lambda_k}} X M u_k$$

Cette dernière relation entraîne :

$$\sqrt{\lambda_k} v_k = X M u_k$$

D'où en prémultipliant des deux côtés par le transposé de  $u_k$  :

$$\sqrt{\lambda_k} v_k {}^t u_k = X M u_k {}^t u_k$$

et en sommant des deux côtés :

$$\sum_{k=1}^p \sqrt{\lambda_k} v_k {}^t u_k = X M \sum_{k=1}^p u_k {}^t u_k$$

Les vecteurs  $u_k$  formant par ailleurs une base orthonormale pour la métrique  $M$  alors

$$\sum_{k=1}^p u_k {}^t u_k M = I$$

En effet, tout vecteur  $u$  de  $\mathbf{R}^p$  peut s'écrire comme combinaison linéaire unique des  $(u_k)$ , soit :

$$u = \sum_{l=1}^p \alpha_l u_l$$

d'où :

$$\sum_{k=1}^p u_k {}^t u_k M u = \sum_{k=1}^p u_k {}^t u_k M \sum_{l=1}^p \alpha_l u_l = \sum_{k=1}^p u_k \sum_{l=1}^p \alpha_l {}^t u_k M u_l$$

or, les vecteurs  $u_k$  étant normés pour  $M$ ,  $u_k M u_l = 1$  si  $l = k$ , 0 sinon

$$\sum_{k=1}^p u_k {}^t u_k M u = \sum_{k=1}^p \alpha_k u_k = u$$

donc :

$$\sum_{k=1}^p u_k {}^t u_k M = I$$

Par ailleurs, il est immédiat que :

$${}^t\left(\sum_{k=1}^p u_k {}^t u_k M\right) = M \sum_{k=1}^p u_k {}^t u_k$$

Par conséquent :

$$X = \sum_{k=1}^p \sqrt{\lambda_k} v_k {}^t u_k$$

Le tableau  $X$  peut ainsi être reconstitué de manière *exacte* à l'aide des  $\lambda_k$  et des vecteurs propres  $u_k$  et  $v_k$ .

De la même manière il est possible de montrer que cette formule de reconstitution peut être étendue aux coordonnées  $C_k$  et  $D_k$  :

$$X = \sum_{k=1}^p \frac{1}{\sqrt{\lambda_k}} C_k {}^t D_k$$



### 2.10.6 Retour au cas initial

Montrons maintenant que moyennant une transformation des données initiales il est possible de se ramener au cas  $M = I$ . Dans le cas général le programme de maximisation s'écrit :

$$\text{Max}_{\{u_1\}} {}^t u_1 M^t X P X M u_1$$

sous la contrainte

$${}^t u_1 M u_1 = 1$$

La matrice  $M$  étant symétrique et définie positive peut s'écrire sous la forme  $M = {}^t T T$ .

La matrice diagonale  $P$  peut quant à elle s'écrire sous la forme  $P^{1/2} P^{1/2}$ , en notant  $P^{1/2}$  la matrice diagonale de terme  $\sqrt{m_i}$ .

Le programme précédent s'écrit alors :

$$\text{Max}_{\{u_1\}} {}^t u_1 {}^t T T^t X P^{1/2} P^{1/2} X^t T T u_1$$

sous la contrainte

$${}^t u_1 {}^t T T u_1 = 1$$

Effectuons les changements de variable suivants :

$$w_1 = T u_1 \quad \text{et} \quad Y = P^{1/2} X^t T$$

alors

$${}^t w_1 = {}^t (T u_1) = {}^t u_1 {}^t T \quad \text{et} \quad {}^t Y = {}^t (P^{1/2} X^t T) = T X^t P^{1/2}$$

La matrice  $P^{1/2}$  étant diagonale est égale à sa transposée.

Le programme initial s'écrit alors :

$$\text{Max}_{\{w_1\}} {}^t w_1 {}^t Y Y w_1$$

sous la contrainte

$${}^t w_1 w_1 = 1$$

Nous sommes ainsi ramenés au cas où  $M = I$  et  $P = I$  présenté au point 2.3.2.

Les méthodes que nous allons voir se différencient essentiellement par la nature des données analysées (le tableau  $X$ ) qui va conditionner le choix de la métrique  $M$ , sachant qu'il est toujours possible par une transformation des données initiales de se ramener au cas  $M = I$ .

### 2.10.7 Formulaire

	Analyse directe ( $\mathbf{R}^p$ )	Analyse duale ( $\mathbf{R}^n$ )
Données	$X_{[n,p]}$	${}^tX_{[p,n]}$
Métrique	$M_{[p,p]}$	$P_{[n,n]}$
Poids	$P_{[n,n]}$	$M_{[p,p]}$
Axes factoriels	${}^tXPXM u_k = \lambda_k u_k$	$XM {}^tXP v_k = \lambda_k v_k$
Norme	$\ u_k\ _M = 1$	$\ v_k\ _P = 1$
Inertie sur l'axe $k$	$I_k = 1$	$I_k = 1$
Facteur de rang $k$	$C_k = XM u_k$	$D_k = {}^tXP v_k$
Norme du facteur	$\ C_k\ _M = \sqrt{\lambda_k}$	$\ D_k\ _P = \sqrt{\lambda_k}$
Relations de transition		
- entre vecteurs	$u_k = \frac{1}{\sqrt{\lambda_k}} {}^tXP v_k$	$v_k = \frac{1}{\sqrt{\lambda_k}} XM u_k$
- entre coordonnées	$C_k = \frac{1}{\sqrt{\lambda_k}} XM D_k$	$D_k = \frac{1}{\sqrt{\lambda_k}} {}^tXP C_k$

### 2.11 Exemple

Le tableau suivant fournit la structure du bilan d'un groupe pétrolier de 1969 à 1984.

année	NET	INT	SUB	LMT	DCT	IMM	EXP	VRD
1969	17.93	3.96	0.88	7.38	19.86	25.45	5.34	19.21
1970	16.21	3.93	0.94	9.82	19.11	26.58	5.01	18.40
1971	19.01	3.56	1.91	9.43	17.87	25.94	5.40	16.88
1972	18.05	3.33	1.73	9.72	18.83	26.05	5.08	17.21
1973	16.56	3.10	2.14	9.39	20.36	23.95	6.19	18.31
1974	13.09	2.64	2.44	8.10	25.05	19.48	11.61	17.59
1975	13.43	2.42	2.45	10.83	22.07	22.13	11.17	15.49
1976	9.83	2.46	1.79	11.81	24.10	22.39	11.31	16.30
1977	9.46	2.33	2.30	11.46	24.45	23.07	11.16	15.77
1978	10.93	2.95	2.25	10.72	23.16	24.17	9.64	16.20
1979	13.02	3.74	2.21	7.99	23.04	19.53	12.60	17.87
1980	13.43	3.60	2.29	7.09	23.59	17.61	16.67	15.72
1981	13.37	3.35	2.58	6.76	23.94	18.04	15.42	16.54
1982	11.75	2.74	3.11	7.37	25.04	18.11	14.71	17.18
1983	12.59	3.05	3.85	7.12	23.40	19.17	11.86	18.97
1984	13.00	3.00	4.00	7.00	24.00	20.00	12.00	17.00

Les postes du bilan sont les suivant :

NET : situation nette, représente l'ensemble des capitaux propres de l'entreprise.

INT : Intérêts. Ensemble des frais financiers supportés par l'entreprise.

SUB : Subventions. Représente le montant total des subventions accordées par l'Etat.

LMT : Dettes à long et moyen terme.

DCT : Dettes à court terme.

IMM : Immobilisations. Représente l'ensemble des terrains et du matériel de l'entreprise.

EXP : Valeurs d'exploitation.

VRD : Valeurs réalisables et disponibles. Ensemble des créances à court terme de l'entreprise.

Les données ont été ventilées en pourcentage par année, la somme des éléments d'une même ligne vaut 100, de manière à éviter les effets dus à l'inflation. On cherche à répondre aux questions suivantes :

- Quelle a été l'évolution de la structure du bilan sur ces 15 années ?
- Peut-on mettre en évidence plusieurs sous périodes ? Si oui, comment se caractérisent-elles ?

Afin d'apprécier pleinement l'apport de l'Analyse des Données il n'est pas superflu de consacrer quelques instants à l'examen de ce tableau, aux dimensions modestes, à l'aide des outils de la Statistique Descriptive usuelle. Examinons maintenant les résultats de l'Analyse Factorielle du tableau précédent.

### 2.11.1 La sélection du nombre d'axes

EDITION DES VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 8.0000  
SOMME DES VALEURS PROPRES .... 8.0000

HISTOGRAMME DES 8 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT. POURCENT.	POURCENT. CUMULE	
1	4.4904	56.13	56.13	*****
2	2.1332	26.67	82.80	*****
3	.6895	8.62	91.41	*****
4	.4795	5.99	97.41	*****
5	.1552	1.94	99.35	***
6	.0437	.55	99.90	*
7	.0084	.10	100.00	*
8	.0000	.00	100.00	*

Le nuage des 15 années est plongé dans un espace de dimension 8 défini par les postes du bilan (analyse directe). Le premier axe conserve plus de la moitié de l'inertie totale du nuage (56.13 %). Il est en conséquence peu probable qu'il soit du au hasard ! En effet alors que le nombre d'axes a été divisé par 8 l'inertie elle, n'est divisée que par 2. Il existe donc une structuration importante des données qui va se manifester sur le premier axe. Le deuxième axe conserve une part importante, 27 % de l'inertie totale. La chute est importante dès le troisième axe qui ne conserve plus que 8 % de l'inertie totale. Ici on peut par exemple décider de ne retenir que les deux premiers axes. D'une part le plan (1 \* 2) conserve plus de 80 % de l'inertie totale, ce qui peut être considéré comme un bon compromis. Nous disposons en effet d'un espace compréhensible par l'oeil et sans subir une déformation trop prononcée du nuage. On peut de plus remarquer que le pourcentage d'inertie sur l'axe 3 (8.62 %) est inférieur au seuil moyen de  $\frac{1}{8}$  (12.5 %). Si le nuage était l'équivalent d'une sphère dans R8 aucune direction ne serait privilégiée et le taux moyen serait alors de  $\frac{1}{8}$ . Il existe une autre raison justifiant ici le rejet de l'axe 3 mais celle-ci fait appel à la méthode utilisée. Ce point sera abordé dans un chapitre ultérieur.

Remarquons que le scree-test de Catell conduit ici à conserver 4 axes (sur 8).

## 2.11.2 Interprétation des deux premiers axes

Appliquons les deux règles que nous avons vues précédemment : d'abord interpréter séparément les axes, puis pour chaque axe procéder à l'interprétation séparée des nuages.

### Interprétation de l'axe 1 - Le nuage des individus

Utilisons pour ce faire les aides à l'interprétation fournies par le logiciel.

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS SUR LES AXES 1 A 5  
INDIVIDUS ACTIFS

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1969	6.67	15.58	3.49	-1.45	.04	.98	.48	18.1	6.6	.0	13.4	10.1	.78	.13	.00	.06	.02
1970	6.67	13.02	3.46	.10	.23	.90	-.35	17.8	.0	.5	11.3	5.4	.92	.00	.00	.06	.01
1971	6.67	10.66	2.94	.25	.34	-1.34	-.16	12.9	.2	1.1	24.9	1.1	.81	.01	.01	.17	.00
1972	6.67	8.35	2.71	.57	-.01	-.80	.13	10.9	1.0	.0	8.9	.7	.88	.04	.00	.08	.00
1973	6.67	3.99	1.68	-.02	-.99	-.30	.21	4.2	.0	9.4	1.3	1.9	.71	.00	.24	.02	.01
1974	6.67	3.82	-1.52	-.45	-.63	.34	.87	3.4	.6	3.8	1.6	32.3	.60	.05	.10	.03	.20
1975	6.67	5.45	-.95	1.85	.20	-.99	.24	1.3	10.7	.4	13.5	2.6	.17	.63	.01	.18	.01
1976	6.67	8.15	-1.27	2.36	.02	.91	.10	2.4	17.5	.0	11.4	.4	.20	.69	.00	.10	.00
1977	6.67	9.49	-1.74	2.50	-.16	.35	-.09	4.5	19.6	.3	1.7	.3	.32	.66	.00	.01	.00
1978	6.67	3.61	-.53	1.64	.05	.28	-.60	.4	8.4	.0	1.1	15.2	.08	.75	.00	.02	.10
1979	6.67	3.29	-.45	-1.44	.39	.70	-.50	.3	6.5	1.5	6.7	10.6	.06	.63	.05	.15	.07
1980	6.67	8.99	-1.81	-1.37	1.94	-.27	-.08	4.8	5.9	36.4	1.0	.3	.36	.21	.42	.01	.00
1981	6.67	6.71	-1.82	-1.56	.95	-.23	.11	4.9	7.6	8.7	.7	.5	.49	.36	.13	.01	.00
1982	6.67	8.23	-2.63	-1.01	-.43	-.14	.24	10.3	3.2	1.8	.3	2.5	.84	.12	.02	.00	.01
1983	6.67	10.67	-1.58	-1.98	-1.93	-.39	-.61	3.7	12.3	36.1	2.1	16.0	.23	.37	.35	.01	.04

INDIVIDUS ILLUSTRATIFS

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1984	6.67	9.32	-2.05	-1.28	-.96	-1.25	-.57	.0	.0	.0	.0	.0	.45	.18	.10	.17	.03

Sélectionnons les éléments dont la contribution est la plus forte et supérieure à la moyenne, soit ici  $\frac{1}{15} = 6.67\%$ .

Individu	Contribution (%)	Coordonnée	Signe de la coordonnée
1969	18.1	3.49	+
1970	17.8	3.46	+
1971	12.9	2.94	+
1972	10.9	2.71	+
1982	10.3	-2.63	-
	$\Sigma = 70.0$		

Remarquons que l'on peut scinder ces éléments en deux groupes selon le signe de la coordonnée sur l'axe 1. L'axe 1 va donc opposer les années 1969 à 1972 à l'année 1982. En fait il isole les premières. Les coordonnées des autres années sont toutes négatives.

**Interprétation de l'axe 1 - Le nuage des variables**

Procédons de même (la contribution moyenne est ici de 12.5 %)

COORDONNEES DES VARIABLES SUR LES AXES 1 A 5															
VARIABLES	COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE COURT	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
VARIABLES ACTIVES															
NET - NET	.86	-.32	.07	-.35	.18	16.4	5.0	0.6	24.9	20.8	.74	.11	.00	.12	.03
INT - INT	.64	-.62	.35	.13	-.25	9.2	17.9	18.2	3.7	39.5	.41	.38	.13	.02	.06
SUB - SUB	-.76	-.24	-.40	-.40	-.19	13.0	2.7	23.0	34.2	22.6	.58	.06	.16	.16	.04
LMT - LMT	.14	.97	-.07	.05	-.11	0.4	44.4	0.6	0.6	8.3	.02	.95	.00	.00	.01
DCT - DCT	-.95	-.02	-.05	.28	.09	20.1	0.0	0.3	16.5	5.7	.90	.00	.00	.08	.01
IMM - IMM	.86	.48	-.10	.02	-.07	16.5	10.7	1.6	0.1	2.8	.74	.23	.01	.00	.00
EXP - EXP	-.92	-.25	.28	.01	-.02	19.0	2.8	11.6	0.0	0.3	.85	.06	.08	.00	.00
VRD - VRD	.50	-.59	-.55	.31	.00	5.5	16.5	44.1	19.9	0.0	.25	.35	.30	.10	.00

Variable	Contribution (%)	Coordonnée	Signe de la coordonnée
DCT	20.1	-0.95	-
EXP	19.0	-0.92	-
NET	16.4	0.86	+
IMM	16.5	0.86	+
SUB	13.0	-0.76	-
	$\Sigma = 85.0$		

Ici les 4 premières variables ont été retenues. L'axe 1 va donc opposer les postes NET et IMM situés à droite de l'axe (coordonnée positive) aux postes DCT et EXP situés à gauche. L'axe 1 va en conséquence opposer les années 1969 à 1972 (1973) marqués par un poids important dans la structure de leur bilan des postes NET et IMM et un poids faible des postes DCT et EXP, aux autres années qui elles présentent le profil inverse.

Pour illustrer ce résultat nous pouvons retourner aux données d'origine en extrayant du tableau de base les éléments, individus et variables, mis en évidence sur l'axe 1.

Individu	NET	IMM	DCT	EXP
1969	17.93	25.45	19.86	5.34
1970	16.21	26.58	19.11	5.01
1971	19.01	25.94	17.87	5.40
1972	18.05	26.05	18.83	5.08
moyenne	13.85	21.98	22.37	10.32
minimum	9.46	17.61	17.87	5.01
Maximum	19.01	26.58	25.05	16.67
1982	11.75	18.11	25.04	14.71

**Interprétation de l'axe 2**

La démarche est identique. Débutons cette fois par le nuage des variables.

## Le nuage des variables

COORDONNEES DES VARIABLES SUR LES AXES 1 A 5

VARIABLES		COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN -	LIBELLE COURT	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
VARIABLES ACTIVES																
NET -	NET	.86	-.32	.07	-.35	.18	16.4	5.0	0.6	24.9	20.8	.74	.11	.00	.12	.03
INT -	INT	.64	-.62	.35	.13	-.25	9.2	17.9	18.2	3.7	39.5	.41	.38	.13	.02	.06
SUB -	SUB	-.76	-.24	-.40	-.40	-.19	13.0	2.7	23.0	34.2	22.6	.58	.06	.16	.16	.04
LMT -	LMT	.14	.97	-.07	.05	-.11	0.4	44.4	0.6	0.6	8.3	.02	.95	.00	.00	.01
DCT -	DCT	-.95	-.02	-.05	.28	.09	20.1	0.0	0.3	16.5	5.7	.90	.00	.00	.08	.01
IMM -	IMM	.86	.48	-.10	.02	-.07	16.5	10.7	1.6	0.1	2.8	.74	.23	.01	.00	.00
EXP -	EXP	-.92	-.25	.28	.01	-.02	19.0	2.8	11.6	0.0	0.3	.85	.06	.08	.00	.00
VRD -	VRD	.50	-.59	-.55	.31	.00	5.5	16.5	44.1	19.9	0.0	.25	.35	.30	.10	.00

Variable	Contribution (%)	Coordonnée	Signe de la coordonnée
LMT	44.4	0.97	+
VRD	16.5	-0.59	-
INT	17.9	-0.62	-
	$\Sigma = 78.8$		

L'axe 2 oppose l'endettement à long et moyen terme LMT aux postes VRD et INT (en haut de l'axe).

## Le nuage des individus

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS SUR LES AXES 1 A 5  
INDIVIDUS ACTIFS

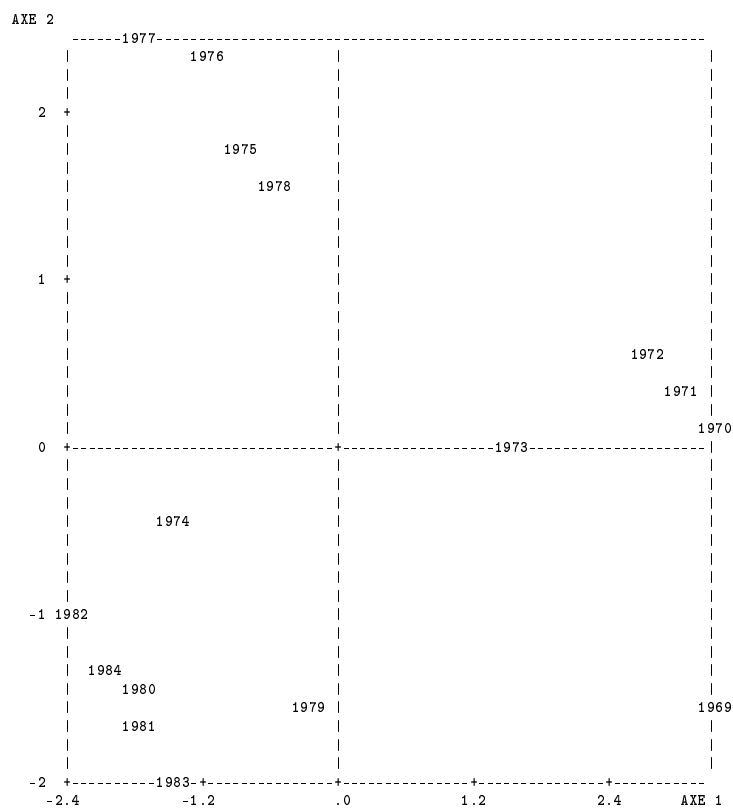
INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1969	6.67	15.58	3.49	-1.45	.04	.98	.48	18.1	6.6	.0	13.4	10.1	.78	.13	.00	.06	.02
1970	6.67	13.02	3.46	.10	.23	.90	-.35	17.8	.0	.5	11.3	5.4	.92	.00	.00	.06	.01
1971	6.67	10.66	2.94	.25	.34	-1.34	-.16	12.9	.2	1.1	24.9	1.1	.81	.01	.01	.17	.00
1972	6.67	8.35	2.71	.57	-.01	-.80	.13	10.9	1.0	.0	8.9	.7	.88	.04	.00	.08	.00
1973	6.67	3.99	1.68	-.02	-.99	-.30	.21	4.2	.0	9.4	1.3	1.9	.71	.00	.24	.02	.01
1974	6.67	3.82	-1.52	-.45	-.63	.34	.87	3.4	.6	3.8	1.6	32.3	.60	.05	.10	.03	.20
1975	6.67	5.45	-.95	1.85	.20	-.99	.24	1.3	10.7	.4	13.5	2.6	.17	.63	.01	.18	.01
1976	6.67	8.15	-1.27	2.36	.02	.91	.10	2.4	17.5	.0	11.4	.4	.20	.69	.00	.10	.00
1977	6.67	9.49	-1.74	2.50	-.16	.35	-.09	4.5	19.6	.3	1.7	.3	.32	.66	.00	.01	.00
1978	6.67	3.61	-.53	1.64	.05	.28	-.60	.4	8.4	.0	1.1	15.2	.08	.75	.00	.02	.10
1979	6.67	3.29	-.45	-1.44	.39	.70	-.50	.3	6.5	1.5	6.7	10.6	.06	.63	.05	.15	.07
1980	6.67	8.99	-1.81	-1.37	1.94	-.27	-.08	4.8	5.9	36.4	1.0	.3	.36	.21	.42	.01	.00
1981	6.67	6.71	-1.82	-1.56	.95	-.23	.11	4.9	7.6	8.7	.7	.5	.49	.36	.13	.01	.00
1982	6.67	8.23	-2.63	-1.01	-.43	-.14	.24	10.3	3.2	1.8	.3	2.5	.84	.12	.02	.00	.01
1983	6.67	10.67	-1.58	-1.98	-1.93	-.39	-.61	3.7	12.3	36.1	2.1	16.0	.23	.37	.35	.01	.04

Individu	Contribution (%)	Coordonnée	Signe de la coordonnée
1977	19.6	2.50	+
1976	17.5	2.36	+
1975	10.7	1.85	+
1983	12.3	-1.91	-
	$\Sigma = 60.1$		

L'axe 2 isole essentiellement les années 1975 à 1977, voire 1978, caractérisées par un poids important du poste LMT et un poids faible des postes VRD et INT. La aussi un retour aux données vient confirmer l'interprétation.

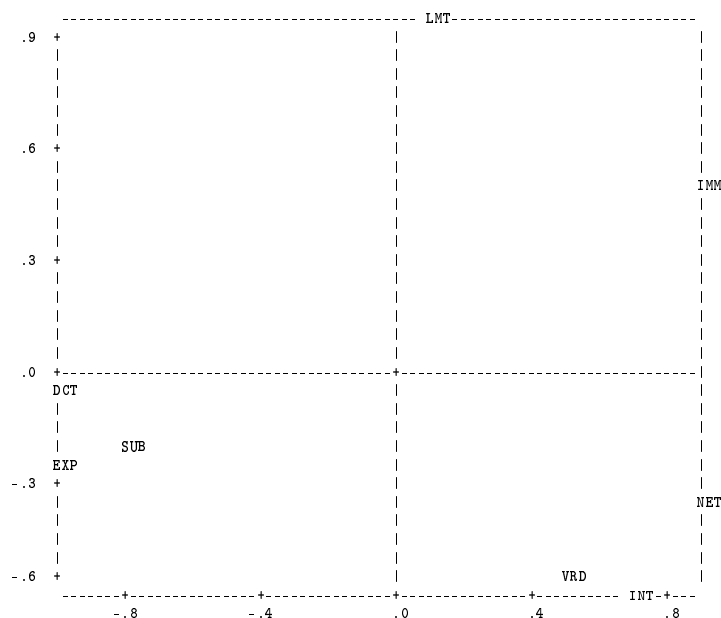
### 2.11.3 Les représentations graphiques

#### Le nuage des individus





### Le nuage des variables



Ce plan fournit un excellent résumé de l'essentiel de l'information contenue dans les données. L'axe 1, horizontal, oppose les années 1969 à 1972 à toutes les autres années. Il met en évidence la rupture majeure induite par le premier choc pétrolier (1973). Le second axe permet de "lire" la stratégie suivie par l'entreprise en réponse à une modification profonde de son environnement. Le groupe s'engage dans une politique d'endettement à long et moyen terme qui prendra fin en 1979 avec le second choc pétrolier. Enfin après 1979 le groupe s'oriente vers une politique d'endettement à court terme.

#### 2.11.4 Conclusion

Sur cet exemple réel, aux dimensions certes modestes, apparaissent déjà les résultats que l'Analyse de Données permet d'obtenir :

- Une mise en évidence des traits majeurs contenus dans les données et qui sont sériés axe par axe.
- L'existence de relations d'attraction ou d'opposition entre variables.
- Une mise en évidence de sous groupes homogènes que l'analyse permet d'interpréter.

Ici schématiquement 3 sous périodes apparaissent :

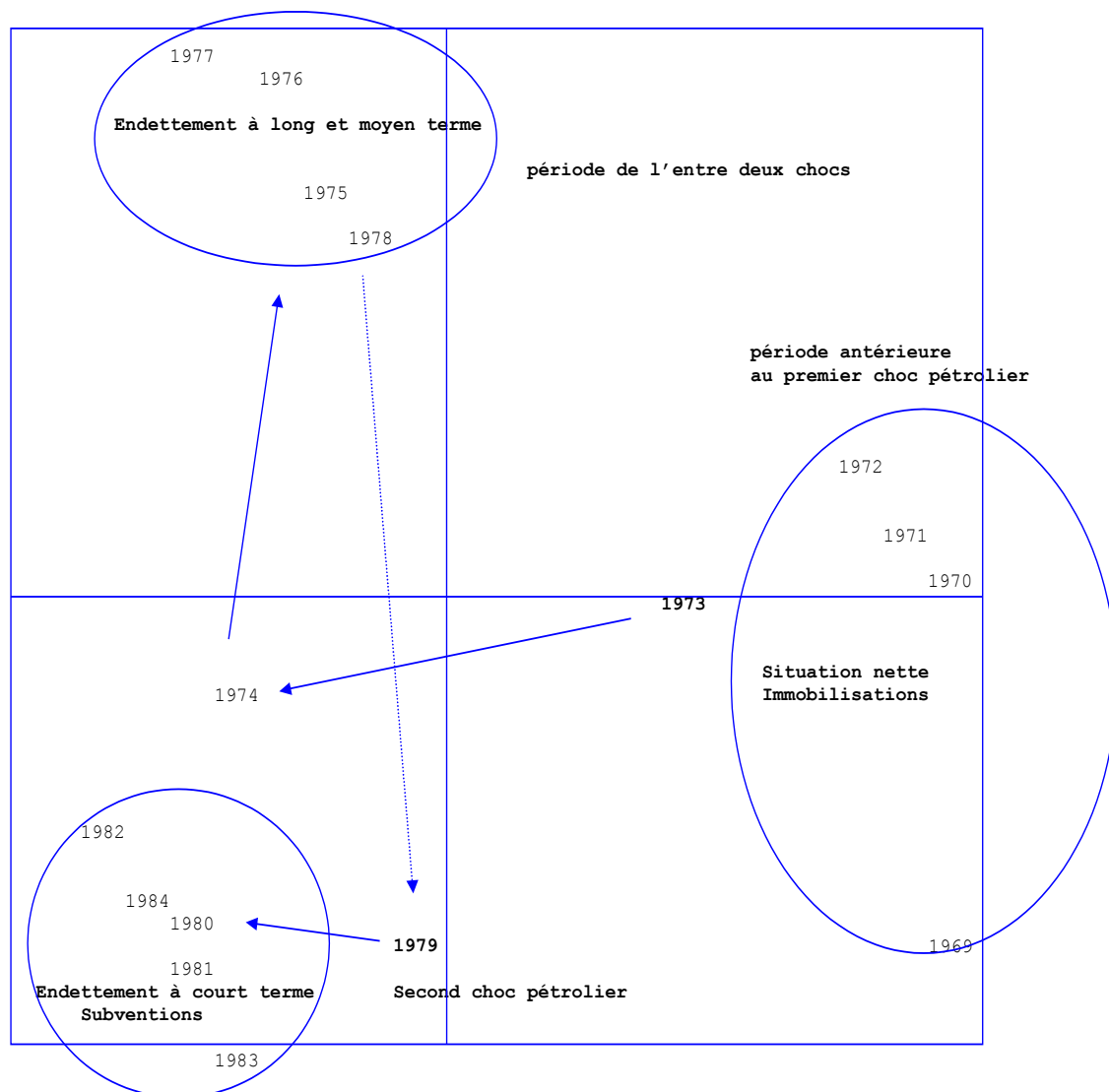
- La période antérieure au premier choc pétrolier 1969-1973. Le bilan est alors marqué par les postes situation nette et immobilisations. En revanche l'endettement à court terme et le poste valeurs d'exploitation sont peu importants.

- La période de l'entre deux chocs, 1975-1978, au cours de laquelle le groupe s'engage dans une politique d'endettement à long et moyen terme.
- La période postérieure au second choc, après 1979, qui voit l'essor de l'endettement à court terme, ainsi que dans une moindre mesure celui des subventions.

Ici la partition peut être opérée à vue sur le plan correspondant dans la mesure où le nombre d'éléments à représenter étant faible, le nuage ne présente pas de continuum de points. Tel n'est pas le cas lorsque le jeu de données est plus important ce qui explique le recours aux techniques de classification. Il sera d'ailleurs intéressant d'observer si ces techniques appliquées à ces données permettent d'observer la même périodisation. Signalons enfin que le plan précédent livré avec les commentaires appropriés constitue un excellent outil de synthèse.

A la page suivante figure un exemple de ce que l'Analyse de Données permet d'obtenir et sous quelle forme le représenter.

Cet exemple n'a d'autre prétention que d'illustrer les propos précédents. Il permet toutefois de souligner les écueils à éviter. En premier lieu livrer des résultats "bruts de fonderie" tout droit sortis de l'ordinateur et sans aucun commentaire. Il importe au contraire de rendre ces représentations graphiques lisibles et de les accompagner d'un texte expliquant à la fois leur obtention et leur signification.



Annexe

Dérivation vectorielle :

Soit  $f$  une application de  $\mathbf{R}^p$  dans  $\mathbf{R}$ ,  $f : u \mapsto f(u)$

Par définition la dérivée de  $f$  par rapport à  $u$  est le vecteur de  $\mathbf{R}$  de terme générique

$$\frac{\partial f}{\partial u_k}$$

Soit  $A$  une matrice carrée d'ordre  $p$  et de terme générique  $a_{ij}$ , alors

$$\frac{\partial(^t u A u)}{\partial u} = A u + ^t A u$$

Démonstration

A étant de terme générique  $a_{ij}$ , il vient :

$$^t u A u = \sum_{i=1}^p \sum_{j=1}^p a_{ij} u_i u_j = \sum_{i=1}^p \left( \sum_{j=1, j \neq k}^p a_{ij} u_i u_j \right) + a_{ik} u_i u_k$$

soit encore

$$^t u A u = \sum_{i=1, i \neq k}^p \left( \sum_{j=1, j \neq k}^p a_{ij} u_i u_j \right) + \sum_{j=1, j \neq k}^p a_{kj} u_k u_j + \sum_{i=1, i \neq k}^p a_{ik} u_i u_k + a_{kk} u_k^2$$

d'où

$$\frac{\partial(^t u A u)}{\partial u} = 0 + \sum_{j=1, j \neq k}^p a_{kj} u_j + \sum_{i=1, i \neq k}^p a_{ik} u_i + 2a_{kk} u_k$$

soit encore

$$\frac{\partial(^t u A u)}{\partial u} = \sum_{j=1}^p a_{kj} u_j + \sum_{i=1}^p a_{ik} u_i$$

Le premier terme correspond au produit de la ligne  $k$  de la matrice  $A$  par le vecteur  $u$  et le second au produit de la transposée de la colonne  $k$  de  $A$  par  $u$ , d'où :

$$\frac{\partial(^t u A u)}{\partial u} = A u + ^t A u$$

Si la matrice  $A$  est symétrique alors  $^t A = A$  et

$$\frac{\partial(^t u A u)}{\partial u} = 2A u$$

# Chapitre 3

## Analyse en Composantes Principales

### 3.1 Introduction

L'Analyse en Composantes Principales ou ACP constitue l'une des plus anciennes méthodes factorielles dont le principe remonte à Hotelling (1933). C'est une technique qui permet d'analyser les tableaux de type individus-variables lorsque ces dernières sont quantitatives. Elle est simple à mettre en oeuvre, en ce sens qu'elle ne requiert aucun codage préalable des données, et de surcroît très répandue, notamment en ce qui concerne son implantation dans la plupart des logiciels statistiques. Ses limites tiennent à ce qu'elle ne permet pas le traitement des variables qualitatives d'une part et en ce qu'elle ne permet de détecter que d'éventuelles liaisons linéaires entre variables d'autre part. La méthode la plus fréquemment utilisée est l'ACP sur données centrées réduites, encore appelée ACP normée.

### 3.2 Le choix de la métrique

Nous avons vu que d'un point de vue formel toute méthode factorielle se résume à se donner un tableau de données  $X[n, p]$ , à munir l'espace d'une métrique  $M[p, p]$  et à assortir les observations de poids contenus dans une matrice  $P[n, n]$ . Soit  $X[n, p] = (x_{ij})$  la matrice des données initiales. Nous allons voir pourquoi il n'est pas toujours souhaitable de travailler sur données brutes ou simplement centrées. Implicitement cette démarche conduit lorsque les variables sont hétérogènes à en privilégier certaines. Considérons l'exemple très simple suivant à des fins pédagogiques. Soient trois variables,

âge de l'exploitant, revenu et surface relevées sur trois exploitations agricoles.

exploitation	AGE	PBT	SAU
1	30	290 000	20
2	50	300 000	30
3	52	320 000	28

Nous avons vu dans le premier chapitre que la recherche des axes principaux pouvait se fonder sur deux critères équivalents, maximisation de l'inertie du nuage projeté ou conservation des distances entre points. Afin d'illustrer le problème posé par l'utilisation directe des données brutes ou centrées, retenons le second critère. Notons en premier lieu que les exploitations 2 et 3 se ressemblent tant par leur structure que par l'âge de l'exploitant qui constitue un critère déterminant. En revanche l'exploitation 1 se distingue des deux autres : elle appartient à un exploitant jeune installé sur une surface faible. Néanmoins le calcul des distances entre ces 3 individus à partir des données brutes ne fait pas ressortir le phénomène précédent. La distance utilisée ici est la distance euclidienne canonique.

d	1	2	3
1	0		
2	$10^4$	0	
3	$3 \cdot 10^4$	$2 \cdot 10^4$	0

L'exploitation 2 apparaît plus proche de 1 que de 3. Cet apparent paradoxe tient à l'importance écrasante prise par la variable revenu dans le calcul des distances. Une variation de 10 % du revenu n'a pas la même incidence qu'une variation équivalente des 2 autres variables. De plus en cas de changement d'échelle de mesure les résultats ne sont pas stables. Mesurons par exemple le revenu non plus en francs mais en dizaines de milliers de francs. Le tableau précédent se modifie de la manière suivante :

d	1	2	3
1	0		
2	22.4	0	
3	23.6	3.46	0

Désormais 2 apparaît plus proche de 3 que de 1. Toutefois le choix de l'unité de mesure est parfaitement arbitraire et si le revenu avait été mesuré en milliers de francs, le résultat initial aurait été conservé. Afin d'éviter ce double inconvénient, hétérogénéité des variables et choix arbitraire de l'échelle de mesure, on décide de travailler sur des données centrées réduites. Cette transformation accorde un poids identique aux variables dans le calcul des distances. Les données sont maintenant sans dimension et l'écart à la moyenne est mesuré en nombre d'écarts-type.

d	1	2	3
1	0		
2	3.2	0	
3	3.8	1.7	0

Le problème précédent aurait pu être mis en évidence de la même manière si au lieu de calculer les distances entre points pris deux à deux on avait calculé l'inertie totale du nuage.

### 3.3 Analyse directe - Formalisation

Notons  $x_i$  le vecteur de  $\mathbf{R}^p$  contenant les observations relatives à l'individu  $i$ . Les termes de  $x_i$  forment la ligne  $i$  de la matrice  $X$ . Soit  $X_j$  le vecteur  $[n, 1]$  contenant les données relatives à la variable  $j$ , soit la colonne  $j$  de la matrice  $X$ . Les données initiales  $x_{ij}$  ont été transformées en  $(x_{ij} - m_j)/\sigma_j$  où  $m_j$  désigne la moyenne de la variable  $j$  et  $\sigma_j$  l'écart-type de cette variable. Ces données transformées seront encore notées  $x_{ij}$ .

Remarque : Formellement, eu égard à la présentation générale du chapitre précédent, nous pourrions définir l'ACP comme l'analyse factorielle du triplet  $(X, M, P)$  avec  $X = \frac{(x_{ij} - m_j)}{\sigma_j}$ ,  $M = I$  et  $P = (p_i)$ .  $I$  étant la matrice identité d'ordre  $p$  et  $P$  la matrice diagonale de terme  $p_i$  avec

$$\sum_{i=1}^n p_i = 1$$

Nous avons vu que l'Analyse Factorielle d'un tel nuage conduit à rechercher les vecteurs propres  $u_j$  de la matrice d'inertie  ${}^tX P X$ . Soit  $u_j$  vérifiant :

$${}^tX P X u_j = \lambda_j u_j \quad (j = 1, p)$$

## 3.4 Propriétés

### 3.4.1 Matrice d'inertie en ACP normée

Le terme  $(j, k)$  de la matrice  ${}^tXPX$  carrée d'ordre  $p$  est égal au produit de la ligne  $j$  de  ${}^tX$ , soit la colonne  $j$  de  $X$ , par la colonne  $k$  de  $PX$ , soit :

$$({}^tXPX)_{jk} = \sum_{i=1}^n p_i \frac{(x_{ij} - m_j)}{\sigma_j} \frac{(x_{ik} - m_k)}{\sigma_k}$$

d'où :

$$({}^tXPX)_{jk} = \frac{\sum_{i=1}^n p_i [(x_{ij} - m_j)(x_{ik} - m_k)]}{\sigma_j \sigma_k}$$

Ce terme n'est autre que le coefficient de corrélation linéaire entre les variables  $X_j$  et  $X_k$ . Ainsi, en ACP normée la matrice d'inertie se confond avec la matrice des corrélations entre variables actives.

### 3.4.2 Inertie totale en ACP normée

Nous avons vu que la trace de la matrice d'inertie était égale à l'inertie totale du nuage  $I_0$ . En conséquence en ACP normée l'inertie totale du nuage est égale au nombre de variables actives, soit  $I_0 = p$ . En ACP « non-normée », c'est à dire sur des données uniquement centrées, la matrice d'inertie correspond alors à la matrice de varianec-covariance entre variables actives, qui ne présente pas d'intérêt particulier.

## 3.5 Le nuage des individus

### 3.5.1 Composante principale

La coordonnée de l'individu  $i$  sur l'axe de rang  $k$  est égale à  $\langle x_i, u_k \rangle$ , notée  $c_{ik}$ . Soit encore sous forme matricielle  $C_k = Xu_k$ .  $C_k$  est un vecteur  $[n, 1]$ ,  $X$  une matrice  $[n, p]$  et  $u_k$  un vecteur  $[p, 1]$ . Cette dernière égalité peut encore s'écrire :

$$C_k = \sum_{j=1}^p u_{jk} X_j$$

où  $u_{jk}$  désigne la coordonnée de rang  $j$  du vecteur-propre d'ordre  $k$ .



Cette expression montre que la variable  $C_k$  s'exprime comme combinaison linéaire des variables d'origine  $X_j$ .  $C_k$  est appelée composante principale de rang  $k$ .

### 3.5.2 Propriétés des composantes principales

1 - En premier lieu remarquons que les variables  $X_j$  étant centrées,  $C_k$  étant combinaison linéaire des  $X_j$ , il est immédiat que  $C_k$  est également centrée. La démonstration est immédiate. En effet :

$$\bar{C}_k = \sum_{i=1}^n p_i c_{ik}$$

or,

$$C_k = \sum_{j=1}^p u_{jk} X_j \Rightarrow c_{ik} = \sum_{j=1}^p u_{jk} \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

d'où

$$\bar{C}_k = \sum_{i=1}^n p_i \sum_{j=1}^p u_{jk} \frac{x_{ij} - \bar{x}_j}{\sigma_j} = \sum_{j=1}^p \frac{u_{jk}}{\sigma_j} \sum_{i=1}^n p_i (x_{ij} - \bar{x}_j)$$

et

$$\sum_{i=1}^n p_i (x_{ij} - \bar{x}_j) = \sum_{i=1}^n p_i x_{ij} - \bar{x}_j \sum_{i=1}^n p_i = 0$$

car

$$\sum_{i=1}^n p_i x_{ij} = \bar{x}_j \quad \text{et} \quad \sum_{i=1}^n p_i = 1$$

2 - Cherchons la variance de  $C_k$ .  $C_k$  étant centrée, sa variance s'écrit :

$$V(C_k) = \sum_{i=1}^n p_i c_{ik}^2 = {}^t C_k P C_k = {}^t u_k {}^t X P X u_k = \lambda_k = I_k$$

La variance de la composante principale de rang  $k$  est donc égale à l'inertie du nuage projeté sur l'axe de rang  $k$ . Notons que la variance de  $C_k$  coïncide avec sa norme. En effet la métrique dans  $\mathbf{R}^n$  étant définie par  $P = I$  il vient :  $\|C_k\|_P^2 = {}^t C_k C_k = V(C_k)$ . Chacun des axes  $u_k$  étant obtenu comme l'axe conservant le maximum d'inertie, au rang  $k$ , les  $C_k$  sont donc de variance maximale (au rang  $k$ ).

## 3 - Corrélation entre deux composantes principales

Cherchons maintenant la corrélation entre deux composantes principales de rangs différents.

$$\text{cor}(C_j, C_k) = \frac{\text{cov}(C_j, C_k)}{[V(C_j) \cdot V(C_k)]^{1/2}}$$

Les composantes principales étant centrées :

$$\text{cov}(C_j, C_k) = \sum_{i=1}^n p_i c_{ij} c_{ik}$$

soit encore :

$${}^t C_j P C_k = {}^t (X u_j) P (X u_k) = {}^t u_j {}^t X P X u_k$$

Or  ${}^t X P X u_k = \lambda_k u_k$  d'où  $\text{cov}(C_j, C_k) = \lambda_k {}^t u_j u_k$

Les vecteurs propres étant orthogonaux par construction  ${}^t u_j u_k = 0$ , d'où :

$$\text{cor}(C_j, C_k) = 0 \quad \text{si } j \neq k$$

Les composantes principales sont donc non corrélées deux à deux.

Ainsi l'ACP normée remplace les variables d'origine  $X_j$  par de nouvelles variables  $C_k$  appelées composantes principales, de variance maximale, non corrélées deux à deux et qui s'expriment comme combinaison linéaire des variables d'origine.

L'ACP peut d'ailleurs être présentée sous cet angle : étant données  $p$  variables quantitatives  $X_j$ , comment obtenir de nouvelles variables, qui s'expriment comme combinaison linéaire des variables  $X_j$ , non-corrélées deux à deux et de variance maximale.

L'ACP normée peut être vue comme une méthode d'analyse d'une matrice de corrélations. En effet, cette technique va permettre de décomposer une matrice de corrélations en blocs homogènes. Chaque « bloc » de corrélations correspondant à un axe de l'analyse.

Remarque : Les composantes principales étant des variables de synthèse, puisqu'elles s'expriment comme combinaisons linéaires des variables d'origine et étant de variance maximale, on peut exiger que leur variance soit précisément supérieure à celle des variables d'origine, soit  $\text{Var}(C_k) > \text{Var}(X_j)$ . Or comme  $\text{Var}(X_j) = 1$ , cela revient à sélectionner les composantes, donc les

axes factoriels, pour lesquels  $Var(C_k) = \lambda_k > 1$ . Par ailleurs, l'inertie totale en ACP normée étant égale au nombre de variables actives  $p$ , l'inertie « moyenne » vaut donc  $\frac{I_0}{p} = 1$ . Par conséquent, seuls les axes pour lesquels l'inertie est supérieure à la moyenne sont retenus. Il faut néanmoins bien voir ici que le critère initial revient à sélectionner les axes pour lesquels l'inertie conservée est supérieure à la moyenne mais pour une autre raison.

### 3.5.3 Qualité de la représentation

Le cosinus carré de l'angle formé par l'individu  $i$  et l'axe  $k$  s'écrit :

$$\cos^2 \theta_{ik} = \frac{(\langle x_i, u_k \rangle)^2}{(\|x_i\| \cdot \|u_k\|)^2}$$

soit :

$$\cos^2 \theta_{ik} = \frac{c_{ik}^2}{\|x_i\|^2}$$

car les vecteurs  $u_k$  sont normés par construction.

### 3.5.4 Contributions

Rappelons que cette quantité n'est autre que la part de l'inertie conservée par l'axe  $k$  due à l'individu  $i$ . Nous savons que

$$\sum_{i=1}^n p_i c_{ik}^2 = \lambda_k$$

d'où :

$$CTR_k(i) = \frac{p_i c_{ik}^2}{\lambda_k}$$

## 3.6 Le nuage des variables

### 3.6.1 Coordonnée d'une variable $X_j$ sur l'axe $k$

D'après le chapitre précédent, le nuage des variables conduit ici à l'analyse du triplet  $({}^tX, P, I)$ . Dans ces conditions, les coordonnées des variables sur l'axe de rang  $k$  sont définies par :

$$D_k = {}^tX P v_k$$

et d'après les relations de transition

$$v_k = \frac{1}{\sqrt{\lambda_k}} X u_k$$

d'où

$$D_k = \frac{1}{\sqrt{\lambda_k}} {}^t X P X u_k = \frac{1}{\sqrt{\lambda_k}} \lambda_k u_k = \sqrt{\lambda_k} u_k$$

Dans ces conditions la coordonnée de la variable  $X_j$  sur l'axe de rang  $k$  s'écrit :

$$d_{jk} = \sqrt{\lambda_k} u_{jk}$$

Par ailleurs, les variables  $X_j$  sont normées (dans  $\mathbf{R}^n$  muni de la métrique  $P$ ). En effet :

$$\|X_j\|_P^2 = {}^t X_j P X_j = \sum_{i=1}^n p_i \left( \frac{x_{ij} - \bar{x}_j}{\sigma_j} \right)^2 = V(X_j) = 1$$

car les variables sont centrées réduites. Ainsi en ACP normée les variables se trouvent sur une hypersphère de rayon unité (dans  $\mathbf{R}^n$ ). En projection sur des plans passant par l'origine les variables se trouvent en conséquence à l'intérieur d'un cercle de rayon unité appelé cercle des corrélations pour une raison que nous allons préciser.

Cherchons la corrélation entre une composante principale  $C_k$  et une variable  $X_j$  :

$$\text{cor}(C_k, X_j) = \frac{{}^t X_j P C_k}{\|X_j\|_P \cdot \|C_k\|_P}$$

Sachant que  $\|C_k\|_P = \lambda_k$ ,  $\|X_j\|_P = 1$  et  $C_k = X u_k$  on en déduit que :

$$\text{cor}(C_k, X_j) = \frac{1}{\sqrt{\lambda_k}} {}^t X_j P X u_k$$

or

$${}^t X_j P X u_k = ({}^t X P X)_j u_k = \lambda_k u_{jk}$$

où  $({}^t X P X)_j$  désigne la ligne  $j$  de la matrice  ${}^t X X$ .

Donc :

$$\text{cor}(C_k, X_j) = \sqrt{\lambda_k} u_{jk}$$

Or cette quantité n'est autre que la coordonnée de la variable  $X_j$  sur l'axe de rang  $k$ .

Ainsi la coordonnée d'une variable sur un axe factoriel n'est autre que la corrélation entre cette variable et la composante principale de rang correspondant.

### 3.6.2 Qualité de la représentation

D'après la définition :

$$\cos_{jk}^2 = \frac{d_{jk}^2}{(\|X_j\|_P \|v_k\|_P)^2} = d_{jk}^2$$

La qualité de la représentation d'une variable sur un axe est égale au carré de la coordonnée de cette variable sur l'axe.

Conséquence pratique : une variable est d'autant mieux représentée sur un plan factoriel qu'elle est proche du bord du cercle des corrélations.

Remarque : Le nuage des variables se trouvant à l'intérieur d'un cercle de rayon unité il est intéressant de conserver la même échelle  $[-1, 1]$  sur l'ensemble des représentations graphiques. On observe ainsi au fur et à mesure que l'on progresse sur les axes une contraction du nuage des variables traduisant leur aptitude décroissante à représenter le nuage.

### 3.6.3 Contributions

Les variables étant assorties de masses égales :

$$CTR_k(j) = \frac{d_{jk}^2}{\lambda_k} = \frac{(\sqrt{\lambda_k} u_{jk})^2}{\lambda_k} = u_{jk}^2$$

La contribution d'une variable à l'inertie d'un axe est égale au carré de la coordonnée correspondante du vecteur propre. Les masses afférentes aux variables étant égales, l'interprétation directe des résultats à partir des représentations graphiques est ici licite bien qu'elle constitue une mauvaise habitude. Cette propriété peut d'ailleurs permettre le calcul des contributions lorsque celles-ci ne sont pas éditées (cas du logiciel SPAD par exemple).

### 3.6.4 Corrélations entre variables

Dans l'espace des variables ( $\mathbf{R}^n$  muni de la métrique  $P$ ) la projection d'une variable  $X_j$  sur une variable  $X_k$  mesure la corrélation entre ces 2 variables.

Soit en effet  $l_{jk}$  la longueur de cette projection.

Alors :

$$l_{jk} = \frac{\langle X_j, X_k \rangle_P}{\|X_k\|_P} = {}^t X_j P X_k$$

soit encore :

$${}^tX_jPX_k = \sum_{i=1}^n p_i \frac{(x_{ij} - m_j)}{\sigma_j} \frac{(x_{ik} - m_k)}{\sigma_k} = \text{cor}(X_j, X_k)$$

Cette propriété va permettre une lecture directe des corrélations sur les graphiques. Si la variable  $X_k$  est bien représentée sur le plan factoriel considéré, donc proche du bord du cercle, la projection de  $X_j$  sur cette variable donne une approximation de la corrélation entre les deux variables.

On remarque que la corrélation coïncide ici avec le cosinus de l'angle formé par ces variables.

Conséquences :

- . deux variables proches et bien représentées sont corrélées positivement
- . deux variables qui s'opposent sont corrélées de manière négative
- . deux variables orthogonales sont non corrélées

Ce résultat peut être obtenu à partir du calcul de la distance entre les variables dans  $\mathbf{R}^n$  muni de la métrique  $P$ .

En effet :

$$d^2(X_j, X_k) = {}^t(X_k - X_j)P(X_k - X_j) = {}^tX_kPX_k - 2{}^tX_kPX_j + {}^tX_jPX_j$$

soit encore, les variables étant normées :

$$d^2(X_j, X_k) = 2(1 - {}^tX_kPX_j) = 2(1 - \text{cor}(X_k, X_j))$$

Si les deux variables sont fortement corrélées positivement alors  $d(X_k, X_j) = 0$  ; elles seront donc proches. En revanche si elles sont faiblement corrélées, la distance vaut  $\sqrt{2}$  et les variables sont à distance « moyenne ».

### 3.6.5 Une propriété spécifique de l'ACP normée

Signalons pour clore ce point une propriété particulière de l'ACP. La transformation subie par les données n'a pas la même signification dans les deux espaces. Dans  $\mathbf{R}^p$  le centrage du nuage correspond à une translation du barycentre à l'origine. La conséquence en est que l'individu « moyen », au sens statistique du terme, se trouve projeté à l'origine des axes. Les individus vont donc se répartir des deux côtés des axes. En revanche dans  $\mathbf{R}^n$  cette transformation correspond à une projection parallèlement à la première bissectrice.

En effet, il est immédiat que dans  $\mathbf{R}^n$  muni de la métrique  $P = (p_i)$  le vecteur directeur de la première bissectrice, noté  $\mathbf{1}_n$ , est normé :

$$\|\mathbf{1}_n\|_{P=(p_i)}^2 = {}^t\mathbf{1}_n P \mathbf{1}_n = 1$$

car

$$P\mathbf{1}_n = \begin{pmatrix} \ddots & & 0 \\ & p_i & \\ 0 & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ p_i \\ \vdots \end{pmatrix}$$

d'où

$${}^t\mathbf{1}_n P \mathbf{1}_n = (\dots 1 \dots) \begin{pmatrix} \vdots \\ p_i \\ \vdots \end{pmatrix} = \sum_{i=1}^n p_i = 1$$

Dans ces conditions, la longueur de la coordonnée de la projection du vecteur  $X_j$  de coordonnée  $(x_{ij})$  sur le vecteur  $\mathbf{1}_n$  est obtenu par le produit scalaire de  $X_j$  par  $\mathbf{1}_n$  avec la métrique  $P = (p_i)$ , soit :

$${}^tX_j P \mathbf{1}_n = (\dots x_{ij} \dots) \begin{pmatrix} \vdots \\ p_i \\ \vdots \end{pmatrix} = \sum_{i=1}^n p_i x_{ij} = \bar{x}_j$$

La transformation  $x_{ij} \rightarrow (x_{ij} - \bar{x}_j)$  correspond bien à une projection sur le sous-espace de  $\mathbf{R}^n$ , noté  $\mathbf{1}_n^\perp$ , orthogonal au sens de la métrique  $P$  à  $\mathbf{1}_n$  (donc parallèlement au vecteur  $\mathbf{1}_n$ ).

Le nuage des variables n'est pas centré ce qui explique que dans certains cas l'ensemble des variables puisse se retrouver d'un même côté d'un axe. Cette configuration se produit lorsque toutes les variables actives sont corrélées de manière positive. En effet, d'après le théorème de Perron-Frobenius une matrice carrée à coefficients réels strictement positifs admet une valeur propre maximale réelle strictement positive et un vecteur propre associé dont toutes les coordonnées sont non nulles et de même signe. Une telle structure traduit alors un phénomène appelé « effet taille ». L'axe correspondant peut s'interpréter comme un gradient opposant les éléments « faibles » aux éléments « forts ». Les individus vont alors se répartir sur cet axe selon les valeurs prises par l'ensemble des variables.

### 3.7 Eléments supplémentaires

Ces éléments sont, rappelons-le, simplement projetés sur les axes mais ne participent pas à leur élaboration.

### 3.7.1 Indivudus supplémentaires

Soit  $x_{i+}$  le vecteur des coordonnées de l'individu supplémentaire  $i+$  dans  $\mathbf{R}^p$ . On commence par faire subir à cet élément les mêmes transformations, soit :

$$x_{i+j} \rightarrow \frac{x_{i+j} - m_j}{s_j}$$

Attention dans le calcul de  $m_j$  et  $s_j$  seuls les éléments actifs interviennent.

La coordonnée sur l'axe  $j$  s'obtient en utilisant l'opérateur de projection, soit :

$$c_{i+j} = {}^t x_{i+} u_j$$

La qualité de la représentation sur l'axe  $j$  s'écrit :

$$\cos^2 \theta_{i+j} = \frac{c_{i+j}^2}{\|x_{i+}\|^2}$$

Rappelons que par définition un élément supplémentaire possède une contribution à l'inertie de l'axe nulle.

### 3.7.2 Variables supplémentaires

Les formules sont identiques à celles obtenues pour les éléments actifs.

## 3.8 Exemple

Afin d'illustrer les propriétés énoncées dans les points précédents nous allons reprendre l'exemple du premier chapitre. Le tableau comporte 16 individus dont un traité en élément supplémentaire (1984) et 8 variables. Ces données ont été traitées par l'ACP normée.

### 3.8.1 Les axes

On peut vérifier en premier lieu que l'inertie totale du nuage  $I_0$  est bien égale au nombre de variables actives. Le calcul de la trace de la matrice avant et après diagonalisation permet d'avoir une idée de la précision des calculs dans la phase d'extraction des valeurs propres.



## EDITION DES VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 8.0000  
 SOMME DES VALEURS PROPRES .... 8.0000

## HISTOGRAMME DES 8 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	4.4904	56.13	56.13	*****
2	2.1332	26.67	82.80	*****
3	.6895	8.62	91.41	*****
4	.4795	5.99	97.41	*****
5	.1552	1.94	99.35	**
6	.0437	.55	99.90	*
7	.0084	.10	100.00	*
8	.0000	.00	100.00	

## INTERVALLES LAPLACIENS D'ANDERSON AU SEUIL 0.95

NUMERO	BORNE INFERIEURE	VALEUR PROPRE	BORNE SUPERIEURE
1	2.1407	4.4904	9.4193
2	1.0170	2.1332	4.4747
3	.3287	.6895	1.4463
4	.2286	.4795	1.0058
5	.0740	.1552	.3256

## ETENDUE ET POSITION RELATIVE DES INTERVALLES

```

1 . . . . . *-----*
2 . . . *-----*
3 *-----*
4 *-----*
5 **--*

```

Les intervalles laplaciens d'Anderson fournissent un intervalle de confiance à 95 % pour les valeurs propres. Ainsi que nous l'avions souligné ces informations sont en pratique peu voire pas utilisées. Le principe en est le suivant : deux valeurs propres consécutives dont les intervalles se chevauchent ne peuvent être considérées comme distinctes. En partant des dernières valeurs propres, qui correspondent à des axes non significatifs (i.e. n'apportant aucune information), on peut donc remonter jusqu'à observer une absence de recouvrement et ne retenir que les axes correspondants. Il apparaît immédiatement que l'application de cette démarche sur l'exemple traité conduit à rejeter tous les axes y compris le premier ce qui semble paradoxal dans la mesure où il conserve plus de la moitié de l'inertie totale. Le problème ici tient d'une part à la taille réduite du nombre d'observations ( $n = 15$ ), il ne faut pas oublier que les résultats d'Anderson ont été établis de manière asymptotique, et que d'autre part le test n'est probablement pas très robuste et que l'on s'éloigne ici de l'hypothèse sous-jacente de normalité. Il convient toutefois de souligner qu'à l'expérience, sur des jeux de données plus importants (plusieurs milliers d'observations), ce test semble plus fiable.

Il existe une méthode originale de sélection des axes qui peut être appliquée ici mais qui malheureusement n'est disponible dans aucun logiciel (à notre

connaissance). La méthode du bootstrap<sup>1</sup> constitue une approche non paramétrique du problème de sélection des axes et d'une manière plus générale de validation des résultats en Analyse Factorielle. L'idée de base du bootstrap, que l'on peut qualifier de méthode de rééchantillonnage, consiste à observer le comportement des variables d'intérêt, ici les valeurs propres, en opérant des tirages avec remise dans l'échantillon de base. Dans le cas de l'ACP cette technique consiste à tirer  $n$  individus avec remise et à réaliser sur chacun des échantillons ainsi obtenus une nouvelle ACP<sup>2</sup>. Au bout de  $N$  tirages on obtiendra ainsi  $N$  valeurs pour chaque valeur propre. Il est alors possible de construire un intervalle de confiance pour chaque valeur propre par observation directe. Les travaux menés par Efron, à qui revient la paternité du bootstrap, montrent que les intervalles de confiance ainsi obtenus coïncident en général avec ceux de la distribution réelle.

Cette méthode a été appliquée à l'exemple traité. Le nombre de tirages effectué est de  $N = 200$ . Les valeurs ainsi obtenues sont ensuite classées par ordre croissant. L'intervalle de confiance au seuil désiré peut alors être obtenu de manière directe. Ainsi dans le cas usuel de 95 % il suffit d'observer la valeur qui sépare les 5 plus petites réalisations (borne inférieure) et celle qui correspondant aux 5 plus grandes (borne supérieure).

Dans le tableau suivant figurent pour chacun des huit axes la moyenne des 200 réalisations de chaque valeur propre ainsi que l'intervalle correspondant à 95 %.

Axe	1	2	3	4	5	6	7	8
Sup	5.5921	2.9358	1.2780	0.6461	0.2455	0.0557	0.0097	0.0000
Inf	3.6421	1.4987	0.3507	0.1473	0.0217	0.0019	0.0000	0.0000
Moy	4.6860	2.1080	0.7146	0.3512	0.1133	0.0240	0.0030	0.0000
$\lambda_k$	4.4904	2.1332	0.6895	0.4795	0.1552	0.0437	0.0084	0.0000

1. Voir par exemple :

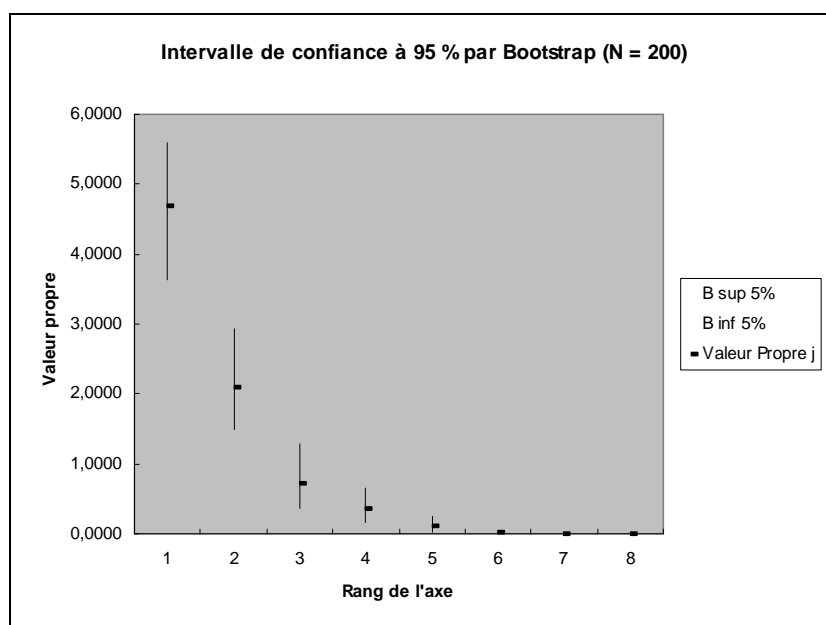
B. EFRON (1979). in « Bootstrap methods : another look at the Jackknife » Ann. Statist. 7, 1-26.

B. EFRON (1982) in « The Jackknife, the bootstrap and other resampling plans » Conference board of the Mathematical Sciences Monograph 38, Society for Industrial and Applied Mathematics.

2. La méthode utilisée ici figure dans l'article « A comparison of principal components from real and random data » D. STAUFFER, E.O. GARTON and R.K. STEINHORST - Ecology, vol. 66, N°6, 1985, pp. 1693-1698.

Une autre approche du même problème figure dans « Stability of Principal Components Analysis studied by the Bootstrap method » J.J. DAUDIN, C. DUBY and P. TRE COURT - Statistics 19 (1988) 2, pp 241-258.

Sur la dernière ligne sont rappelées les valeurs propres obtenues dans l'analyse du tableau d'origine. Il apparaît immédiatement que le recouvrement commence à partir du troisième axe, la borne inférieure obtenue pour la troisième composante (0.3507) étant inférieure à la borne supérieure de la quatrième (0.6461). En revanche il n'y a pas recouvrement entre la seconde et la troisième ou entre la seconde et la première. Cette démarche conduit donc à ne retenir que les deux premiers axes ce qui correspond bien à ce que l'approche empirique a suggéré.



La validation par bootstrap peut s'appliquer de manière générale quelle que soit la méthode, ACP, Analyse des Correspondances ou Analyse des Correspondances Multiples.

Par ailleurs, une méthode a été proposée pour tester si les valeurs propres sont significativement différentes entre elles à partir d'un certain rang. La statistique du test se fonde sur le rapport entre la moyenne arithmétique des dernières valeurs propres et leur moyenne géométrique. Sous l'hypothèse  $H_0$

d'égalité des dernières valeurs propres, la quantité :

$$Q = \left( n - \frac{2p+11}{6} \right) k \ln \left[ \frac{\frac{1}{k} \sum_{j=p-k+1}^p \lambda_k}{\left( \prod_{j=p-k+1}^p \lambda_k \right)^{\frac{1}{k}}} \right]$$

suit une loi du  $\chi^2$  à  $\frac{k(k+1)}{2} - 1$  degrés de liberté. Si  $Q > \chi^2$  alors on rejette  $H_0$ . L'application de ce test ne semble pas donner sur cet exemple des résultats probants.

Signalons que dans le cas de l'ACP normée une autre règle peut être adoptée. On peut ici décider de ne retenir que les deux premiers axes, dans la mesure où l'inertie qui leur est associée est supérieure à celle des variables initiales. Il suffit pour cela de prendre les axes dont la valeur propre correspondante est supérieure à 1 (il y a dans le cas de l'ACP normée identité entre variance et inertie). Cette méthode peut dans certains cas être prise en défaut notamment lorsqu'il y a un effet taille sur le premier axe.

### 3.8.2 Le nuage des individus

En ACP celui-ci ne possède pas de propriété particulière. On trouvera éditées l'ensemble des éléments nécessaires à l'interprétation des résultats, à savoir coordonnées, contributions et qualités de représentation.

### 3.8.3 Le nuage des variables

Les sorties ci-après diffèrent de celles de la première partie qui avaient été reprises par rapport aux sorties brutes du logiciel SPAD.

COORDONNEES DES VARIABLES SUR LES AXES 1 A 5

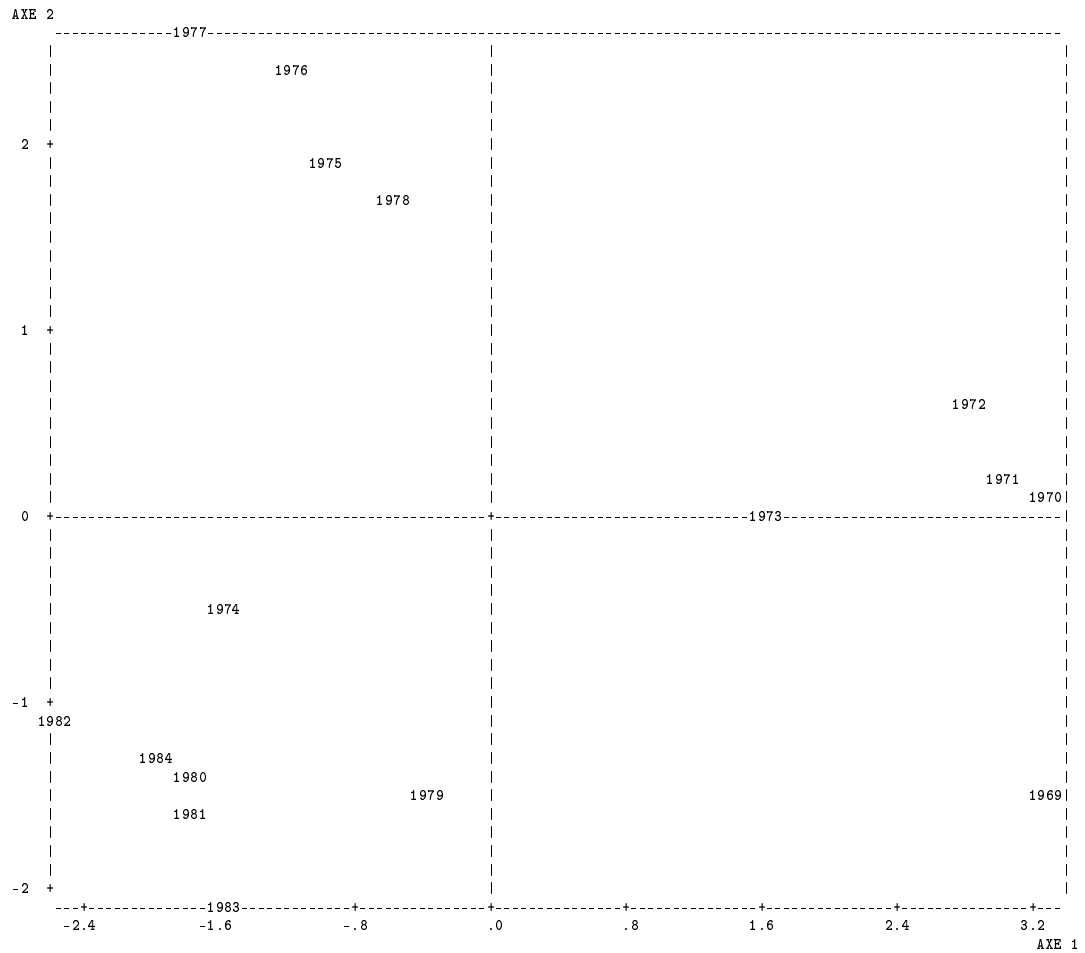
VARIABLES		COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
VARIABLES ACTIVES																
NET - NET		.86	-.32	.07	-.35	.18	.86	-.32	.07	-.35	.18	.40	-.22	.08	-.50	.46
INT - INT		.64	-.62	.35	.13	-.25	.64	-.62	.35	.13	-.25	.30	-.42	.43	.19	-.63
SUB - SUB		-.76	-.24	-.40	-.40	-.19	-.76	-.24	-.40	-.40	-.19	-.36	-.16	-.48	-.58	-.48
LMT - LMT		.14	.97	-.07	.05	-.11	.14	.97	-.07	.05	-.11	.06	.67	-.08	.08	-.29
DCT - DCT		-.95	-.02	-.05	.28	.09	-.95	-.02	-.05	.28	.09	-.45	-.01	-.06	.41	.24
IMM - IMM		.86	.48	-.10	.02	-.07	.86	.48	-.10	.02	-.07	.41	.33	-.12	.03	-.17
EXP - EXP		-.92	-.25	.28	.01	-.02	-.92	-.25	.28	.01	-.02	-.44	-.17	.34	.02	-.06
VRD - VRD		.50	-.59	-.55	.31	.00	.50	-.59	-.55	.31	.00	.23	-.41	-.66	.45	.01

Ce dernier ne fournit pas, pour des raisons mystérieuses, les cosinus carrés et les contributions à l'inertie. La coordonnée d'une variable sur un axe n'est autre que la corrélation entre cette variable et la composante principale. Ici la première composante principale est corrélée de manière positive avec la variable NET (0.86) et de manière négative avec DCT (-0.95). En revanche elle est peu corrélée avec la variable LMT (0.14). La qualité de la représentation d'une variable étant égale au carré de sa coordonnée sur l'axe correspondant il suffit pour obtenir le cosinus carré d'élever la coordonnée au carré. Ainsi la qualité de représentation de NET sur l'axe 1 est égale à  $(0.86)^2$  soit 0.74. En revanche elle n'est que de 0.02 pour LMT.

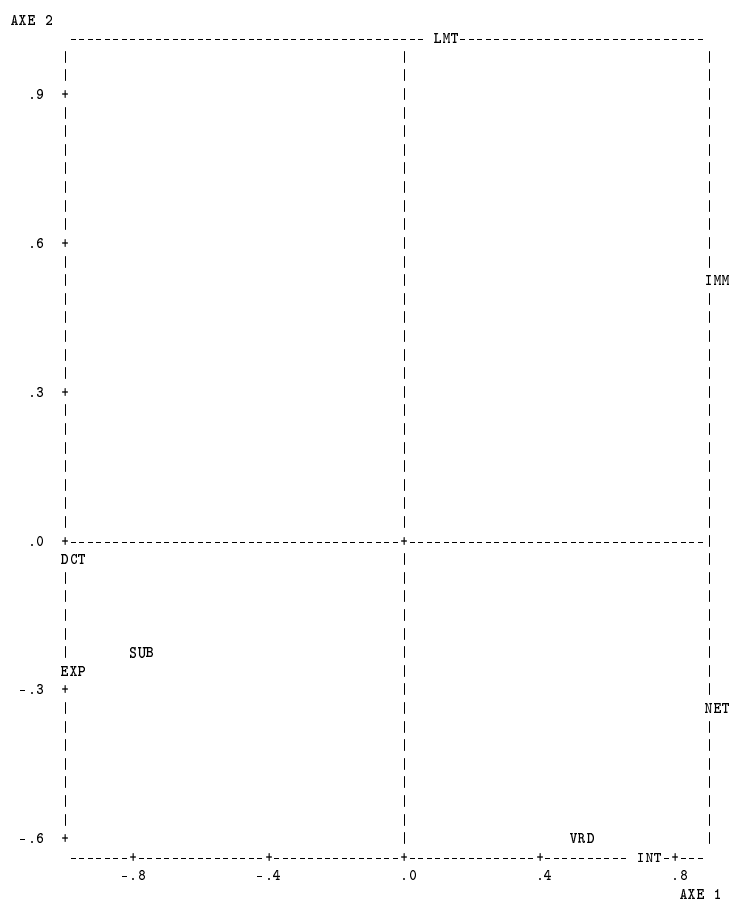
Pour obtenir la contribution d'une variable à l'inertie de l'axe il suffit soit d'élever la coordonnée du vecteur propre correspondant (colonne « anciens axes unitaires »)  $u_{jk}$  au carré ou de prendre le carré de la coordonnée  $d_{jk}$  et de diviser par l'inertie conservée par l'axe  $k$  égale à  $\lambda_k$ .  
 Cherchons à titre d'exemple la contribution de la variable NET sur l'axe 1 :  
 $CTR_1(NET) = (0.40)^2 = 0.16 = 16\%$  ou  $(0.86)^2/4.4904 = 16.4\%$

### 3.8.4 Les représentations graphiques

Le nuage des individus représenté sur le plan 1 \* 2 :



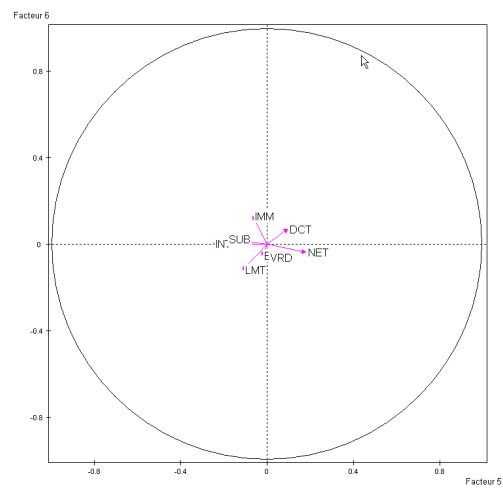
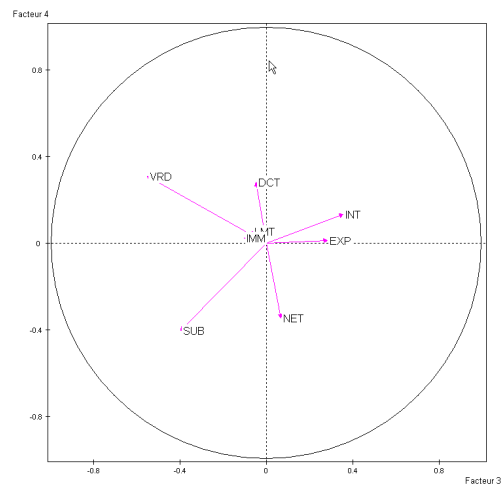
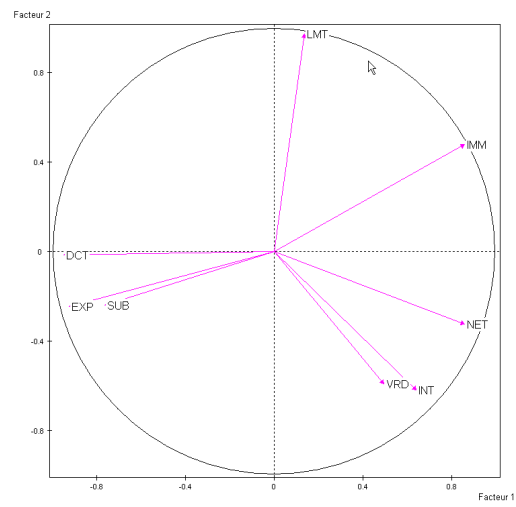
Le nuage des variables :



Rappelons qu'en ACP normée les variables se trouvent à l'intérieur d'un cercle de rayon unité. Vous pouvez vérifier dans le tableau précédent qu'aucune coordonnée n'est supérieure à 1.

Compte tenu des propriétés évoquées auparavant on voit que les représentations graphiques permettent une lecture directe de certains résultats. La variable DCT proche du bord du cercle sur l'axe 1 est bien représentée sur cet axe. En revanche LMT est mal représentée sur cet axe. Par contre la représentation de cette dernière variable est satisfaisante sur le premier plan ( $qlt = 0.96$ ). Le graphique suivant montre la représentation des variables et le cercle des corrélations dans le plan formé par les axes 1 et 2.

La représentation des plans suivants appariant les axes 3 et 4 puis les axes 5 et 6, montre bien l'aptitude décroissante des axes selon leur rang à représenter le nuage.





Nous avons vu que dans l'espace d'origine la longueur de la projection d'une variable sur une autre mesure exactement la corrélation entre ces deux variables. Cette propriété se conserve plus ou moins bien en projection selon la qualité de représentation des variables correspondantes. Ainsi les variables NET et INT sont-elles bien représentées et proches l'une de l'autre. On peut donc affirmer que leur proximité en projection reflète leur proximité dans l'espace et qu'en conséquence elles sont corrélées positivement. L'examen de la matrice des corrélations confirme ce point ( $\text{cor}(\text{NET}, \text{INT}) = 0.68$ ).

MATRICE DES CORRELATIONS

	NET	INT	SUB	LMT	DCT	IMM	EXP	VRD
NET	1.00							
INT	.68	1.00						
SUB	-.50	-.49	1.00					
LMT	-.24	-.50	-.31	1.00				
DCT	-.89	-.60	.62	-.14	1.00			
IMM	.55	.24	-.72	.58	-.81	1.00		
EXP	-.70	-.33	.65	-.38	.86	-.94	1.00	
VRD	.48	.53	-.14	-.45	-.35	.20	-.46	1.00

Les variables NET et DCT étant également bien représentées mais s'opposant, leur corrélation est forte mais négative (-0.89). En revanche DCT et LMT étant presque orthogonales sont peu corrélées (-0.14).

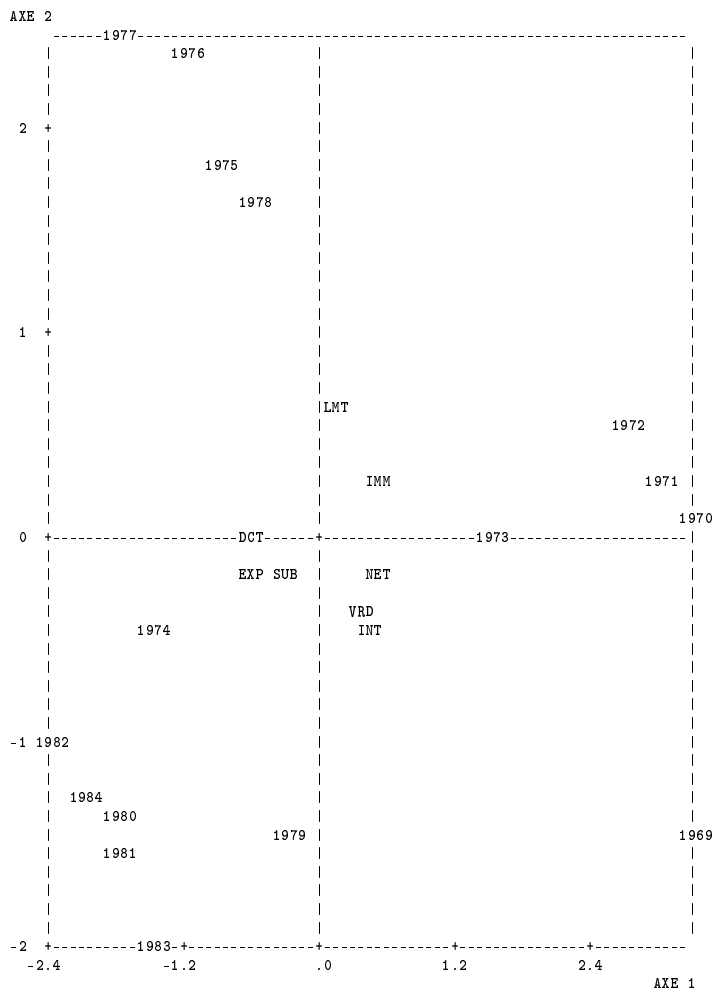
## 3.9 Compléments

### 3.9.1 Variables qualitatives en ACP

Bien que nous ayons dit que l'ACP ne concerne que les variables continues il est néanmoins possible de traiter des données qualitatives. Il suffit pour cela de créer un individu fictif possédant les caractéristiques moyennes des éléments concernés pour chacune des modalités de la variable qualitative. Si l'on désire par exemple faire figurer la variable sexe on crée alors deux individus fictifs, le premier possédant comme caractéristiques les moyennes observées sur les éléments masculins de l'échantillon et le second les caractéristiques moyennes des éléments féminins. Ainsi, et de manière paradoxale, le traitement des variables qualitatives se réalise en ACP dans l'espace des individus. Bien entendu, en raison de cette démarche, les variables qualitatives seront traitées uniquement en éléments supplémentaires. Nous verrons ultérieurement une méthode qui elle permet le traitement en tant qu'éléments actifs, des variables qualitatives.

### 3.9.2 Représentations simultanées en ACP

De même il est possible d'obtenir en ACP une représentation simultanée des 2 nuages. Considérons l'individu dont toutes les composantes sur les variables sont nulles sauf celle d'ordre  $j$  qui vaut 1. Cet élément caractérise la variable  $X_j$  et coïncide avec le vecteur  $e_j$  de la base canonique de  $\mathbf{R}^p$ . Sa projection sur l'axe de rang  $k$  est égale à  ${}^t e_j u_k$  soit  $u_{jk}$ , la coordonnée de rang  $j$  du vecteur propre  $u_k$ . De cette manière, il est possible de représenter sur un même graphique les  $n$  points du nuage des individus et les directions prises par les  $p$  variables. Il ne s'agit pas, à l'inverse de l'AFC par exemple, d'une véritable représentation simultanée des points des deux ensembles. En effet, si les individus sont bien des points, en ACP les variables correspondent plus à des directions (vecteurs). Ce qui compte ce n'est pas la proximité entre un point et une variable mais l'éloignement de l'individu de l'origine dans la direction prise par la variable.



## Chapitre 4

# Analyse Factorielle des Correspondances

### 4.1 Introduction

L'Analyse Factorielle des Correspondances (AFC) traite des données différentes de celles requises par l'ACP. Cette dernière analyse des données quantitatives tandis que la première permet le traitement des tableaux croisés encore appelés tableaux de contingence. Un tel tableau ventile une population ou un effectif selon deux critères qualitatifs.

Un tableau de contingence doit vérifier la propriété suivante : la somme des éléments en ligne possède une signification, de même que la somme des éléments en colonne.

L'objet de l'AFC est d'étudier la nature de la liaison éventuelle entre les deux caractères. Considérons le tableau suivant répartissant les étudiants de l'Université de Rennes I selon la catégorie socio-professionnelle de leurs parents et leur discipline d'inscription :

$n_{ij}$	DRO	ECO	LET	SCI	MED	PHA	DEN	PLU	IUT
agr	270	193	108	455	208	117	22	3	319
oua	8	6	1	9	1	0	8	1	9
ind	346	202	103	391	307	145	68	7	214
lib	923	409	194	1003	1431	429	260	37	265
cad	436	236	155	562	426	137	74	27	247
emp	132	78	33	155	121	20	25	2	57
ouv	546	287	212	604	347	131	36	26	478
ser	57	32	26	88	32	8	11	3	48
aut	150	72	54	103	138	32	22	13	74
san	3	3	2	2	5	1	0	0	0
non	93	65	86	169	295	47	13	30	17

Remarquons en premier lieu qu'il s'agit bien d'un tableau de contingence. La somme des éléments d'une ligne donne le nombre d'inscrits dont les parents possèdent une CSP donnée et la somme des éléments d'une colonne le total inscrit dans une discipline.

Les questions que l'on se pose sont ici les suivantes :

- Existe-t-il une attraction ou au contraire une répulsion entre certaines disciplines et certaines CSP ?
- Peut-on par ailleurs mettre en évidence des relations entre éléments d'un même ensemble, donc dresser une typologie des lignes et des colonnes ?

La première idée que l'on peut avoir est d'utiliser l'ACP. L'emploi de cette méthode se révèle inapproprié en regard de l'objectif fixé. L'analyse révèle un effet taille sur le premier axe avec un taux d'inertie supérieur à 90 %. Toutefois cet axe ne fait qu'échelonner les CSP selon leur nombre d'inscrits, ce qui ne correspond pas au but recherché. Ce qui montre en passant qu'un fort taux d'inertie ne gage pas nécessairement de l'intérêt de l'axe correspondant. Afin d'éliminer cet effet nous allons travailler non pas sur les effectifs bruts mais sur les profils.

## 4.2 Notation

Le nombre de modalités du caractère en ligne est égal à  $n$  et celui du caractère en colonne à  $p$ . Soit  $n_{ij}$  l'effectif dans la modalité  $i$  du premier caractère (en ligne) et de la modalité  $j$  du second caractère (en colonne). Soit  $n_{.}$  l'effectif

total et  $f_{ij}$  la fréquence correspondant aux modalités  $i$  et  $j$ , donc :

$$f_{ij} = \frac{n_{ij}}{n_{..}}$$

Soit par ailleurs  $n_{i.}$  l'effectif total dans la modalité  $i$  du premier caractère et  $n_{.j}$  l'effectif dans la modalité  $j$  du second, soit :

$$n_{i.} = \sum_{j=1}^p n_{ij}$$

$$n_{.j} = \sum_{i=1}^n n_{ij}$$

Notons  $f_{i.}$  et  $f_{.j}$  les fréquences correspondantes :

$$f_{i.} = \sum_{j=1}^p f_{ij} = \frac{n_{i.}}{n_{..}}$$

$$f_{.j} = \sum_{i=1}^n f_{ij} = \frac{n_{.j}}{n_{..}}$$

Dans ces conditions les profils-lignes s'écrivent :

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

et les profils-colonnes :

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

Il s'agit en fait des distributions conditionnelles.

Le tableau suivant fournit les profils-lignes, c'est à dire la répartition des inscrits par discipline à l'intérieur d'une CSP donnée. Les profils sont exprimés en pourcentage (la somme des éléments d'une même ligne est égale à 100).

$f_{ij}/f_{i.}$	DRO	ECO	LET	SCI	MED	PHA	DEN	PLU	IUT
agr	15.9	11.4	6.4	26.8	12.3	6.9	1.3	0.2	18.8
oua	18.6	14.0	2.3	20.9	2.3	0.0	18.6	2.3	20.9
ind	19.4	11.3	5.8	21.9	17.2	8.1	3.8	0.4	12.0
lib	18.6	8.3	3.9	20.3	28.9	8.7	5.3	0.7	5.4
cad	19.0	10.3	6.7	24.4	18.5	6.0	3.2	1.2	10.7
emp	21.2	12.5	5.3	24.9	19.4	3.2	4.0	0.3	9.1
ouv	20.5	10.8	7.9	22.6	13.0	4.9	1.3	1.0	17.9
ser	18.7	10.5	8.5	28.9	10.5	2.6	3.6	1.0	15.7
aut	22.8	10.9	8.2	15.7	21.0	4.9	3.3	2.0	11.2
san	18.8	18.8	12.5	12.5	31.3	6.3	0.0	0.0	0.0
non	11.4	8.0	10.6	20.7	36.2	5.8	1.6	3.7	2.1
Moyenne	18.7	10.0	6.1	22.3	20.9	6.7	3.4	0.9	10.9

Le profil-moyen figure à la dernière ligne et correspond à la répartition de l'ensemble des inscrits par discipline, c'est à dire toutes CSP confondues.

De même, les profils-colonnes (exprimés en pourcentage) figurent dans le tableau suivant :

$f_{ij}/f_{.j}$	DRO	ECO	LET	SCI	MED	PHA	DEN	PLU	IUT
agr	9.1	12.2	11.1	12.8	6.3	11.0	4.1	2.0	18.5
oua	0.3	0.4	0.1	0.3	0.0	0.0	1.5	0.7	0.5
ind	11.7	12.8	10.6	11.0	9.3	13.6	12.6	4.7	12.4
lib	31.1	25.8	19.9	28.3	43.2	40.2	48.2	24.8	15.3
cad	14.7	14.9	15.9	15.9	12.9	12.8	13.7	18.1	14.3
emp	4.5	4.9	3.4	4.4	3.7	1.9	4.6	1.3	3.3
ouv	18.4	18.1	21.8	17.1	10.5	12.3	6.7	17.4	27.7
ser	1.9	2.0	2.7	2.5	1.0	0.7	2.0	2.0	2.8
aut	5.1	4.5	5.5	2.9	4.2	3.0	4.1	8.7	4.3
san	0.1	0.2	0.2	0.1	0.2	0.1	0.0	0.0	0.0
non	3.1	4.1	8.8	4.8	8.9	4.4	2.4	20.1	1.0

### 4.3 Une métrique spécifique en AFC

De la même manière que s'était posé en ACP le choix de la distance, métrique identité ou métrique inverse des variances, se pose en AFC le choix

d'une distance adaptée aux profils. En AFC la distance utilisée est celle dite du « chi-deux » qui s'exprime de la manière suivante :

La distance entre deux profils-ligne  $i_1$  et  $i_2$  s'écrit :

$$d_{\chi^2}^2(i_1, i_2) = \sum_{j=1}^p \frac{1}{f_{.j}} \left( \frac{f_{i_1 j}}{f_{i_1.}} - \frac{f_{i_2 j}}{f_{i_2.}} \right)^2$$

La métrique du chi-deux se distingue de la métrique usuelle par la pondération par l'inverse de  $f_{.j}$  qui vise à rééquilibrer le poids de chaque discipline dans le calcul des distances entre profils des CSP. La métrique usuelle s'écrit en effet :

$$d^2(i_1, i_2) = \sum_{j=1}^p \left( \frac{f_{i_1 j}}{f_{i_1.}} - \frac{f_{i_2 j}}{f_{i_2.}} \right)^2$$

Examinons par exemple la distance entre les deux CSP « ouv » et « agr » :

$f_{ij}/f_{i.}$	DRO	ECO	LET	SCI	MED	PHA	DEN	PLU	IUT
agr	15.9	11.4	6.4	26.8	12.3	6.9	1.3	0.2	18.8
ouv	20.5	10.8	7.9	22.6	13.0	4.9	1.3	1.0	17.9
$f_{.j}$	18.7	10.0	6.1	22.3	20.9	6.7	3.4	0.9	10.9

Examinons plus particulièrement le rôle joué dans le calcul de cette distance par les deux disciplines « SCI » et « PLU ». La distance usuelle s'écrit :

$$\begin{aligned} d^2(agr, ouv) &= \dots + \frac{(26.8 - 22.6)^2}{100^2} \dots + \frac{(0.2 - 1.0)^2}{100^2} + \dots \\ &= \dots + \frac{17.64}{100^2} + \dots + \frac{0.64}{100^2} + \dots \end{aligned}$$

La contribution de la variable « SCI » à la distance entre les deux CSP est d'environ 28 fois celle de la modalité « PLU », or l'écart sur la première est de 1,2 alors qu'il est de 1 à 5 pour la seconde. Ainsi, la distance usuelle aura tendance à favoriser les modalités les plus fréquentes. Ce qui est bien le cas ici, puisque « SCI » vaut en moyenne 22,3 % alors que « PLU » seulement 0,9 %.

Examinons maintenant la distance du « chi-deux » :

$$\begin{aligned} d_{\chi^2}^2(agr, ouv) &= \dots + \left(\frac{100}{22.3}\right) \frac{(26.8 - 22.6)^2}{100^2} \dots + \left(\frac{100}{0.9}\right) \frac{(0.2 - 1.0)^2}{100^2} + \dots \\ &= \dots + \frac{79.10}{100} + \dots + \frac{71.11}{100} + \dots \end{aligned}$$

La contribution des deux modalités est désormais proche. La distance du chi-deux tend donc bien à rééquilibrer le rôle des modalités à faible effectif, bien que de manière imparfaite.

De manière symétrique la distance entre deux profils colonnes  $j_1$  et  $j_2$  va s'écrire :

$$d_{\chi^2}^2(j_1, j_2) = \sum_{i=1}^n \frac{1}{f_{i.}} \left( \frac{f_{ij_1}}{f_{.j_1}} - \frac{f_{ij_2}}{f_{.j_2}} \right)^2$$

Le choix de la distance du chi-deux se fonde sur la propriété dite d'équivalence distributionnelle. Celle-ci implique que si l'on agrège deux profils lignes identiques  $i_1$  et  $i_2$  on ne modifie pas la distance entre les profils-colonnes. Cette propriété s'énonce de manière symétrique pour les profils-colonnes.

Démonstration : montrons que l'agrégation de deux lignes  $i_1$  et  $i_2$  possédant des profils identiques ne modifie pas la distance entre les profils-colonnes.

Notons  $i_0$  le profil-ligne résultant de la réunion des profils  $i_1$  et  $i_2$ .

Dans le calcul de la distance entre profils-colonnes  $j_1$  et  $j_2$  seuls les termes relatifs à  $i_1$  et  $i_2$  sont modifiés. Il faut donc montrer :

$$\frac{1}{f_{i_1.}} \left( \frac{f_{i_1 j_1}}{f_{.j_1}} - \frac{f_{i_1 j_2}}{f_{.j_2}} \right)^2 + \frac{1}{f_{i_2.}} \left( \frac{f_{i_2 j_1}}{f_{.j_1}} - \frac{f_{i_2 j_2}}{f_{.j_2}} \right)^2 = \frac{1}{f_{i_0.}} \left( \frac{f_{i_0 j_1}}{f_{.j_1}} - \frac{f_{i_0 j_2}}{f_{.j_2}} \right)^2$$

Il est immédiat que :

$$f_{i_0.} = f_{i_1.} + f_{i_2.}$$

Le terme en  $i_0$  peut donc encore s'écrire :

$$f_{i_0.} \left( \frac{f_{i_0 j_1}}{f_{i_0.} f_{.j_1}} - \frac{f_{i_0 j_2}}{f_{i_0.} f_{.j_2}} \right)^2 = f_{i_1.} \left( \frac{f_{i_1 j_1}}{f_{i_0.} f_{.j_1}} - \frac{f_{i_1 j_2}}{f_{i_0.} f_{.j_2}} \right)^2 + f_{i_2.} \left( \frac{f_{i_2 j_1}}{f_{i_0.} f_{.j_1}} - \frac{f_{i_2 j_2}}{f_{i_0.} f_{.j_2}} \right)^2$$

Les lignes  $i_1$  et  $i_2$  étant identiques nous avons :

$$\frac{f_{i_1 j}}{f_{i_1.}} = \frac{f_{i_2 j}}{f_{i_2.}} = \frac{f_{i_1 j} + f_{i_2 j}}{f_{i_1.} + f_{i_2.}} = \frac{f_{i_0 j}}{f_{i_0.}}$$



Ce qui établit l'égalité recherchée.

La conséquence pratique de cette propriété est qu'elle garantit l'utilisateur relativement à la construction des classes. Ainsi on ne gagnera rien en subdivisant des classes homogènes. De la même manière on ne perdra rien en regroupant des catégories identiques. Dans l'exemple précédent si les inscrits issus de deux CSP différentes se répartissent de manière identique selon les disciplines, leur fusion en une seule CSP ne va pas modifier la distance entre les disciplines.

## 4.4 Résolution

Nous allons maintenant montrer que l'on peut se replacer dans le cadre général de l'analyse d'un nuage quelconque. Toutefois en ce qui concerne l'analyse dite duale nous aborderons une autre présentation conforme à celle que l'on trouve dans la plupart des ouvrages.

### 4.4.1 Analyse dans $\mathbf{R}^p$ . Le nuage des profils-lignes

Soit  $F$  le tableau  $[n, p]$  des fréquences  $f_{ij}$ . Soient par ailleurs  $D_n$  la matrice diagonale d'ordre  $n$  de terme générique  $f_{i.}$  et  $D_p$  la matrice diagonale d'ordre  $p$  d'élément  $f_{.j}$ . L'AFC peut être présentée comme l'analyse du tableau  $X = D_n^{-1}F$  avec la métrique  $M = D_p^{-1}$  et la matrice des masses  $P = D_n$ . Chaque profil-ligne est en effet affecté du poids  $f_{i.}$ .

Analyse des profils-colonnes dans  $\mathbf{R}^p$  :

$$\text{Données : } X_{[n,p]} = D_n^{-1}F = \begin{pmatrix} \vdots & & \\ \cdots & \frac{f_{ij}}{f_{i.}} & \cdots \\ & f_{i.} & \\ \vdots & & \end{pmatrix}$$

$$\text{Métrique : } M_{[p,p]} = D_p^{-1} = \begin{pmatrix} \ddots & & 0 \\ & \frac{1}{f_{.j}} & \\ 0 & & \ddots \end{pmatrix}$$

$$\text{Poids : } P_{[n,n]} = D_n = \begin{pmatrix} \ddots & & 0 \\ & f_{i.} & \\ 0 & & \ddots \end{pmatrix}$$

Nous savons que l'axe factoriel de rang  $k$  est obtenu comme vecteur propre de la matrice  ${}^tXPM$  soit ici  ${}^tFD_n^{-1}FD_p^{-1}$ .

Le vecteur propre  $u_k$  vérifie donc la relation :

$${}^tFD_n^{-1}FD_p^{-1}u_k = \lambda_k u_k \quad (4.1)$$

Les coordonnées des profils-lignes sur l'axe  $k$  sont obtenus par la relation  $C_k = XM u_k$  soit ici :

$$C_k = D_n^{-1}FD_p^{-1}u_k$$

Remarque : La matrice à diagonaliser  ${}^tFD_n^{-1}FD_p^{-1}$  n'est pas symétrique. Il est toutefois possible de se ramener au cas d'une matrice symétrique pour lesquelles il existe des algorithmes de diagonalisation spécifiques. Ce point essentiellement informatique n'est pas abordé ici.

#### 4.4.2 Analyse dans $\mathbf{R}^n$ . Le nuage des profils-colonnes

Il existe une symétrie parfaite entre les deux analyses mais il convient de noter que le tableau des données n'est pas ici le transposé de celui de l'analyse directe.

Le tableau des données a pour terme  $f_{ij}/f_{.j}$  soit sous forme matricielle  $D_p^{-1}{}^tF$ . La métrique est  $D_n^{-1}$  de terme  $(1/f_{.i})$ . La matrice des poids est  $D_p$  de terme  $(f_{.j})$ .

Analyse des profils-colonnes dans  $\mathbf{R}^n$  :

$$\begin{aligned} \text{Données : } X_{[p,n]} &= D_p^{-1}{}^tF = \begin{pmatrix} & \vdots & \\ \cdots & \frac{f_{ij}}{f_{.j}} & \cdots \\ & \vdots & \end{pmatrix} \\ \text{Métrique : } M_{[n,n]} &= D_n^{-1} = \begin{pmatrix} \ddots & & 0 \\ & \frac{1}{f_{.i}} & \\ 0 & & \ddots \end{pmatrix} \\ \text{Poids : } P_{[p,p]} &= D_p = \begin{pmatrix} \ddots & & 0 \\ & f_{.j} & \\ 0 & & \ddots \end{pmatrix} \end{aligned}$$

L'axe factoriel de rang  $k$  est obtenu comme vecteur propre de rang  $k$  de la matrice  $FD_p^{-1t}FD_n^{-1}$ , soit :

$$FD_p^{-1t}FD_n^{-1}v_k = \mu_k v_k \quad (4.2)$$

Multiplions cette dernière égalité à gauche par  ${}^tFD_n^{-1}$  :

$${}^tFD_n^{-1}FD_p^{-1}({}^tFD_n^{-1}v_k) = \mu_k({}^tFD_n^{-1}v_k)$$

${}^tFD_n^{-1}v_k$  apparaît donc comme vecteur propre de la matrice  ${}^tFD_n^{-1}FD_p^{-1}$  associé à la valeur propre  $\mu_k$ . Comme  $\lambda_k$  est la plus grande valeur propre au rang  $k$  de cette matrice, on en déduit que  $\mu_k \leq \lambda_k$ . En procédant de même on peut montrer que  $\lambda_k \leq \mu_k$  d'où l'égalité  $\mu_k = \lambda_k$ . Les valeurs propres sont donc identiques dans les deux analyses pour des axes de rang homologue. Par ailleurs, le vecteur  ${}^tFD_n^{-1}v_k$  n'étant pas normé et sa  $D_p^{-1}$  norme valant  $\lambda_k$ , on en déduit :

$$u_k = \frac{1}{\sqrt{\lambda_k}} {}^tFD_n^{-1}v_k \quad (4.3)$$

de même :

$$v_k = \frac{1}{\sqrt{\lambda_k}} FD_p^{-1}u_k \quad (4.4)$$

Ainsi, bien que n'étant pas placé dans les conditions du schéma général présenté au premier chapitre, on en retrouve pas moins les propriétés usuelles, à savoir égalité des valeurs propres pour des axes de même rang et formules de transition entre vecteurs. Les relations de transition s'étendent de la même manière aux coordonnées, comme nous le verrons plus loin.

### 4.4.3 Interprétation des résultats

Examinons en premier lieu les résultats de la diagonalisation et le choix du nombre d'axes à retenir.

EDITION DES VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION ... .0878  
SOMME DES VALEURS PROPRES .... .0878

HISTOGRAMME DES 8 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT. POURCENT.	POURCENT. CUMULE	
1	.0631	71.80	71.80	*****
2	.0144	16.41	88.21	*****
3	.0048	5.44	93.65	*****
4	.0027	3.06	96.71	****
5	.0016	1.85	98.56	***
6	.0007	.85	99.42	*
7	.0004	.49	99.91	*
8	.0001	.09	100.00	*

Le premier axe explique 71.8 % de l'inertie totale et on obtient près de 90 % de l'inertie totale en ne retenant que les 2 premiers axes. Décidons en conséquence de n'interpréter que ces 2 axes.

#### Examen de l'axe 1 - Les profils-lignes

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS SUR LES AXES 1 A 5

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
agr	10.69	.13	.33	-.03	-.12	-.03	-.02	18.5	.6	34.8	3.1	3.3	.86	.01	.12	.01	.00
oua	.27	1.07	.27	-.42	.65	-.34	-.51	.3	3.3	24.0	12.0	43.1	.07	.17	.40	.11	.24
ind	11.24	.02	.07	-.09	-.01	.02	.00	.8	6.0	.1	1.4	.1	.28	.46	.00	.02	.00
lib	31.22	.09	-.28	-.08	-.01	.01	-.01	40.0	14.3	1.3	1.7	.6	.92	.08	.00	.00	.00
cad	14.51	.01	.05	.03	.02	-.05	.02	.5	.8	1.4	11.3	2.7	.31	.11	.07	.30	.04
emp	3.93	.04	.03	-.05	.10	-.09	.13	.0	.7	8.0	12.5	40.7	.02	.06	.24	.21	.41
ouv	16.82	.10	.30	.06	.02	.05	.00	24.8	4.0	1.4	15.9	.0	.93	.03	.00	.03	.00
ser	1.92	.13	.30	.04	.09	-.15	.00	2.7	.2	3.2	16.7	.0	.71	.01	.06	.18	.00
aut	4.15	.05	.01	.11	.16	.12	-.01	.0	3.4	23.6	21.2	.1	.00	.22	.51	.26	.00
san	.10	.39	-.25	.24	.01	.15	.33	.1	.4	.0	.8	6.6	.16	.14	.00	.06	.27
non	5.14	.34	-.39	.43	-.04	-.04	-.03	12.1	66.3	2.0	3.5	2.8	.44	.55	.01	.01	.00

I	Contribution (%)	Coordonnée	Signe de la coordonnée
lib	40.0	-0.28	-
ouv	24.8	0.30	+
agr	18.5	0.33	+
	$\Sigma=83.3$		

L'élément « non » n'a pas été retenu, ce qui comporte une part d'arbitraire. Notons que sa contribution est de 12.1 % soit proche de la contribution moyenne (100 / 11) et que d'autre part avec les trois éléments précédents nous obtenons 83.3 % de l'inertie totale de l'axe 1.

Le premier axe oppose donc les professions libérales aux catégories ouvriers et

agriculteurs. Cette opposition est dominante puisque le premier axe conserve plus de 70 % de l'inertie totale.

Les profils-colonnes

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES FREQUENCES SUR LES AXES 1 A 5																		
FREQUENCES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES					
IDEN - LIBELLE COURT P.REL DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5			
FREQUENCES ACTIVES																		
DRO - DRO	18.69	.01	.03	-.06	.08	.05	.04	.3	4.1	22.2	18.4	17.6	.07	.22	.39	.18	.11	
ECO - ECO	9.98	.02	.12	-.01	.03	-.01	.03	2.3	.1	1.8	.7	6.3	.69	.01	.04	.01	.05	
LET - LET	6.14	.09	.14	.27	.03	.01	.00	2.0	30.1	1.3	.1	.0	.22	.75	.01	.00	.00	
SCI - SCI	22.33	.01	.08	.00	-.04	-.08	.02	2.2	.0	8.7	51.0	3.4	.42	.00	.13	.42	.02	
MED - MED	20.88	.13	-.35	.05	-.03	.02	.00	41.0	3.9	3.8	2.2	.0	.96	.02	.01	.00	.00	
PHA - PHA	6.73	.07	-.16	-.13	-.14	.07	-.05	2.7	7.5	26.5	11.3	8.9	.36	.23	.27	.06	.03	
DEN - DEN	3.40	.27	-.34	-.31	.19	-.09	-.11	6.3	22.2	26.1	10.4	23.7	.44	.35	.14	.03	.04	
PLU - PLU	.94	.64	-.27	.70	.21	-.02	-.18	1.1	31.8	8.4	.2	19.4	.11	.76	.07	.00	.05	
IUT - IUT	10.90	.25	.49	-.02	-.02	.04	-.06	42.1	.3	1.1	5.6	20.5	.97	.00	.00	.01	.01	

J	Contribution (%)	Coordonnée	Signe de la coordonnée
MED	41.0	-0.35	-
IUT	42.1	0.49	+
	$\Sigma = 83.1$		

L'opposition entre CSP dégagée recouvre une opposition entre les professions libérales sur représentées en Médecine et sous représentées en IUT d'une part et les ouvriers et agriculteurs d'autre part qui présentent le profil inverse.

Ce résultat garanti par les formules de transition peut être mis en évidence de manière empirique en opérant un retour aux données de base. Si la discipline dans laquelle s'inscrit un étudiant était indépendante de la CSP possédée par ses parents alors l'effectif observé dans une case serait égal au produit des marges, soit :

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

Appliquons cette démarche aux éléments  $i$  et  $j$  dégagées par l'AFC :

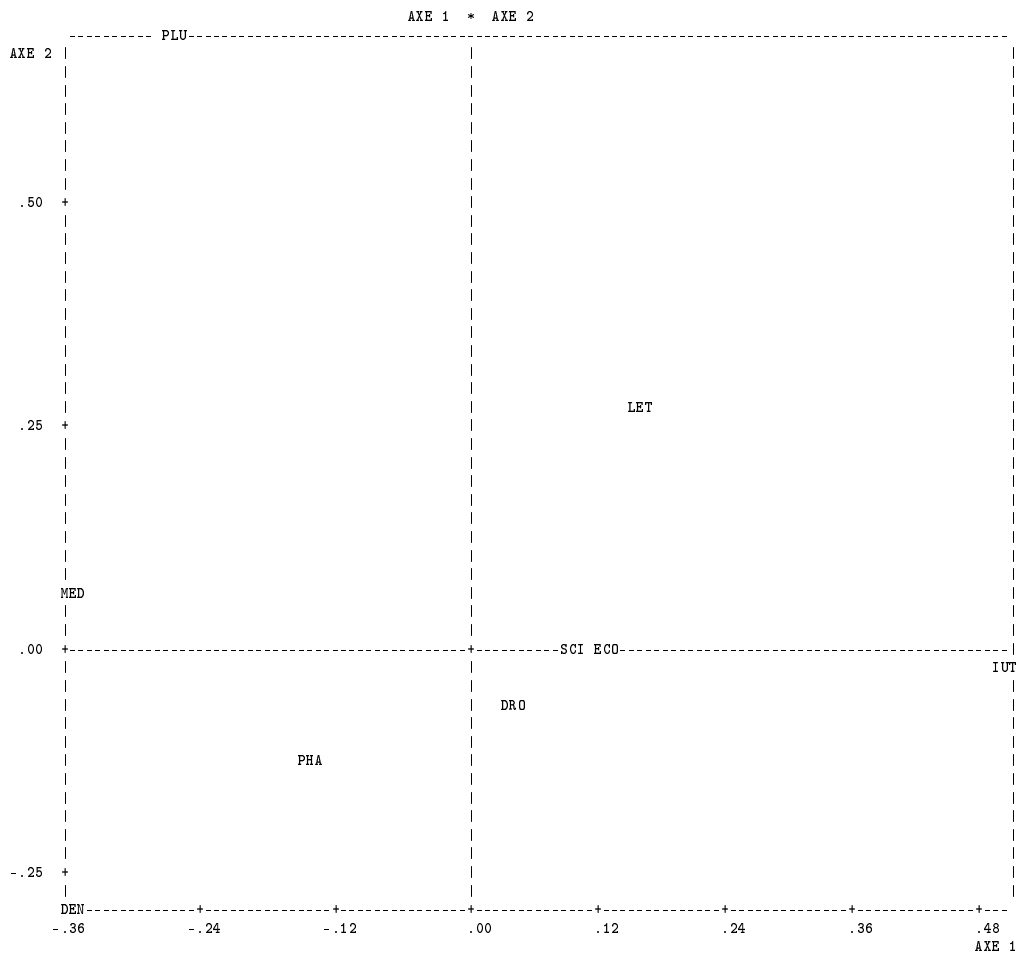
$n_{ij}$ observé	MED	IUT
$n_{ij} = n_{i.}n_{.j}/n_{..}$		
lib	1431	265
	1034	540
agr	208	319
	354	185
ouv	347	278
	557	291

L'examen du tableau vient confirmer les résultats précédents. La catégorie professions libérales apparaît bien sur-représentée en Médecine, on observe 1 431 inscrits alors que si la discipline dans laquelle un étudiant s'inscrit était indépendante de la CSP de ses parents, on devrait en observer 1 034 soit environ 1.5 fois moins. Inversement on attendrait 540 inscrits issus de cette même CSP en IUT alors qu'ils ne sont que 265, soit deux fois moins. Ce constat s'inverse pour les catégories agr et ouv.

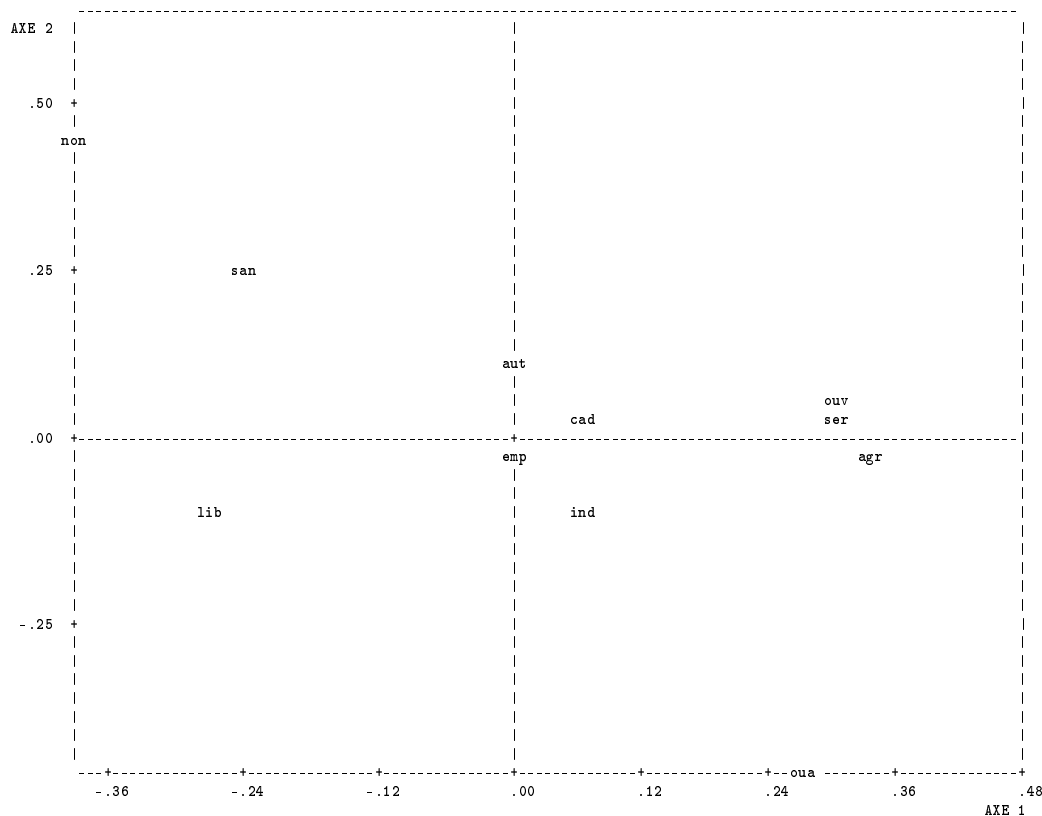
L'axe 2 isole quant à lui les non réponses et les distingue des professions libérales alors que ces deux catégories étaient associées sur l'axe 1. On remarque au passage sur cet axe l'attraction des professions libérales pour les études dentaires ce qui confirme l'intérêt manifesté par cette catégorie pour les disciplines de santé.

Les représentations graphiques :

Le nuage des profils-colonnes



## Le nuage des profils-lignes



## Remarque :

Sur les dangers de l'interprétation directe des graphiques.

Sur l'axe 1 les catégories ouvrier et personnel de service sont proches et excentrées (voir les coordonnées), personnel de service étant plus éloigné de l'origine que ouvrier. L'interprétation directe du graphique conduit à affirmer que ouv et ser se distinguent du profil moyen, ce qui est vrai (le profil moyen se projetant à l'origine), et que ces deux catégories contribuent fortement à l'élaboration de l'axe 1, ce qui est faux. L'examen des contributions révèle que ouv possède une contribution de 24.8 % et ser de 2.8 % soit dix fois moins (ou presque).

L'explication tient à ce que dans la notion d'inertie interviennent deux éléments, la masse et la distance (au carré). Ici la différence vient de la masse qui en AFC pour les profils-lignes est  $f_{i.}$ , soit le poids de la CSP dans l'ensemble total des inscrits. Or la catégorie ser est peu « fournie » avec 305 étudiants contre 2 667 pour ouv. Le poids de la première n'est ainsi que de 1.92 contre 16.82 pour la seconde. Le même constat peut être fait avec non et lib.

## 4.5 Propriété fondamentale

### 4.5.1 Les relations de transition en AFC

Les coordonnées des profils-lignes sur l'axe factoriel de rang  $k$  s'écrivent :

$$C_k = D_n^{-1} F D_p^{-1} u_k$$

De la même manière les coordonnées des profils-colonnes sur l'axe de rang  $k$  s'écrivent :

$$D_k = D_p^{-1t} F D_n^{-1} v_k$$

Compte tenu de la relation

$$u_k = \frac{1}{\sqrt{\lambda_k}} {}^t F D_n^{-1} v_k$$

$C_k$  peut encore s'écrire :

$$C_k = \frac{1}{\sqrt{\lambda_k}} D_n^{-1} F D_p^{-1t} F D_n^{-1} v_k$$

soit :

$$C_k = \frac{1}{\sqrt{\lambda_k}} D_n^{-1} F D_k \quad (4.5)$$

De même on peut montrer que :

$$D_k = \frac{1}{\sqrt{\lambda_k}} D_p^{-1t} F C_k \quad (4.6)$$

Examinons les relations induites par ces deux égalités sur chacune des composantes :

$$c_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^p \frac{f_{ij}}{f_{i.}} d_{jk} \quad (4.7)$$

$$d_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^n \frac{f_{ij}}{f_{.j}} c_{ik} \quad (4.8)$$

$c_{ik}$  désignant la coordonnée du profil-ligne  $i$  sur l'axe de rang  $k$  et  $d_{jk}$  la coordonnée du profil-colonne  $j$  sur ce même axe. La relation (4.7) exprime, au facteur  $\frac{1}{\sqrt{\lambda_k}}$  près, que chaque profil-ligne est au barycentre des projections des profils-colonnes affectés du poids de la colonne  $j$  dans la ligne  $i$  (le terme  $\frac{f_{ij}}{f_{i.}}$ ). De même la relation (4.8) exprime que chaque profil-colonne se trouve,



au facteur  $\frac{1}{\sqrt{\lambda_k}}$  près, au barycentre des projections des profils-ligne.

Nous avons ainsi en AFC une double représentation barycentrique : sur les axes factoriels chaque point d'un nuage est au barycentre des points de l'autre nuage.

Dans le cas de l'exemple traité cela signifie par exemple qu'un point CSP est au barycentre des points discipline, chaque discipline étant pondérée par son pourcentage d'inscrits dans la CSP correspondante.

Illustrons cette propriété à l'aide d'un exemple tiré du tableau analysé.

La coordonnée de la CSP ouv sur le premier axe est égale à 0.30.

Considérons le profil-ligne  $(\frac{f_{ij}}{f_i})$  de cette CSP, soit :

J	DRO	ECO	LET	SCI	MED	PHA	DEN	PLU	IUT
ouv	0.21	0.11	0.08	0.23	0.13	0.05	0.01	0.01	0.18

Considérons maintenant les coordonnées  $d_{j1}$  des profils-colonnes sur le premier axe factoriel :

J	DRO	ECO	LET	SCI	MED	PHA	DEN	PLU	IUT
$d_{j1}$	0.03	0.12	0.14	0.08	-0.35	-0.16	-0.34	-0.27	0.49

Le produit terme à terme, soit :

$$\sum_{i=1}^n \frac{f_{ij}}{f_i} d_{j1} = 0.08$$

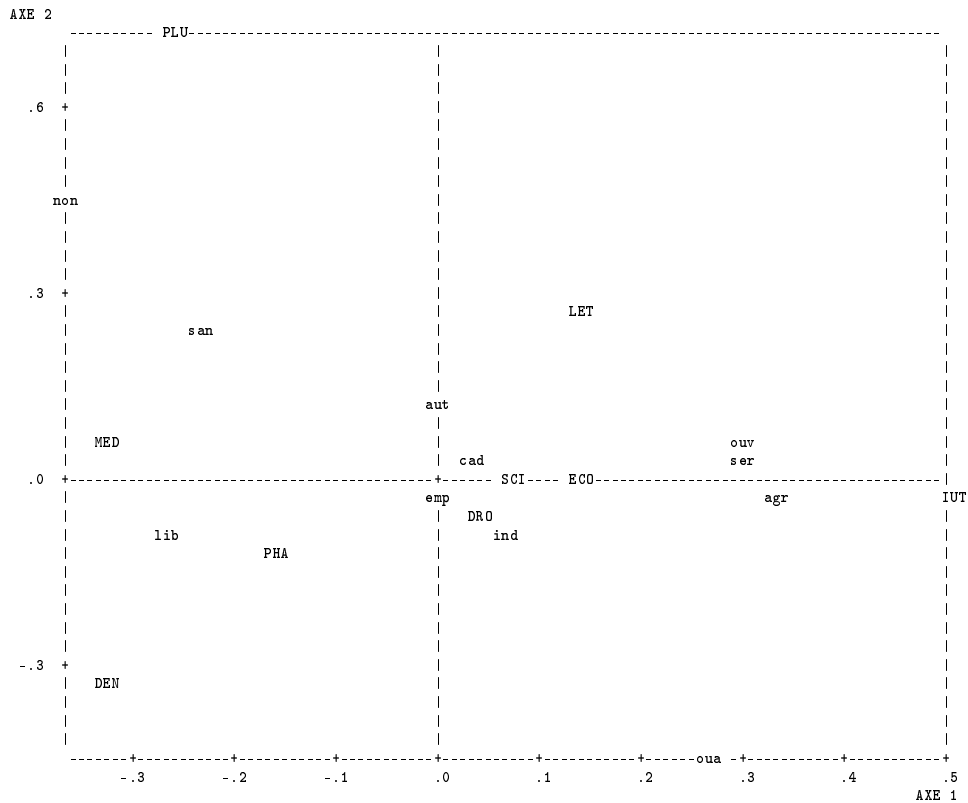
Alors

$$\frac{1}{\sqrt{\lambda_1}} \sum_{i=1}^n \frac{f_{ij}}{f_i} d_{j1} = \frac{1}{\sqrt{0.0631}} * 0.08 = 0.30$$

ce qui coïncide avec  $c_{i1}$ .

De manière symétrique, d'après la relation (4.8), chaque discipline se trouve au barycentre des coordonnées des points CSP, chaque CSP étant pondérée par son poids relatif dans la discipline correspondante. Cette propriété légitime en AFC les représentations simultanées. Alors qu'en ACP une telle représentation constitue un simple artifice, au contraire en AFC les représentations simultanées des points des deux espaces sont licites. Il convient toutefois d'être très prudent dans l'interprétation et d'éviter de commenter d'éventuelles proximités entre points de deux espaces différents. Rappelons que la phase d'interprétation passe toujours par la consultation des aides à l'interprétation.

### 4.5.2 La représentation simultanée des lignes et des colonnes en AFC

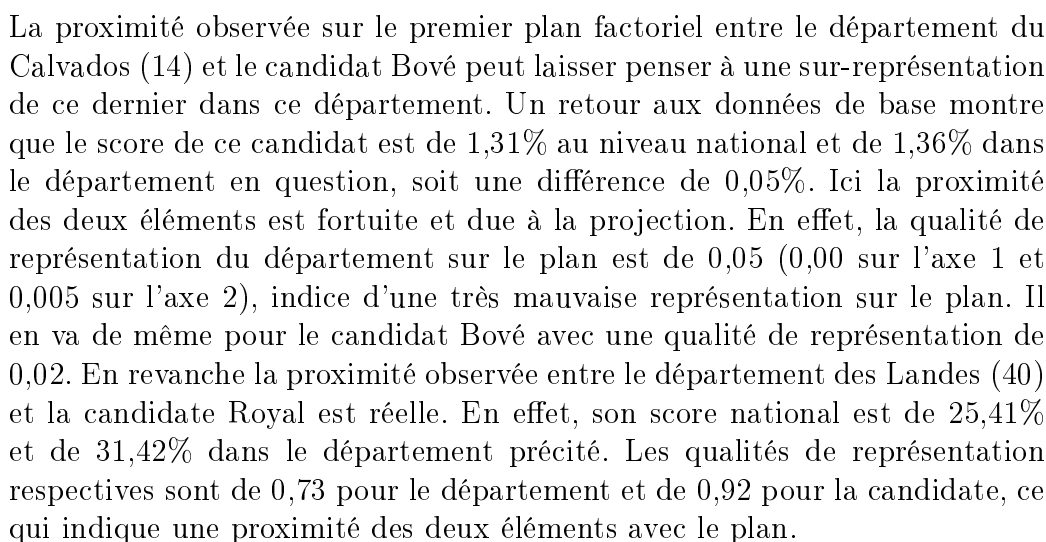


Sur le premier plan ci-dessus les points relatifs aux deux espaces ont été représentés sur le même graphique.<sup>1</sup>

L'exemple suivant constitue une illustration du danger des conclusions hâtives tirées de la proximité entre points des deux espaces. Le tableau analysé fournit la répartition des suffrages exprimés au premier tour de l'élection présidentielle de 2007. En ligne figurent les départements en en colonne les candidats.

1. Sur ce point voir :

« Techniques de la Description Statistique » TABARD, MORINEAU et LEBART - p 60  
« Analyse des Données » VOLLE - p 143



## 4.6 Formes classiques des nuages

L'une des formes les plus fréquentes du nuage en AFC sur le premier plan factoriel est la forme parabolique.

Il s'agit d'un effet classique appelé effet « Gutman ». Dans ce cas le premier axe oppose les extrêmes entre eux et l'axe 2 oppose les extrêmes aux moyens. Une telle structure peut indiquer l'existence d'une liaison de nature quadratique entre variables, qui aurait échappé à l'ACP par exemple, cette dernière ne détectant que d'éventuelles liaisons linéaires. Par ailleurs il est toujours intéressant d'observer les accidents par rapport à l'allure théorique.

Sur les plans suivants le nuage possède la forme suivante : plan (1\*3)  
plan (2\*3)

## 4.7 Compléments

Par rapport à l'analyse générale le nuage n'a pas été centré. Nous allons montrer que cette opération n'est pas nécessaire dans la mesure où la première valeur propre est en AFC égale à 1 et se trouve associée à un axe trivial joignant le barycentre du nuage à l'origine.

Montrons dans un premier temps que les valeurs propres sont en AFC inférieures à 1. D'après la relation (4.8) nous pouvons écrire :

$$\sqrt{\lambda_k} c_{ik} = \sum_{j=1}^n \frac{f_{ij}}{f_i} d_{jk}$$

Les coefficients  $\frac{f_{ij}}{f_i}$  étant  $\leq 1$  nous pouvons écrire :

$$\sqrt{\lambda_k} c_{ik} \leq \text{Max}_{\{j\}} d_{jk} \quad \text{car} \quad \sum_{j=1}^n \frac{f_{ij}}{f_i} = 1$$

Cette inégalité est vraie pour tout  $i$ , donc :

$$\sqrt{\lambda_k} \text{Max}_{\{i\}} c_{ik} \leq \text{Max}_{\{j\}} d_{jk}$$

Soit encore :

$$\lambda_k \text{Max}_{\{i\}} c_{ik} \leq \sqrt{\lambda_k} \text{Max}_{\{j\}} d_{jk}$$

De manière symétrique la relation (4.8) conduit à :

$$\sqrt{\lambda_k} \text{Max}_{\{j\}} d_{jk} \leq \text{Max}_{\{i\}} c_{ik}$$

Des deux relations précédentes on en déduit :

$$\lambda_k \text{Max}_{\{i\}} c_{ik} \leq \text{Max}_{\{i\}} c_{ik} \quad \text{d'où} \quad \lambda_k \leq 1$$

Remarque :

C'est cette propriété qui rend possible la double représentation barycentrique.

En effet, les valeurs propres étant inférieures à 1, le terme en  $\frac{1}{\sqrt{\lambda_k}}$  est supérieur à 1, ce qui va entraîner une dilatation des coordonnées.

Montrons maintenant que dans l'analyse du nuage non centré le premier axe factoriel est associé à la valeur propre  $\lambda = 1$ .

Soit  $G$  le barycentre du nuage dans  $\mathbf{R}^p$ .  $G$  a pour composante  $f_{.j}$  (il s'agit de la distribution marginale des disciplines toutes CSP confondues), soit sous forme matricielle  $D_p \mathbf{1}_p$  où  $\mathbf{1}_p$  désigne la première bissectrice dans  $\mathbf{R}^p$  ( $\mathbf{1}_p$  est le vecteur dont toutes les composantes sont égales à 1). Montrons d'abord que le barycentre du nuage dans  $\mathbf{R}^p$  est bien le vecteur de composante  $f_{.j}$ .

Rappelons que les profils-lignes ont pour terme  $\frac{f_{ij}}{f_{i.}}$  et pour masse  $f_{i.}$  donc :

$$(OG)_j = \sum_{i=1}^n \frac{f_{ij}}{f_{i.}} f_{i.} = \sum_{i=1}^n f_{ij} = f_{.j}$$

Montrons maintenant que ce vecteur est vecteur propre de la matrice d'inertie associé à la valeur propre  $\lambda = 1$ . Il faut donc vérifier que :

$${}^t F D_n^{-1} F D_p^{-1} (D_p \mathbf{1}_p) = D_p \mathbf{1}_p$$

soit encore :

$${}^t F D_n^{-1} F \mathbf{1}_p = D_p \mathbf{1}_p$$

La matrice  ${}^t F D_n^{-1}$  de dimension  $[p, n]$  a pour terme d'ordre  $ji$  :

$${}^t F D_n^{-1} = \begin{pmatrix} \vdots & & \\ \dots & f_{ij} & \dots \\ \vdots & & \end{pmatrix} \begin{pmatrix} \ddots & & 0 \\ & 1/f_{i.} & \\ 0 & & \ddots \end{pmatrix} = \begin{pmatrix} \vdots & & \\ \dots & f_{ij}/f_{i.} & \dots \\ \vdots & & \end{pmatrix}$$

Par ailleurs,  $F \mathbf{1}_p$  est un vecteur  $[p, 1]$  de composante :

$$F \mathbf{1}_p = \begin{pmatrix} \vdots & & \\ \dots & f_{ij} & \dots \\ \vdots & & \end{pmatrix} \begin{pmatrix} \vdots \\ 1 \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \sum_{j=1}^p f_{ij} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ f_{i.} \\ \vdots \end{pmatrix}$$

d'où

$${}^t F D_n^{-1} F \mathbf{1}_p = \begin{pmatrix} \vdots \\ \dots & f_{ij}/f_{i.} & \dots \\ \vdots \end{pmatrix} \begin{pmatrix} \vdots \\ f_{i.} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \sum_{i=1}^n f_{ij} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ f_{.j} \\ \vdots \end{pmatrix}$$

Le vecteur  $D_p \mathbf{1}_p$  est donc vecteur propre de la matrice d'inertie  ${}^t F D_n^{-1} F D_p^{-1}$  associé à la valeur propre  $\lambda = 1$ . Comme nous avons montré que toutes les valeurs propres étaient en AFC inférieures ou égales à 1, on en déduit donc que  $\lambda_1 = 1$ . Ainsi, en AFC le premier axe factoriel correspond à l'axe joignant l'origine au barycentre du nuage. Nous allons voir toutefois que cet axe ne présente pas d'intérêt et qu'il est donc éliminé. Il est parfois qualifié d'axe « trivial ». En effet, sur cet axe tous les éléments du nuage des profils-lignes se projettent en  $G$ . Montrons tout d'abord que le vecteur  $OG = D_p \mathbf{1}_p$  est unitaire pour la norme  $D_p^{-1}$  :

$$\begin{aligned} \|OG\|_{D_p^{-1}}^2 &= {}^t(D_p \mathbf{1}_p) D_p^{-1} \mathbf{1}_p = {}^t \mathbf{1}_p D_p \mathbf{1}_p = \begin{pmatrix} 1 & \dots & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} f_{.1} \\ \vdots \\ f_{.j} \\ \vdots \\ f_{.p} \end{pmatrix} \\ &= \sum_{j=1}^p f_{.j} = 1 \end{aligned}$$

Dans ces conditions, la longueur de la projection de tout profil-ligne  $i$  est donnée par le produit scalaire, avec la métrique  $M = D_p^{-1}$ , du profil par le vecteur  $D_p \mathbf{1}_p$ , soit :

$$\langle x_i, OG \rangle_{D_p^{-1}} = {}^t x_i D_p^{-1} D_p \mathbf{1}_p = {}^t x_i \mathbf{1}_p = \sum_{j=1}^p \frac{f_{ij}}{f_{i.}} = 1$$

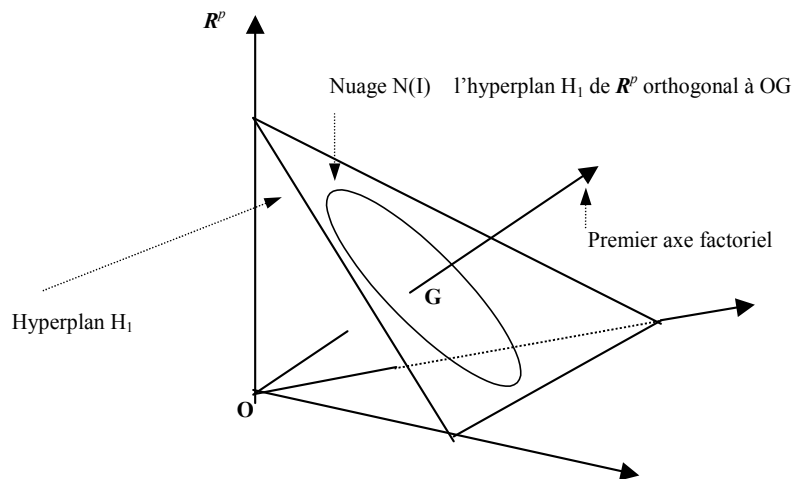
Ainsi, l'ensemble des éléments du nuage  $N(I)$  des profils-lignes se projette en  $G$  sur l'axe  $OG$ . La raison en est que l'ensemble des éléments de  $N(I)$  se trouve dans un hyperplan de  $\mathbf{R}^p$   $D_p^{-1}$ -orthogonal à l'axe  $OG$ . En effet, chaque profil-ligne est tel que ses coordonnées vérifient :

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

Le nuage des profils-lignes est donc contenu dans un hyperplan de  $\mathbf{R}^p$ . De plus cet hyperplan est orthogonal au vecteur  $OG$ . En ne centrant pas le nuage,

donc en réalisant l'analyse par rapport à l'origine, le premier axe conduit à projeter le nuage sur  $OG$ .

On peut donc soit ne pas centrer le nuage et éliminer le premier axe, soit le centrer et récupérer  $p - 1$  axes. Dans ce cas la dernière valeur propre est égale à 0. Dans tous les cas il faut retenir qu'en AFC si le tableau traité est de dimension  $[n, p]$  on récupère  $\text{Min}(n, p) - 1$  axes non triviaux.



Remarques :

1- L'essentiel d'un programme d'Analyse Factorielle d'un tableau  $[n, p]$  se résumant en une diagonalisation d'une matrice d'ordre  $p$ , on s'arrange en AFC pour placer en colonne le caractère présentant le plus petit nombre de modalités.

2- Certains logiciels, par exemple ADDAD dans une version antérieure, édi-  
taient en AFC la première valeur propre afin de fournir un indicateur de la  
qualité de la diagonalisation, sachant que celle-ci en théorie est exactement  
égale à 1.

Montrons maintenant que si  $V$  désigne la matrice d'inertie sur données brutes et  $W$  sur données centrées alors : le vecteur  $OG = (f_{.j})$  noté  $g$  est vecteur propre de  $V$  associé à la valeur propre 1 et vecteur propre de  $W$  associé à la valeur propre 0.

Sur données brutes (dans  $\mathbf{R}^p$ ) :

$$\begin{aligned} X_{[n,p]} &= \begin{pmatrix} f_{ij} \\ f_{i.} \end{pmatrix} = D_n^{-1} F \\ M_{[p,p]} &= \begin{pmatrix} 1 \\ f_{.j} \end{pmatrix} = D_p^{-1} \\ P_{[n,n]} &= (f_{i.}) = D_n \end{aligned}$$

La démonstration a été faite au point précédent. donc  $Vg = g$ . Ainsi  $g$  est vecteur propre de  $V$  associé à  $\lambda_1 = 1$ .

Après centrage, les données s'écrivent :

$$X = \begin{pmatrix} f_{ij} \\ f_{i.} \end{pmatrix} - f_{.j}, \text{ soit sous forme matricielle } X = D_n^{-1} F - \mathbf{1}_n {}^t g$$

Cherchons l'expression de  $W$  en fonction  $V$  :

$$\begin{aligned} W &= {}^t X P X M = {}^t (D_n^{-1} F - \mathbf{1}_n {}^t g) D_n (D_n^{-1} F - \mathbf{1}_n {}^t g) D_p^{-1} \\ &= ({}^t F D_n^{-1} - g {}^t \mathbf{1}_n) D_n (D_n^{-1} F - \mathbf{1}_n {}^t g) D_p^{-1} \\ &= ({}^t F D_n^{-1} F - {}^t F \mathbf{1}_n {}^t g - g {}^t \mathbf{1}_n F + g {}^t \mathbf{1}_n D_n \mathbf{1}_n {}^t g) D_p^{-1} \\ &= ({}^t F D_n^{-1} F - {}^t F \mathbf{1}_n {}^t g - g {}^t \mathbf{1}_n F + g {}^t \mathbf{1}_n D_n \mathbf{1}_n {}^t g) D_p^{-1} \\ &= {}^t F D_n^{-1} F D_p^{-1} - {}^t F \mathbf{1}_n {}^t g D_p^{-1} - g {}^t \mathbf{1}_n F D_p^{-1} + g {}^t \mathbf{1}_n D_n \mathbf{1}_n {}^t g D_p^{-1} \end{aligned}$$

Or

$${}^t \mathbf{1}_n D_n \mathbf{1}_n = \sum_{i=1}^n f_{i.} = 1 \Rightarrow -g {}^t \mathbf{1}_n D_n \mathbf{1}_n {}^t g D_p^{-1} = -g {}^t g D_p^{-1}$$

Par ailleurs :

$${}^t \mathbf{1}_n F = {}^t ({}^t F \mathbf{1}_n) = {}^t g \Rightarrow -g {}^t \mathbf{1}_n F D_p^{-1} = -g {}^t g D_p^{-1}$$

et

$$-{}^t F \mathbf{1}_n {}^t g D_p^{-1} = -g {}^t g D_p^{-1}$$



donc :

$$W = {}^t F D_n^{-1} F D_p^{-1} - g^t g D_p^{-1} = V - g^t g D_p^{-1}$$

Dans ces conditions  $Wg = Vg - g^t g D_p^{-1} g = g - g = 0$  car  ${}^t g D_p^{-1} g = 1$  et  $g$  apparaît bien comme vecteur propre de la matrice  $W$  associé à la valeur propre  $\lambda = 0$ .

Montrons maintenant que si  $u$  est vecteur propre de  $V$ ,  $u \neq g$ , associé à la valeur propre  $\lambda$ , alors  $u$  est aussi vecteur propre de  $W$  associé à  $\lambda$ . Soit  $u$  vérifiant  $Vu = \lambda u$ .

Cherchons  $Wu$  :

$Wu = (V - g^t g D_p^{-1})u = Vu - g^t g D_p^{-1} u$  or  $Vu = \lambda u$  et les vecteurs propres étant orthogonaux deux à deux pour la métrique  $M = D_p^{-1}$ , alors  ${}^t g D_p^{-1} u = 0$  d'où  $Wu = Vu = \lambda u$ .

## 4.8 Lien avec le test du $\chi^2$

L'énoncé introductif du problème posé en AFC, à savoir l'analyse d'une liaison éventuelle entre les deux variables, évoque immédiatement le test du  $\chi^2$  relatif à l'indépendance de deux caractères. L'objet de cette section est de préciser le lien entre l'AFC et ce test.

rappelons qu'en AFC la matrice à diagonaliser s'écrit :

$$V = {}^t F D_n^{-1} F D_p^{-1}$$

Le terme  $(V)_{jk}$  de cette matrice est égal à :

$$(V)_{jk} = \sum_{i=1}^n \frac{f_{ij}}{f_{i.}} \frac{f_{ik}}{f_{.k}}$$

Le terme diagonal d'ordre  $j$  s'écrit :

$$(V)_{jj} = \sum_{i=1}^n \frac{f_{ij}}{f_{i.}} \frac{f_{ij}}{f_{.j}}$$

La trace de  $V$  est donc égale à :

$$\sum_{j=1}^p \sum_{i=1}^n \frac{f_{ij}^2}{f_{i.} f_{.j}}$$

Cette quantité n'est autre que l'inertie totale du nuage  $I_0$  mesurée par rapport à l'origine. Nous avons vu précédemment que la valeur propre maximale

est égale à 1. L'inertie du nuage est donc, après élimination du facteur trivial, égale à  $I_0 - 1$ . Cette quantité est en fait l'inertie du nuage  $N(I)$  mais mesurée par rapport au barycentre du nuage, soit  $I_G$ . cette dernière quantité mesure en quelque sorte l'inertie « utile », celle dont la décomposition par l'AFC présente un intérêt.

La quantité

$$\sum_{j=1}^p \sum_{i=1}^n \frac{f_{ij}^2}{f_{i.}f_{.j}} - 1$$

peut encore s'écrire

$$\sum_{j=1}^p \sum_{i=1}^n \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \quad (1)$$

En effet :

$$\sum_{j=1}^p \sum_{i=1}^n \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \sum_{j=1}^p \sum_{i=1}^n \left( \frac{f_{ij}^2}{f_{i.}f_{.j}} + f_{i.}f_{.j} - 2f_{ij} \right)$$

or

$$\sum_{j=1}^p \sum_{i=1}^n f_{ij} = 1$$

et

$$\sum_{j=1}^p \sum_{i=1}^n f_{i.}f_{.j} = \sum_{j=1}^p f_{.j} \sum_{i=1}^n f_{i.} = 1$$

d'où

$$\sum_{j=1}^p \sum_{i=1}^n \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \sum_{j=1}^p \sum_{i=1}^n \frac{f_{ij}^2}{f_{i.}f_{.j}} - 1$$

Or la quantité (1) n'est autre, au facteur  $\frac{1}{n_{..}}$  près, que la valeur de la statistique du  $\chi^2$  entre les deux caractères.

Ainsi en AFC, l'inertie totale du nuage après élimination de l'axe trivial, n'est autre que la valeur de la statistique du  $\chi^2$  entre les deux caractères :

$$I_0 - 1 = I_G = \frac{S_{\chi^2}}{n_{..}}$$

Vérifions cette propriété sur l'exemple traité. La statistique  $S_{\chi^2}$  calculée à partir du tableau est égale à 1 392. Par ailleurs l'effectif total  $n_{..}$  est égal à 15856 et l'inertie totale  $I_0$  est égale à 1.0878. On vérifie bien que

$$I_0 - 1 = 0.0878 = 1392/15856.$$

L'intérêt de l'AFC tient à ce qu'elle permet de préciser la nature de la relation entre les deux caractères lorsqu'on rejette l'hypothèse d'indépendance. Elle indique quelles sont les « cases » du tableau responsables de l'écart à la situation d'indépendance. Elle permet également à travers la séquence des valeurs propres de quantifier l'importance des cases correspondantes. Enfin elle fournit la possibilité de visualiser les relations à travers les plans factoriels. Il convient toutefois de souligner que le test du  $\chi^2$  pourrait être enrichi de manière à fournir des informations sur l'écart à la situation d'indépendance.

$I$	$J$				$i.$
	1	...	$j$	...	$p$
1					
$\vdots$					$\vdots$
$i$	... $n_{ij}$ ...				$n_{i.}$
$\vdots$					$\vdots$
$n$					
$.j$	... $n_{.j}$ ...				$n_{..}$

Dans le calcul de la statistique du  $\chi^2$  la case  $(i, j)$  possède la contribution

$$\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}}\right)^2}{\frac{n_{i.}n_{.j}}{n_{..}}}. \text{ où } n_{ij} \text{ est l'effectif observé et } \frac{n_{i.}n_{.j}}{n_{..}} \text{ l'effectif théorique attendu sous l'hypothèse d'indépendance.}$$

Le test du  $\chi^2$  repose sur la sommation de ces quantités sur l'ensemble des cases mais est utilisé en règle générale de manière binaire, rejet ou acceptation de l'hypothèse d'indépendance. En cas de rejet de cette hypothèse il est possible de préciser la nature de la dépendance en examinant les contributions de chaque case  $(i, j)$  au  $\chi^2$  total,

$$\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}}\right)^2}{\frac{n_{i.}n_{.j}}{n_{..}}}$$

soient les quantités :  $\frac{n_{..}}{\chi^2}$

Les cases pour lesquelles ces quantités sont les plus élevées sont responsables de l'écart à la situation d'indépendance. Elles coïncident avec les modalités des deux caractères possédant les plus fortes contributions absolues.

	DRO	ECO	LET	SCI	<i>MED</i>	PHA	DEN	PLU	<i>IUT</i>
<i>agr</i>	7	3	0	15	60	0	22	10	<u>98</u>
<i>oua</i>	0	1	1	0	7	3	29	1	4
<i>ind</i>	0	3	0	0	11	5	1	6	2
<i>lib</i>	0	15	40	10	<u>153</u>	28	50	2	<u>140</u>
<i>cad</i>	0	0	1	5	6	2	0	1	0
<i>emp</i>	2	4	1	2	1	11	1	3	2
<i>ouv</i>	5	2	14	0	79	13	33	0	<u>121</u>
<i>ser</i>	0	0	3	6	16	8	0	0	7
<i>aut</i>	6	1	5	13	0	3	0	8	0
<i>san</i>	0	1	1	1	1	0	1	0	2
<i>non</i>	23	3	26	1	92	1	8	65	58

La démarche précédente peut d'ailleurs être complétée en calculant des contributions « signées ». Il suffit d'affecter les quantités figurant dans le tableau du signe de  $n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}}$ . Si le signe est positif cela indique que l'effectif observé est supérieur à celui attendu dans l'hypothèse d'indépendance et inversement lorsqu'il est négatif.

	DRO	ECO	LET	SCI	<i>MED</i>	PHA	DEN	PLU	<i>IUT</i>
<i>agr</i>	-7	3	0	15	-60	0	-22	-10	<u>98</u>
<i>oua</i>	0	1	-1	0	-7	-3	29	1	4
<i>ind</i>	0	3	0	0	-11	5	1	-6	2
<i>lib</i>	0	-15	-40	-10	<u>153</u>	28	50	-2	<u>-140</u>
<i>cad</i>	0	0	1	5	-6	-2	0	1	0
<i>emp</i>	2	4	-1	2	-1	-11	1	-3	-2
<i>ouv</i>	5	2	14	0	-79	-13	-33	0	<u>121</u>
<i>ser</i>	0	0	3	6	-16	-8	0	0	7
<i>aut</i>	6	1	5	-13	0	-3	0	8	0
<i>san</i>	0	1	1	-1	1	0	-1	0	-2
<i>non</i>	-23	-3	26	-1	92	-1	-8	65	-58

Remarque :

La quantité

$$\sum_{j=1}^p \sum_{i=1}^n \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

égale à l'inertie totale du nuage après élimination du facteur trivial s'appelle le *lien* entre les caractères  $I$  et  $J$ . Cette quantité peut être interprétée comme une mesure de la quantité d'information, au sens de la théorie de l'information, apportée par la connaissance du contenu du tableau (les termes  $n_{ij}$ ) par rapport à la seule connaissance de ses marges ( $n_{i.}$  et  $n_{.j}$ ). L'expression du lien indique qu'en cas d'indépendance entre les deux caractères le gain d'information est nul. En effet dans ce cas le contenu d'une case,  $n_{ij}$ , est obtenu par simple produit des marges correspondantes. A l'inverse plus les effectifs observés s'écartent de la situation d'indépendance et plus le gain d'information s'accroît, ce qu'exprime bien la quantité précédente<sup>2</sup>. Cette expression de l'inertie du nuage montre bien que l'AFC constitue fondamentalement une méthode d'analyse de la structure des écarts à la situation d'indépendance.

## 4.9 AFC et analyse d'un nuage de points quelconque

La présentation que nous avons adoptée correspond à la démarche usuelle de la plupart des ouvrages. Elle diffère, ainsi que nous l'avons déjà souligné, de l'analyse générale du premier chapitre dans la mesure où les tableaux ne sont pas les transposés l'un de l'autre dans les deux nuages. Par ailleurs métrique et masses ne sont pas non plus permutées dans les deux analyses.

	Analyse dans $\mathbf{R}^p$	Analyse dans $\mathbf{R}^n$
Données	$X_{[n,p]} = \begin{pmatrix} f_{ij} \\ f_{i.} \end{pmatrix} = D_n^{-1} F$	$Y_{[p,n]} = \begin{pmatrix} f_{ij} \\ f_{.j} \end{pmatrix} = D_p^{-1t} F$
Métrique	$M_{[p,p]} = \begin{pmatrix} 1 \\ f_{.j} \end{pmatrix} = D_p^{-1}$	$M_{[n,n]} = \begin{pmatrix} 1 \\ f_{i.} \end{pmatrix} = D_n^{-1}$
Poids	$P_{[n,n]} = (f_{i.}) = D_n$	$P_{[p,p]} = (f_{.j}) = D_p$

Il est immédiat que  $Y$  n'est pas le transposé de  $X$ , de même métrique et matrice des masses ne sont pas permutées. Pour revenir à la formulation générale et retrouver cette propriété il suffit de poser dans  $\mathbf{R}^p$  :

---

2. Sur ce point voir VOLLE ch 3 - p 48 et suivantes

$$X = \left( \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \right) = D_n^{-1} F D_p^{-1} - \mathbf{1}_n {}^t \mathbf{1}_p$$

et d'utiliser :

$$M = (f_{.j}) = D_p$$

et

$$P = (f_{i.}) = D_n$$

Montrons maintenant que la matrice d'inertie dans ce cas possède les mêmes valeurs propres et que les vecteurs propres peuvent se déduire de ceux obtenus dans l'analyse précédente.

Appelons  $W_1$  la matrice d'inertie égale à  ${}^t X P X M$  dans le cas général. Alors :

$$\begin{aligned} W_1 &= {}^t (D_n^{-1} F D_p^{-1} - \mathbf{1}_n {}^t \mathbf{1}_p) D_n (D_n^{-1} F D_p^{-1} - \mathbf{1}_n {}^t \mathbf{1}_p) D_p \\ &= (D_p^{-1} {}^t F D_n^{-1} - \mathbf{1}_p {}^t \mathbf{1}_n) D_n (D_n^{-1} F D_p^{-1} - \mathbf{1}_n {}^t \mathbf{1}_p) D_p \\ &= (D_p^{-1} {}^t F - \mathbf{1}_p {}^t \mathbf{1}_n D_n) (D_n^{-1} F - \mathbf{1}_n {}^t \mathbf{1}_p D_p) \\ &= D_p^{-1} {}^t F D_n^{-1} F - D_p^{-1} {}^t F \mathbf{1}_n {}^t \mathbf{1}_p D_p - \mathbf{1}_p {}^t \mathbf{1}_n F + \mathbf{1}_p {}^t \mathbf{1}_n D_n \mathbf{1}_n {}^t \mathbf{1}_p D_p \end{aligned}$$

Or :

$${}^t \mathbf{1}_n D_n \mathbf{1}_n = 1$$

$${}^t F \mathbf{1}_n = g \text{ et } {}^t \mathbf{1}_p D_p = {}^t g \text{ donc } -D_p^{-1} {}^t F \mathbf{1}_n {}^t \mathbf{1}_p D_p = -D_p^{-1} g {}^t g$$

$$-\mathbf{1}_p {}^t \mathbf{1}_n F = -\mathbf{1}_p {}^t g$$

$$\text{et } \mathbf{1}_p {}^t \mathbf{1}_n D_n \mathbf{1}_n {}^t \mathbf{1}_p D_p = \mathbf{1}_p {}^t g$$

$$\text{d'où } W_1 = D_p^{-1} {}^t F D_n^{-1} F - D_p^{-1} g {}^t g = D_p^{-1} ({}^t F D_n^{-1} F - g {}^t g)$$

Dans l'analyse précédente les vecteurs propres vérifiaient la relation  $Wu = \lambda u$  avec  $W = {}^t F D_n^{-1} F D_p^{-1} - g {}^t g D_p^{-1}$  soit :

$$({}^t F D_n^{-1} F - g {}^t g) D_p^{-1} u = \lambda u$$

Posons  $v = D_p^{-1} u$  alors :

$$D_p^{-1} ({}^t F D_n^{-1} F - g {}^t g) v = \lambda v$$

soit :

$$W_1 v = \lambda v$$

On en déduit que les valeurs propres de  $W$  et  $W_1$  sont les mêmes et que les vecteurs propres vérifient la relation  $v = D_p^{-1}u$ . Le vecteur  $u$  étant normé pour la métrique  $D_p^{-1}$  il est immédiat que  $v$  est unitaire pour  $M = D_p$ .

## 4.10 Sélection des axes

Bien que nous ayons expliqué dans une partie préliminaire qu'il convenait de sélectionner les axes plus sur la base de leur signification éventuelle que sur des tests statistiques, il convient de souligner l'existence d'abaques utilisables en AFC<sup>3</sup>. Le graphique suivant fournit selon les dimensions du tableau la valeur que peut atteindre la première valeur propre dans l'hypothèse d'indépendance des lignes et des colonnes.

Ici pour un tableau [11, 9] le pourcentage d'inertie sur le premier axe factoriel peut atteindre 40

Les méthodes de bootstrap peuvent aussi être utilisées en AFC<sup>4</sup>. Signalons toutefois que leur utilisation requiert la connaissance des données de base, c'est à dire des tableaux d'indicatrices des deux modalités dont le croisement fournit le tableau de contingence analysé.

Il existe en AFC un test de sélection des axes fondé sur la formule de reconstitution des données vue au chapitre premier.

Rappelons que les coordonnées des profils-colonnes sur l'axe de rang  $k$  sont définies par :

$$D_k = D_p^{-1t} F D_n^{-1} v_k$$

de même pour les profils-lignes :

$$C_k = D_n^{-1} F D_p^{-1} u_k$$

Comme par ailleurs, d'après les relations de transition entre vecteurs :

$$v_k = \frac{1}{\sqrt{\lambda_k}} F D_p^{-1} u_k$$

on en déduit que :

$$D_k = \sqrt{\lambda_k} D_p^{-1} u_k$$

---

3. Voir « Techniques de la Description Statistique » TABARD, MORINEAU et LEBART - p 223 et suivantes

4. Voir « Statistique Exploratoire Multidimensionnelle » LEBART, MORINEAU et PIRON - p 389 et suivantes

d'où

$$C_k^t D_k = \sqrt{\lambda_k} D_n^{-1} F D_p^{-1} u_k^t u_k D_p^{-1}$$

En sommant sur  $k$ , on obtient :

$$\sum_{k=1}^p \frac{1}{\sqrt{\lambda_k}} C_k^t D_k = D_n^{-1} F D_p^{-1} \sum_{k=1}^p u_k^t u_k D_p^{-1}$$

Les vecteurs  $u_k$  formant une base orthonormée pour la métrique  $D_p^{-1}$  alors :

$$\sum_{k=1}^p u_k^t u_k D_p^{-1} = I$$

soit encore :

$$\sum_{k=1}^n \frac{1}{\sqrt{\lambda_k}} C_k^t D_k = D_n^{-1} F D_p^{-1}$$

Au niveau de chaque coordonnée  $(i, j)$  cette dernière relation s'écrit :

$$\sum_{k=1}^p \frac{1}{\sqrt{\lambda_k}} c_{ik} d_{jk} = \frac{f_{ij}}{f_{i.} f_{.j}}$$

d'où

$$n_{ij} = \frac{n_{i.} n_{.j}}{n_{..}} \sum_{k=1}^p \frac{1}{\sqrt{\lambda_k}} c_{ik} d_{jk}$$

Comme nous avons que  $\lambda_1 = 1$  et que  $c_{i1} = d_{j1} = 1$ , la relation précédente devient :

$$n_{ij} = \frac{n_{i.} n_{.j}}{n_{..}} \left( 1 + \sum_{k=2}^p \frac{1}{\sqrt{\lambda_k}} c_{ik} d_{jk} \right)$$

où  $c_{ik}$  désigne la coordonnée du profil-ligne  $i$  sur l'axe de rang  $k$  et  $d_{jk}$  celle du profil-colonne  $j$  sur l'axe de même rang. D'où l'idée d'un test afin de sélectionner les axes qui consiste à comparer à l'aide du  $\chi^2$  la distribution observée  $N$  à sa reconstitution  $N^{(q)}$  obtenue par la formule précédente<sup>5</sup> avec les  $q$  premiers axes ( $q_{leqp}$ ). La quantité

$$Q = \sum_{i=1}^n \sum_{j=1}^p \frac{\left( n_{ij} - n_{ij}^{(q)} \right)^2}{n_{ij}^{(q)}}$$

---

5. « Data analysis in socio-economic statistics with special consideration of Correspondance Analysis » E. MALINVAUD - Jouy en Josas 1987



où

$$n_{ij}^{(q)} = \frac{n_{i.}n_{.j}}{n_{..}} \left( 1 + \sum_{k=2}^{q \leq p} \frac{1}{\sqrt{\lambda_k}} c_{ik} d_{jk} \right)$$

est comparée à un  $\chi^2$  à  $(n - q - 1)(p - q - 1)$  degrés de liberté où  $q$  désigne le nombre d'axes utilisés pour construire  $N^{(q)}$ . Tant que le test ne permet pas de conclure à l'identité des deux distributions, on peut estimer que le nombre d'axes retenus  $q$  est insuffisant.

Appliquons ce test à l'exemple traité. En nous limitant au premier axe, le test revient à comparer la distribution observée à celle obtenue sous l'hypothèse d'indépendance des deux caractères, soit le test classique du  $\chi^2$ . La valeur de la statistique du test, 1392, conduit à rejeter l'hypothèse d'identité des deux distributions. Considérons la distribution obtenue avec les cinq premiers axes :

$n_{ij}^{(q)}$	MED	DEN	PLU	PHA	DRO	SCI	ECO	LET	IUT
agr	281	189	110	462	201	117	20	2	307
oua	13	6	1	12	0	-1	6	0	5
ind	354	184	92	397	327	132	65	6	227
lib	930	428	201	996	1411	433	259	37	258
cad	415	239	155	555	439	128	78	26	263
emp	123	67	36	152	112	30	32	6	66
ouv	537	303	216	600	344	138	37	27	467
ser	56	37	24	87	32	8	10	4	48
aut	153	68	54	107	139	32	21	13	75
san	3	1	1	2	5	1	0	0	1
non	100	63	86	175	293	44	10	29	13

La somme des carrés des écarts entre la distribution observée et la distribution précédente vaut 47, qui comparée avec la valeur d'un  $\chi^2$  à  $(9 - 5)(11 - 5) = 24$  ddgrés de liberté, soit 42, conduit à rejeter l'hypothèse d'identité des deux distributions. Le tableau suivant récapitule la démarche :

Nombre d'axes $q$	Inertie (%)	$S^{(q)}$	DDL	$\chi^2$ tabulé	Décision
1	71.80	1392	80	102	Rejet
2	16.41	410	63	92	Rejet
3	5.44	155	48	74	Rejet
4	3.06	95	35	57	Rejet
5	1.85	23	24	31	Acceptation

Ainsi dans cet exemple, il serait nécessaire de sélectionner les cinq premiers axes, le dernier conservant moins de 2 % de l'inertie totale ...

### 4.11 Utilisation des résultats

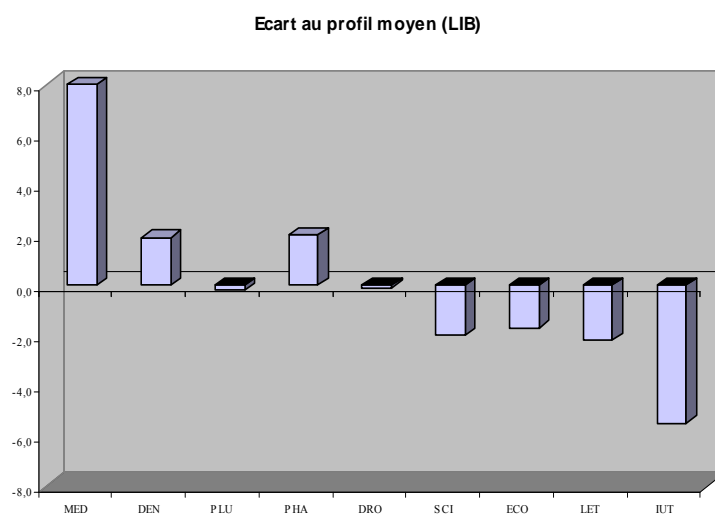
On peut utiliser l'AFC comme aide à la réalisation de certains graphiques. En reclassant les lignes et les colonnes du tableau selon leur coordonnée sur le premier axe on obtient la structure qui répond à la question initiale de mise en évidence des typologies à l'intérieur des deux sous-ensembles et des attractions entre lignes et colonnes.

Reclassement des profils-lignes selon la coordonnée sur le premier facteur (lignes et colonnes)

	MED	DEN	PLU	PHA	DRO	SCI	ECO	LET	IUT	1#F
non	36.2	1.6	3.7	5.8	11.4	20.7	8.0	10.6	2.1	-0.39
lib	28.9	5.3	0.7	8.7	18.6	20.3	8.3	3.9	5.4	-0.28
san	31.3	0.0	0.0	6.3	18.8	12.5	18.8	12.5	0.0	-0.25
moy	20.9	3.4	0.9	6.7	18.7	22.3	10.0	6.1	10.9	0.00
aut	21.0	3.3	2.0	4.9	22.8	15.7	10.9	8.2	11.2	0.01
emp	19.4	4.0	0.3	3.2	21.2	24.9	12.5	5.3	9.1	0.03
cad	18.5	3.2	1.2	6.0	19.0	24.4	10.3	6.7	10.7	0.05
ind	17.2	3.8	0.4	8.1	19.4	21.9	11.3	5.8	12.0	0.07
oua	2.3	18.6	2.3	0.0	18.6	20.9	14.0	2.3	20.9	0.27
ouv	13.0	1.3	1.0	4.9	20.5	22.6	10.8	7.9	17.9	0.30
ser	10.5	3.6	1.0	2.6	18.7	28.9	10.5	8.5	15.7	0.30
agr	12.3	1.3	0.2	6.9	15.9	26.8	11.4	6.4	18.8	0.33
1#F	-0.35	-0.34	-0.27	-0.16	0.03	0.08	0.12	0.14	0.49	

En interclassant le profil-moyen (barycentre) qui se projette à l'origine on éclaire la structure du tableau.

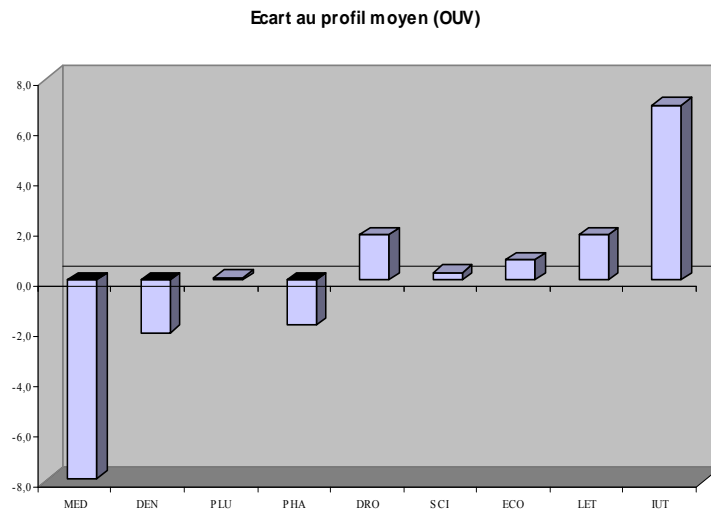
Il est à noter que certains praticiens se sont penchés sur cette question<sup>6</sup>. Les méthodes empiriques qu'ils proposent aboutissent à un résultat dont l'AFC donne la solution optimale. Dans cette optique apparaît bien le rôle de l'AFC qui est d'augmenter la valeur pratique de l'information et surtout de fournir des guides de lecture des grands tableaux<sup>7</sup>.



---

6. Voir par exemple « La Graphique dans la presse » Serge et Madeleine BONIN - Editions du CFPJ

7. Voir aussi « Analyse des Données » M. VOLLE - p 65 et suivantes



## 4.12 Extensions

Jusqu'ici nous avons uniquement considéré l'AFC dans l'optique du traitement des tableaux de contingence. Il est toutefois possible d'utiliser l'AFC sur d'autres types de tableaux.

### 4.12.1 L'analyse factorielle des correspondances multiples

En premier lieu l'AFC peut s'appliquer aux tableaux mis sous forme disjonctive complète. Les variables continues sont découpées en modalités et l'individu est codé 1 dans la modalité qu'il possède et 0 dans les autres. Cette technique constitue l'Analyse Factorielle des Correspondances Multiples (ACM) et un chapitre spécifique lui sera consacré.

### 4.12.2 L'AFC sur tableaux de notes

Supposons que l'on dispose d'un tableau dont le terme  $(i, j)$  indique la note obtenue par l'étudiant  $i$  dans la discipline  $j$ . Bien qu'a priori ce tableau puisse être traité par l'ACP on peut chercher à utiliser l'AFC. Le problème dans ce cas est que les masses afférentes aux lignes étant les totaux correspondants (les  $f_{i.}$ ) l'utilisation de l'AFC conduit à pondérer les individus en accordant à ceux ayant obtenu les meilleures notes les poids les plus élevés. Afin de remédier à ce problème on crée une matière complément qui pour chaque matière contiendra le complément à la note maximale. De cette manière les marges en ligne du tableau seront constantes.

Exemple :

I	MATH	STAT	...	MATH+	MATH-	STAT+	STAT-
i	16	6	...	16	4	6	14

### 4.12.3 L'AFC appliquée à certains tableaux chronologiques

Considérons un tableau de contingence croisant la population d'un département par âge et par canton à l'issue d'un recensement. De tels tableaux existent pour chaque recensement. Il peut alors être intéressant d'analyser la suite de tableaux ainsi obtenus. L'AFC peut être utilisée en cumulant l'ensemble des tableaux. L'individu traité devient alors le canton \* année. Diverses approches peuvent être envisagées. On peut par exemple empiler les tableaux et mettre le tableau « moyen » en supplémentaire.

Signalons toutefois le développement plus récent de méthodes spécifiques dédiées au traitement des données « ternaires » ou « cubiques » :

L'analyse factorielle multiple (AFM)<sup>8</sup>

La méthode STATIS<sup>9</sup>

L'analyse factorielle conjointe<sup>10</sup>

---

8. Voir in « Analyses Factorielles simples et multiples » ESCOFIER et PAGES - Dunod 1985

9. Analyse Conjointe de tableaux quantitatifs » C. LAVIT - Masson 1988

10. « Analyse Factorielle Conjointe d'une famille de triplets indexés » F. LECHEVAL-LIER - Thèse de doctorat - Rennes 1990

Une présentation de ces diverses méthodes figure dans l'ouvrage « L'Analyse des Données évolutives » Technip 1996



# Chapitre 5

## Analyse des Correspondances Multiples

### 5.1 Introduction

En première instance l'Analyse des Correspondances Multiples (ACM) peut être présentée comme un simple prolongement de l'AFC aux tableaux de données mis sous forme binaire ou tableaux logiques.

Cette première approche coïncide d'ailleurs avec la démarche historique. Toutefois en raison des résultats prometteurs qu'a fournis cette utilisation de l'AFC, l'intérêt s'est porté plus en détail sur cette technique qui désormais constitue une méthode à part entière. De plus elle s'est progressivement imposée comme outil privilégié dans le traitement des données d'enquête. L'ACM permet en effet le traitement des ensembles de données « mixtes », c'est à dire comprenant à la fois des variables quantitatives et des variables qualitatives. Le traitement conjoint de ces deux types de données repose sur leur transformation préalable appelée codage disjonctif complet. Cette phase préliminaire, inexistante par exemple en ACP, est d'une part fondamentale et demeure d'autre part fondamentalement empirique.

Il est à noter que si sur le plan technique, au moins dans un premier temps, l'Analyse des Correspondances Multiples constitue un simple prolongement de l'AFC, sur le plan des objectifs elle s'apparente plus à l'Analyse en Composantes Principales à travers l'analyse des tableaux individus-variables.

Dans un premier temps nous allons suivre la démarche classique et présenter l'ACM comme une extension de l'AFC aux tableaux mis sous forme disjonctive complète. Nous présenterons ensuite quelques propriétés spécifiques qui expliquent les précautions à prendre notamment lors de la phase préliminaire

de codage, à laquelle nous consacrerons une partie. Ensuite nous montrerons que les résultats de l'ACM peuvent être acquis indifféremment par l'analyse de deux types de tableaux. Enfin ce chapitre se clôt par la présentation de propriétés particulières attendant au cas binaire.

## 5.2 L'ACM simple prolongement de l'AFC

### 5.2.1 Le traitement des variables en ACM

En présence d'ensemble de données incluant à la fois des variables qualitatives et des variables quantitatives, l'utilisation de l'ACP n'aurait aucun sens. Considérons à cet égard la variable catégorie socio-professionnelle. Celle-ci est un simple code et la CSP codée 12 n'est pas deux fois plus importante que celle codée 6. La notion de moyenne est pour ce type de variable dénuée de signification. En ACM la variable n'est pas traitée telle quelle mais à travers ses modalités. Elle est en effet découpée en autant de modalités qu'elle possède et tout individu est alors codé 1 dans la modalité qu'il possède et zéro dans les autres (qu'il ne possède pas, les modalités étant exclusives). On réalise ainsi une tranformation de l'information, appelée codage disjonctif complet. Il s'agit bien d'un codage dans la mesure où l'information initiale est transformée, de plus il est disjonctif dans la mesure où tout individu possède au plus une modalité, et complet car tout individu a au moins une modalité.

Pour les variables quantitatives on procède de la même manière en découpant au préalable la variable en classes. Par exemple en considérant trois classes d'âge, moins de 35 ans (age1), de 35 à 49 ans (age2) et 50 ans et plus (age3).

Individu	age	age1	age2	age3	sexe	sexe1	sexe2
1	42	0	1	0	H	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	25	1	0	0	F	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	65	0	0	1	F	0	1

### 5.2.2 Notation

Soient  $n$  le nombre d'individus ou d'observations et  $Q$  le nombre de variables (ou questions, d'où la notation).

Chaque variable possède  $J_q$  modalités et le nombre total de modalités est égal à  $J$ .



### 5.2.3 Tableau de codage disjonctif complet

Le tableau logique (c'est à dire ne contenant que des uns ou des zéros) noté  $Z$  à  $n$  lignes et  $J$  colonnes correspondant est appelé tableau de codage disjonctif complet (CDC).

Remarque :

Ici il y a transformation des données initiales avec perte d'information dans le seul cas des variables continues. En effet la donnée de base est que l'individu 1 est âgé de 42 ans. L'information contenue dans  $Z$  est que l'individu 1 a un âge compris entre 35 et 49 ans. Nous verrons toutefois que pour certaines variables qualitatives il est parfois nécessaire de regrouper des modalités.

### 5.2.4 Tableau de codage condensé

A la place du tableau  $Z$  il est possible d'utiliser un tableau  $R[n, Q]$  appelé tableau de codage condensé. Sa taille est en effet celle du tableau d'origine. Le terme générique  $r_{iq}$  de ce tableau contient le numéro de la modalité de la variable  $q$  possédée par l'individu  $i$ . L'intérêt essentiel de ce tableau tient à son moindre encombrement. Ce tableau contient la même information que  $Z$ . Un tel tableau est notamment utilisé par un logiciel comme SPAD alors que d'autres exigent la constitution explicite du tableau  $Z$ .

### 5.2.5 Propriétés du tableau de codage disjonctif complet

La somme des éléments d'une même ligne est constante et vaut  $Q$  (le nombre de variables). Donc :

$$z_{i.} = \sum_{j=1}^J z_{ij} = \sum_{q=1}^Q \sum_{j=1}^{J_q} z_{ij} = Q$$

car

$$\sum_{j=1}^{J_q} z_{ij} = 1$$

en raison du caractère disjonctif et complet du codage.

Par conséquent :

$$z_{..} = \sum_{i=1}^n \sum_{j=1}^J z_{ij} = nQ$$

La somme de tous les éléments du tableau est égale à  $nQ$ .

La somme des éléments d'une même colonne n'est pas en revanche constante mais égale à l'effectif  $n_j$  possédant la modalité  $j$  de la variable  $q$  considérée :

$$z_{.j} = \sum_{i=1}^n z_{ij} = n_j$$

Modalité	...	1	...	$j_q$	...	$J_q$	...	
$\vdots$				$\vdots$				
$i$			...	$z_{ij}$	...			$z_{i.} = Q$
$\vdots$				$\vdots$				
				$z_{.j} = n_j$				$z_{..} = nQ$

### 5.2.6 Tableau de Burt

On appelle tableau de Burt, noté  $B$ , le produit du transposé de  $Z$  par  $Z$ , soit  $B = {}^tZZ$ .

Le tableau de Burt est carré et sa taille est égale au nombre total de modalités  $J$  possédées par les  $Q$  variables.

En utilisant les propriétés élémentaires du produit des matrices par blocs et le fait que le produit de  $Z$  par son transposé conduit à effectuer des produits d'indicateurs, on voit que le tableau de Burt se compose de :

- Blocs diagonaux  $B_{qq} = {}^tZ_qZ_q$  qui sont eux-mêmes diagonaux et qui contiennent sur la diagonale l'effectif dans chaque modalité pour la variable  $q$ .
- Blocs non diagonaux  $B_{qq'} (q \neq q')$  égaux à  ${}^tZ_qZ_{q'}$  qui ne sont autres que les tableaux de contingence croisant les variables  $q$  et  $q'$ .

Il est immédiat que le tableau de Burt est non seulement carré mais symétrique. Le bloc  $B_{q'q} = {}^tZ_{q'}Z_q$  n'est autre que le transposé du bloc  $B_{qq'} = {}^tZ_qZ_{q'}$ .

### 5.2.7 AFC du tableau de codage disjonctif complet

L'Analyse des Correspondances est née de la simple application d'un programme d'AFC à un tableau de données mises sous forme disjonctive complète. Dans cette optique le tableau traité est un tableau individus-variables

dans lequel les modalités viennent se substituer aux variables d'origine. En ce sens et de manière presque paradoxale l'ACM s'apparente plus dans sa problématique à l'ACP qu'à l'AFC même si techniquement elle constitue une extension de cette dernière.

Rappelons brièvement les résultats de l'AFC :

Dans  $\mathbf{R}^p$  le tableau analysé est  $X = D_n^{-1}F$  de terme générique  $\frac{f_{ij}}{f_{i.}}$

La métrique ambiante est  $M = D_p^{-1}$  de terme  $\frac{1}{f_{.j}}$

La matrice des masses est  $P = D_n$  de terme  $f_{i.}$

Dans ces conditions :

$$f_{i.} = \frac{z_{i.}}{z_{..}} = \frac{Q}{nQ} = \frac{1}{n}$$

$$f_{.j} = \frac{z_{.j}}{z_{..}} = \frac{n_j}{nQ}.$$

L'AFC du tableau  $Z$  se résume alors à l'analyse du triplet suivant :

- Données :  $X = \left(\frac{z_{ij}}{z_{i.}}\right) = \frac{1}{Q}Z$

- Métrique :  $M = \left(\frac{z_{..}}{z_{.j}}\right) = nQ\left(\frac{1}{n_j}\right) = nQD_J^{-1}$  où  $D_J$  désigne la matrice diagonale de terme  $n_j$  correspondant à l'effectif de la modalité  $j$

- Poids :  $P = \left(\frac{z_{i.}}{z_{..}}\right) = \left(\frac{Q}{nQ}\right) = \left(\frac{1}{n}\right) = \frac{1}{n}I$

La matrice à diagonaliser s'écrit :

$${}^tXPM = \frac{1}{Q}{}^tZZD_J^{-1}$$

Sur le plan technique notons que la matrice à diagonaliser  $\frac{1}{Q}{}^tZZD_J^{-1}$  est d'ordre  $J$  ( $J$  désignant le nombre total de modalités). Toutefois l'ensemble des variables possédant en commun le vecteur 1, la somme des modalités d'une même variables étant constante, le rang de cette matrice est égal  $J - (Q - 1)$ , soit  $J - Q$  après élimination du facteur trivial. Il est en conséquence possible

de se ramener à la diagonalisation d'une matrice d'ordre  $J - Q$  seulement. Cette propriété est exploitée par le logiciel Spad par exemple. A l'inverse certains logiciels n'utilisent pas cette propriété et l'ACM est alors réalisée par le même programme qu'en AFC, l'utilisateur devant alors construire explicitement le tableau de codage disjonctif complet  $Z$ .

### 5.2.8 Résumé

	Analyse dans $\mathbf{R}^J$	Analyse dans $\mathbf{R}^n$
Données	$X_{[n,J]} = \begin{pmatrix} z_{ij} \\ n_{i.} \end{pmatrix} = \frac{1}{Q}Z$	$Y_{[J,n]} = \begin{pmatrix} z_{ij} \\ z_{.j} \end{pmatrix} = D_J^{-1}Z$
Métrique	$M_{[J,J]} = \begin{pmatrix} z_{..} \\ z_{.j} \end{pmatrix} = nQD_J^{-1}$	$M_{[n,n]} = \begin{pmatrix} z_{..} \\ z_{i.} \end{pmatrix} = nI$
Poids	$P_{[n,n]} = \begin{pmatrix} z_{i.} \\ z_{..} \end{pmatrix} = \frac{1}{n}I$	$P_{[J,J]} = \begin{pmatrix} z_{.j} \\ z_{..} \end{pmatrix} = \frac{1}{nQ}D_J$

## 5.3 Propriétés

### 5.3.1 Inertie totale du nuage

Nous savons que l'inertie totale  $I_0$  du nuage est égale à la trace de la matrice d'inertie, soit ici  $I_0 = tr(\frac{1}{Q} {}^tZZD_J^{-1})$

Or  ${}^tZZ = B$  et  $b_{jj} = n_j$  effectif de la modalité  $j$ , d'où :

$$I_0 = tr(\frac{1}{Q} {}^tZZD_J^{-1}) = \frac{1}{Q} \sum_{j=1}^J \frac{b_{jj}}{n_j} = \frac{J}{Q}$$

De plus, si l'on se souvient que 1 est valeur propre triviale en AFC, on en déduit que l'inertie totale  $I_G$  est égale à  $\frac{J}{Q} - 1$ .

Sachant par ailleurs qu'en AFC les valeurs propres sont inférieures à 1, on comprend pourquoi les taux d'inertie sont souvent faibles en ACM et conduisent à une estimation pessimiste de la part d'information expliquée.

Prenons par exemple le cas de 10 variables. En ACP il n'est pas rare d'obtenir avec un effet taille 50 % de l'inertie totale sur le premier axe factoriel. Les mêmes données traitées par l'ACM sur la base de 5 modalités par variable, conduiront à une inertie totale de  $\frac{10 * 5}{10} - 1 = 4$ . Les valeurs propres étant inférieures à 1, le premier axe ne pourra conserver qu'au plus 25 % de l'inertie totale soit deux fois moins qu'en ACP pour un même phénomène.

### 5.3.2 Inertie d'une modalité

Cherchons en premier lieu les coordonnées du barycentre  $G$  nuage de  $\mathbf{R}^n$  formé par les  $J$  modalités de l'ensemble des variables  $q$ . Les coordonnées de la modalité  $j$  sont  $\frac{z_{ij}}{z_{.j}} = \frac{z_{ij}}{n_j}$  et le poids de la modalité  $j$  est  $\frac{z_{.j}}{z_{..}} = \frac{n_j}{nQ}$ .

Le barycentre a donc pour coordonnées :

$$\sum_{j=1}^J \frac{n_j}{nQ} \frac{z_{ij}}{n_j} = \frac{1}{nQ} \sum_{j=1}^J z_{ij} = \frac{1}{nQ} \sum_{q=1}^Q \sum_{j=1}^{J_q} z_{ij} = \frac{Q}{nQ} = \frac{1}{n} \text{ car } \sum_{j=1}^{J_q} z_{ij} = 1$$

(Tout individu possédant une modalité et une seule pour la variable  $q$ ).

Le barycentre  $G$  a donc comme coordonnées dans  $\mathbf{R}^n$ ,  $\frac{1}{n}$ . Il est également possible de montrer que toutes les variables possèdent le même barycentre.

Rappelons que dans  $\mathbf{R}^n$  la métrique est  $M = D_n^{-1} = nI$  et la matrice des masses  $P = D_p = \left( \frac{n_j}{nQ} \right) = \frac{1}{nQ} D_j$

d'où :

$$\begin{aligned} d_M^2(j, G) &= \left\| \frac{z_{ij}}{n_j} - \frac{1}{n} \right\|^2 = n \sum_{i=1}^n \left( \frac{z_{ij}}{n_j} - \frac{1}{n} \right)^2 \\ &= n \left( \sum_{i=1}^n \frac{z_{ij}^2}{n_j^2} + \frac{n}{n^2} - \frac{2}{nn_j} \sum_{i=1}^n z_{ij} \right) \end{aligned}$$

Or,  $z_{ij}$  prenant les valeurs 0 ou 1 il est immédiat que  $z_{ij}^2 = z_{ij}$ . Par conséquent :

$$\sum_{i=1}^n \frac{z_{ij}^2}{n_j^2} = \sum_{i=1}^n \frac{z_{ij}}{n_j^2} = \frac{1}{n_j^2} \sum_{i=1}^n z_{ij} = \frac{z_{.j}}{n_j^2} = \frac{n_j}{n_j^2} = \frac{1}{n_j}$$

d'autre part :

$$-\frac{2}{nn_j} \sum_{i=1}^n z_{ij} = -\frac{2z_{.j}}{nn_j} = -\frac{2n_j}{nn_j} = -\frac{2}{n}$$

d'où :

$$d_M^2(j, G) = n \left( \frac{1}{n_j} + \frac{1}{n} - \frac{2}{nn_j} \sum_{i=1}^n z_{ij} \right) = n \left( \frac{1}{n_j} - \frac{1}{n} \right)$$

Par ailleurs, le poids de la modalité  $j$  est  $m_j = \frac{z_{.j}}{nQ} = \frac{n_j}{nQ}$ .

L'inertie de la modalité  $j$  est donc :

$$m_j d_M^2(j, G) = \frac{n_j}{nQ} n \left( \frac{1}{n_j} - \frac{1}{n} \right) = \frac{1}{Q} \left( 1 - \frac{n_j}{n} \right)$$

soit :

$$I_j = \frac{1}{Q} \left( 1 - \frac{n_j}{n} \right)$$

Il apparaît immédiatement que l'inertie d'une modalité est une fonction décroissante de l'effectif de cette modalité. Le maximum  $\frac{1}{Q}$  est obtenu pour une modalité d'effectif nul.

Il sera donc nécessaire en pratique d'éviter les modalités à effectifs faibles. Ce problème sera évoqué lors de la partie consacrée au codage.

Remarque :

On vérifie que :

$$\sum_{j=1}^J I_j = \frac{J}{Q} - \frac{1}{nQ} \sum_{j=1}^J n_j = \frac{J}{Q} - 1 = I_G$$

### 5.3.3 Inertie d'une variable

Par définition, l'inertie d'une variable est égale à la somme des inerties de ses modalités.

$$I_q = \sum_{j=1}^{J_q} I_j = \sum_{j=1}^{J_q} \frac{1}{Q} \left( 1 - \frac{n_j}{n} \right) = \frac{1}{Q} (J_q - 1)$$

L'inertie totale d'une variable est donc une fonction croissante du nombre de ses modalités. D'où la nécessité d'équilibrer le nombre des modalités, du moins pour les variables actives.

### 5.3.4 Distance entre individus

Le nuage des individus est dans  $\mathbf{R}^J$  muni de la métrique  $M = D_p^{-1} = \left(\frac{nQ}{n_j}\right) = (nQ)D_j^{-1}$ . La distance entre deux individus  $i$  et  $i'$  s'écrit donc :

$$d^2(i, i') = \sum_{j=1}^J \frac{nQ}{n_j} \left( \frac{z_{ij}}{n_{i.}} - \frac{z_{i'j}}{n_{i'.}} \right)^2$$

Or  $n_{i.} = Q$  pour tout  $i = [1, n]$  d'où :

$$d^2(i, i') = \frac{1}{Q} \sum_{j=1}^J \frac{n}{n_j} (z_{ij} - z_{i'j})^2$$

Deux individus seront d'autant plus proches qu'ils possèdent de modalités communes, les termes  $(z_{ij} - z_{i'j})^2$  valant 0 ou 1. On remarque également que plus une modalité est rare et plus elle va contribuer à éloigner les individus les possédant. Par ailleurs, deux individus distincts mais possédant les mêmes modalités pour l'ensemble des variables, ne sont plus différenciés. Il s'agit là d'un effet direct du codage disjonctif complet qui va avoir tendance à gommer les différences entre individus. Cet effet peut même poser problème, notamment pour les variables continues, dans le cas d'individus atypiques. Les valeurs extrêmes seront alors regroupées dans la dernière modalité, ce qui peut expliquer la non-détection d'individus atypiques, à l'inverse de l'ACP.

### 5.3.5 Distance entre modalités

Les modalités ont comme coordonnées  $\frac{z_{ij}}{n_j}$  dans  $\mathbf{R}^n$  muni de la métrique  $M = D_n^{-1} = nI$ . La distance entre deux modalités  $j$  et  $k$  s'écrit en conséquence :

$$d_M^2(j, k) = \sum_{i=1}^n n \left( \frac{z_{ij}}{n_j} - \frac{z_{ik}}{n_k} \right)^2$$

En utilisant le fait que  $z_{ij} = z_{ij}^2$  et que  $z_{.j} = n_j$ , alors :

$$\sum_{i=1}^n \left( \frac{z_{ij}}{n_j} \right)^2 = \frac{1}{n_j^2} \sum_{i=1}^n z_{ij} = \frac{1}{n_j^2} z_{.j} = \frac{1}{n_j}$$

Par ailleurs, le terme  $z_{ij}z_{ik}$  vaut 1 si  $i$  possède les modalités  $j$  et  $k$ , 0 sinon. Dans ces conditions :

$$\sum_{i=1}^n z_{ij}z_{ik} = n_{jk}$$

où  $n_{jk}$  désigne le nombre d'individus possédant les modalités  $j$  et  $k$ .  
d'où :

$$d_M^2(j, k) = n \sum_{i=1}^n \left( \frac{1}{n_j} + \frac{1}{n_k} - \frac{2n_{jk}}{n_j n_k} \right)$$

soit encore :

$$d_M^2(j, k) = \frac{n}{n_j n_k} (n_j + n_k - 2n_{jk}) = \frac{n}{n_j n_k} ((n_j - n_{jk}) + (n_k - n_{jk}))$$

le terme  $n_j - n_{jk}$  est égal au nombre d'individus possédant la modalité  $j$  moins le nombre de ceux possédant  $j$  et  $k$ . Il est donc égal au nombre de ceux possédant  $j$  et ne possédant pas  $k$ . Soit  $n_{j\bar{k}}$  ce nombre. De même, soit  $n_{\bar{j}k} = n_k - n_{jk}$  le nombre d'individus possédant la modalité  $k$  et ne possédant pas  $j$ . Dans ces conditions,  $n_{j,k} = n_{j\bar{k}} + n_{\bar{j}k}$  représente le nombre d'individus possédant la modalité  $j$  ou bien (ou exclusif) la modalité  $k$ . Alors :

$$d_M^2(j, k) = \frac{n}{n_j n_k} n_{j,k}$$

$n_{j,k}$  désignant le nombre d'individus possédant une seule des deux modalités  $j$  ou  $k$ .

Compte tenu du caractère disjonctif du codage il apparaît immédiatement que deux modalités d'une même variable ne peuvent être à distance nulle. Par ailleurs, il est important de remarquer que deux modalités possédées par les mêmes individus ne se distinguent plus. Il s'agit là d'un effet direct du codage disjonctif complet.

### 5.3.6 Propriétés des modalités

L'ACM se résumant dans cette approche à une AFC du tableau individus-modalités, les relations barycentriques de l'AFC confèrent aux modalités un rôle particulier. La coordonnée d'une modalité  $j$  sur l'axe de rang  $k$  s'écrit :

$$d_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^n \frac{z_{ij}}{n_j} c_{ik}$$

Les termes  $z_{ij}$  prenant la valeur 0 ou 1 cette relation devient :

$$d_{jk} = \frac{1}{\sqrt{\lambda_k}} \frac{1}{n_j} \sum_{i \in E_j} c_{ik}$$

où  $E_j$  désigne l'ensemble des  $n_j$  individus possédant la modalité  $j$ . Ainsi, la coordonnée sur l'axe de rang  $k$  d'une modalité se trouve, au facteur  $\frac{1}{\sqrt{\lambda_k}}$  près, au barycentre des coordonnées sur ce même axe, des seuls individus possédant cette modalité. Cette propriété est à la base d'un test pour les modalités illustratives qui sera présenté plus tard.



## 5.4 Le codage des variables

Cette phase préliminaire, inexistante pour certaines méthodes telle l'ACP, est fondamentale. Ce qui importe ici n'est pas la constitution du tableau de codage disjonctif complet contenant 0 ou 1, cette opération est réalisée informatiquement, mais le choix des bornes des classes.

Nous distinguerons à cet effet deux cas, celui des variables continues et celui des variables qualitatives. Rappelons en premier lieu qu'il convient d'éviter les modalités à effectif faible (voir point précédent).

### 5.4.1 Les variables continues

En ce qui concerne les variables continues, le mieux est de commencer à réfléchir sur les bornes qui paraissent pertinentes en regard du problème étudié et à observer les distributions correspondantes (histogrammes). Prenons à titre d'exemple le cas de la variable temps de travail sur une exploitation agricole. Il semble plus pertinent de distinguer les exploitants à temps complet des autres. Il faut en conséquence prévoir une modalité explicite « à temps plein sur l'exploitation ». Le codage direct sans réflexion préalable sur le sujet risque de déboucher sur des classes du type « de  $\frac{3}{4}$  de temps à temps complet sur l'exploitation » alors que l'on peut penser que les exploitants à temps complet sur l'exploitation se distinguent des autres. De même si l'on s'intéresse à l'activité du conjoint sur l'exploitation il apparaît préférable de créer deux modalités « sans activité sur l'exploitation » et « moins de  $\frac{1}{4}$  de temps sur l'exploitation » plutôt que de regrouper les deux. Il est en effet probable que la situation des conjoints n'ayant aucune activité sur l'exploitation se distingue de celle des conjoints ayant une activité certes modeste mais néanmoins réelle.

Cette phase préliminaire requiert, comme l'Analyse des Données elle-même, une bonne connaissance du domaine étudié. Si dans certains cas cette démarche n'aboutit pas, il est toujours possible de recourir à la technique qui consiste à découper la variable en modalités d'effectifs égaux. Dans cette optique on part des effectifs pour en déduire les bornes de classes. Le problème inhérent à cette approche est de conduire à des modalités peu pertinentes. Cette démarche doit plutôt être envisagée en dernier recours.

### 5.4.2 Le codage des variables qualitatives

En principe ici le choix des classes ne se pose pas. Le seul problème qui puisse se présenter est que les modalités naturelles conduisent à des effectifs déséqui-

librés ou faibles. En ce cas il est nécessaire de procéder à des regroupements ce qui parfois peut s'avérer problématique. Ainsi dans le cas de l'enquête sur la structure des exploitations agricoles réalisée tous les deux ans auprès d'un échantillon, cinq modalités sont prévues pour le régime d'imposition. Certaines modalités contenant des effectifs trop faibles, il a fallu opérer des regroupements. Une même modalité réunit ainsi des régimes très différents (forfait collectif, régime transitoire et statut particulier). En fait on distingue ici trois modalités, le réel normal (grandes exploitations), le réel simplifié (unités moyennes) et une catégorie autres.

Il semble plus souhaitable d'opérer des regroupements plutôt que d'utiliser une méthode qui consiste à ventiler de manière aléatoire les effectifs des modalités à effectif faible dans les autres modalités, solution qui est parfois proposée.

Enfin la contribution à l'inertie d'une variable étant une fonction croissante du nombre de ses modalités il est préférable de ne pas multiplier les modalités pour certaines variables. Ce dernier point est toutefois la plupart du temps ignoré et ne présente pas la même importance que le problème des effectifs faibles dans certaines modalités.

Remarque :

Un point important concernant l'ACM est de s'assurer de la validité des résultats. En d'autres termes les résultats obtenus sont-ils stables ou dépendent-ils au contraire du codage retenu ? Afin de s'assurer de leur robustesse, une bonne méthode consiste à bousculer les bornes de classe voire opérer des regroupements de modalités. De telles modifications, si les résultats observés sont robustes, ne devraient pas apporter de grands bouleversements.

## 5.5 Equivalence pour l'acquisition des facteurs

Nous allons montrer que les résultats d'une analyse des correspondances multiples peuvent être obtenus soit par une AFC du tableau de codage disjonctif complet soit par une AFC du tableau de Burt.

### 5.5.1 AFC du tableau $Z$ (Rappel)

L'AFC de  $Z$  conduit à chercher les vecteurs propres de la matrice  ${}^tZZD_j^{-1}$ . Dans  $\mathbf{R}^n$  l'analyse duale conduit à diagonaliser la matrice  $ZD_j^{-1}{}^tZ$ . Nous savons que ces deux matrices admettent les mêmes valeurs propres et que les vecteurs propres dans un espace se déduisent de ceux dans l'autre espace par les formules de transition.

### 5.5.2 AFC du tableau de Burt $B(= {}^tZZ)$

Ici analyse directe et analyse duale coïncident le tableau étant symétrique et les matrices diagonales  $D_n$  et  $D_p$  identiques.

La somme des éléments d'une même ligne (ou d'une même colonne) du tableau  $B$  vaut  $Q$  fois  $n_j$ .

En effet dans chaque sous-tableau le total de la ligne  $j$  donne  $n_j$ . Comme il y a  $Q$  sous-tableaux le total vaut donc  $Qn_j$ .

La somme des éléments d'un sous-tableau vaut  $n$ . Comme il y a  $Q^2$  sous-tableaux le total des éléments de  $B$  est égal à  $nQ^2$ . D'où  $F = \left(\frac{1}{nQ^2}\right) B$

Le terme générique  $f_{.j}$  de  $F$  s'écrit  $\left(\frac{b_{.j}}{nQ^2}\right) = \left(\frac{Qn_j}{nQ^2}\right) = \frac{n_j}{nQ}$ , soit sous forme matricielle  $D_p = \frac{1}{nQ} D_j$  où  $D_j$  désigne la matrice diagonale de terme  $n_j$ .

d'où :

$$\begin{aligned} {}^t F D_n^{-1} F D_p^{-1} &= \frac{1}{nQ^2} {}^t B (nQ) D_j^{-1} \left(\frac{1}{nQ^2}\right) B (nQ) D_j^{-1} \\ &= \frac{1}{Q^2} {}^t B D_j^{-1} B D_j^{-1} \\ &= \frac{1}{Q^2} {}^t Z Z D_j^{-1} {}^t Z Z D_j^{-1} \end{aligned}$$

Il est immédiat que cette dernière matrice et la matrice  ${}^t Z Z D_j^{-1}$  possèdent les mêmes vecteurs propres.

En effet, soit  $u$  vérifiant  ${}^t Z Z D_j^{-1} u = \lambda u$ , alors :

$${}^t Z Z D_j^{-1} {}^t Z Z D_j^{-1} u = {}^t Z Z D_j^{-1} \lambda u = \lambda^2 u$$

L'analyse de  $Z$  ou de  $B$  fournit donc les mêmes vecteurs propres et chaque valeur propre de  $B$  est le carré de son homologue du tableau  $Z$  :

$$\lambda(B) = \lambda^2(Z)$$

### 5.5.3 Relation entre les coordonnées des modalités dans l'analyse de $Z$ et de $B$

Les coordonnées des colonnes sur l'axe de rang  $k$  s'écrivent en AFC :  $\sqrt{\lambda_k} D_p^{-1} u_k$   
Or,  $D_p^{-1} = nQ D_j^{-1}$

Dans l'AFC de  $Z$  on obtient donc  $\sqrt{\lambda_k} n Q D_j^{-1} u_k$

Dans celle de  $B$  :  $\lambda_k n Q D_j^{-1} u_k$

Les coordonnées des modalités dans l'AFC de  $B$  sont égales à celles issues de l'AFC de  $Z$  multipliées par  $\sqrt{\lambda_k}$ .

## 5.6 Un cas particulier

Lorsque toutes les variables possèdent deux modalités l'ACM présente quelques propriétés particulières. En premier lieu l'inertie totale du nuage  $I_G$  est égale à  $\frac{2Q}{Q} - 1$  soit 1. En second lieu si l'on se rappelle que les diverses modalités d'une même variable ont leur barycentre à l'origine, les représentations graphiques forment un système de haltères. Fk

f2.

o Fj f1.

En effet pour que ce système soit en équilibre à l'origine il faut que le poids de la modalité 1 soit plus important que celui de la modalité 2. Or en AFC le poids afférent à chaque modalité n'est autre que la fréquence de cette modalité dans la population. En conséquence une modalité est d'autant plus rare qu'elle s'éloigne de l'origine. Dans notre exemple si la variable représentée est le sexe il apparaît immédiatement que la population comporte plus d'hommes que de femmes (La modalité 1 désignant les hommes).

## 5.7 Le cas binaire

Nous allons montrer l'équivalence des analyses suivantes dans le cas de deux variables :

- 1- AFC du tableau de codage disjonctif complet  $Z[n, J]$  ( $J = J_1 + J_2$ )
- 2- AFC du tableau de Burt  $B[J, J] = {}^t Z Z$
- 3- AFC du tableau de contingence  $C[J_1, J_2]$

L'équivalence entre (1) et (2) ayant été montrée il suffit ici de montrer l'équivalence entre (1) et (3). Le tableau  $Z$  peut s'écrire sous la forme  $[Z_1, Z_2]$  avec  $J = J_1 + J_2$  où  $J_1$  et  $J_2$  désignent le nombre de modalités de chacun des deux caractères.

### 5.7.1 AFC du tableau disjonctif complet $Z$

Notons respectivement  $D_1$  et  $D_2$  les matrices diagonales respectivement d'ordre  $J_1$  et  $J_2$  dont les termes sont les effectifs dans chaque modalité. Il est immé-

diat que :

$$D_1 = {}^t Z_1 Z_1$$

$$D_2 = {}^t Z_2 Z_2$$

Il suffit ensuite d'appliquer l'équation précédente,  $\frac{1}{Q} {}^t Z Z D_J^{-1} u = \mu u$  au cas  $Q = 2$ . En utilisant les règles du produit par blocs on obtient :

$$\frac{1}{2} \begin{bmatrix} {}^t Z_1 Z_1 D_1^{-1} & {}^t Z_1 Z_2 D_2^{-1} \\ {}^t Z_2 Z_1 D_1^{-1} & {}^t Z_2 Z_2 D_2^{-1} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mu \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Soit :

$$\begin{aligned} \frac{1}{2} u_1 + \frac{1}{2} {}^t Z_1 Z_2 ({}^t Z_2 Z_2)^{-1} u_2 &= \mu u_1 \\ \frac{1}{2} {}^t Z_2 Z_1 ({}^t Z_1 Z_1)^{-1} u_1 + \frac{1}{2} u_2 &= \mu u_2 \end{aligned}$$

De la première équation il vient :

$$({}^t Z_1 Z_2) ({}^t Z_2 Z_2)^{-1} u_2 = (2\mu - 1) u_1$$

soit :

$$u_1 = \frac{1}{2\mu - 1} ({}^t Z_1 Z_2) ({}^t Z_2 Z_2)^{-1} u_2$$

En reportant dans la seconde équation on obtient :

$$\frac{1}{2\mu - 1} {}^t Z_2 Z_1 ({}^t Z_1 Z_1)^{-1} ({}^t Z_1 Z_2) ({}^t Z_2 Z_2)^{-1} u_2 = (2\mu - 1) u_2$$

ce qui entraîne :

$${}^t Z_2 Z_1 ({}^t Z_1 Z_1)^{-1} ({}^t Z_1 Z_2) ({}^t Z_2 Z_2)^{-1} u_2 = (2\mu - 1)^2 u_2$$

### 5.7.2 AFC du tableau de contingence $C$

L'AFC revient à analyser le tableau de contingence  ${}^t Z_1 Z_2$ . Plaçons-nous dans  $\mathbf{R}^p$  :

Le triplet est  $X = D_n^{-1} F$ ,  $M = D_p^{-1}$ ,  $P = D_n$

soit ici  $F = \frac{1}{n} {}^t Z_1 Z_2$ ,  $D_p = \frac{1}{n} {}^t Z_2 Z_2$ ,  $D_n = \frac{1}{n} {}^t Z_1 Z_1$

d'où  $X = ({}^tZ_1Z_1)^{-1}({}^tZ_1Z_2)$ ,  $M = n({}^tZ_2Z_2)^{-1}$  et  $P = \frac{1}{n}{}^tZ_1Z_1$

L'AFC revient à chercher les vecteurs propres de  ${}^tFD_n^{-1}FD_p^{-1}u = \lambda u$ , soit :

$$\frac{1}{n}{}^tZ_2Z_1n({}^tZ_1Z_1)^{-1}\frac{1}{n}{}^tZ_1Z_2n({}^tZ_2Z_2)^{-1}u = \lambda u$$

soit encore :

$${}^tZ_2Z_1({}^tZ_1Z_1)^{-1}{}^tZ_1Z_2({}^tZ_2Z_2)^{-1}u = \lambda u$$

En rapprochant les deux égalités on voit que les vecteurs propres sont les mêmes et que les valeurs propres vérifient la relation :

$$\lambda = (2\mu - 1)^2 \text{ ou } \mu = \frac{1 \pm \sqrt{\lambda}}{2}$$

L'AFC du tableau de contingence conduit à  $J_2 - 1$  facteurs (en supposant sans perte de généralité  $J_2 < J_1$ ). L'AFC du tableau  $Z$  conduit à  $J_1 + J_2 - 2$  facteurs non triviaux.

Dans cette dernière analyse il y a en fait  $J_2 - 1$  facteurs associés aux valeurs propres de la forme  $\frac{1 + \sqrt{\lambda}}{2}$  et  $J_2 - 1$  associés à celles de la forme  $\frac{1 - \sqrt{\lambda}}{2}$ . Enfin  $J_1 - J_2$  facteurs associés à la valeur propre  $\frac{1}{2}$  ( $\lambda = 0$ ) viennent compléter l'ensemble des solutions. En conséquence dans l'AFC du tableau  $Z$  seules les valeurs propres supérieures à  $\frac{1}{2}$  sont à prendre en compte.

Exemple : Considérons le cas du tableau de contingence traité dans le chapitre consacré à l'AFC et ventilant les étudiants selon la profession de leur parents (11 modalités) et leur discipline d'inscription (9 modalités). L'AFC de ce tableau conduit bien à  $J_2 - 1 = 8$  facteurs après élimination du facteur trivial.

Supposons qu'au lieu d'analyser le tableau de contingence  $C = ({}^tZ_1Z_2)$  nous ayons analysé par l'AFC le tableau de codage disjonctif  $Z (= [Z_1Z_2])$  complet comportant 15856 lignes et  $J_1 + J_2$  soit 20 colonnes (à condition de disposer des données individuelles ce qui est un autre problème).

Ces deux tableaux véhiculent exactement la même information il est en conséquence logique d'obtenir les mêmes résultats que l'on analyse l'un ou l'autre. Dans l'analyse de  $Z$  nous aurions obtenu  $J_2 - 1$  soit 8 valeurs propres supérieures à  $\frac{1}{2}$ , 8 autres inférieures à  $\frac{1}{2}$  et enfin  $J_1 - J_2$  soit 2 égales à  $\frac{1}{2}$ .

## 5.8 Conclusion

Dans ce chapitre nous n'avons en fait qu'esquissé le problème. La construction de l'ACM en tant que méthode spécifique requiert la présentation de l'analyse canonique généralisée dont elle constitue un cas particulier.

## 5.9 Exemples

Deux exemples vont être présentés. Le premier concerne la reprise de l'exemple utilisé lors de la présentation de l'ACP et a une fonction essentiellement pédagogique. Il permet en effet d'illustrer les principales caractéristiques et les problèmes rencontrés en ACM, à savoir choix des bornes de classe lors de la phase préliminaire de codage, chute du taux d'inertie sur les premiers axes et complexité, ici très relative, de l'interprétation. Le second exemple concerne un cas réel d'application de l'ACM au traitement d'une enquête sur les haies de Bretagne réalisée en 1996.

### 5.9.1 Premier exemple

Les huit variables continues ont chacune été découpées en trois classes. Ce choix n'a pas été effectué au hasard et tient probablement compte des résultats de l'ACP réalisée auparavant sur le même sujet. L'ACP constituant d'ailleurs un excellent préalable dans la mesure où d'une part elle permet une éventuelle diminution des variables continues à prendre en compte et où d'autre part elle peut fournir des indications quant à la constitution des classes. Le choix des modalités a été réalisé ici en constituant des effectifs égaux, méthode pourtant à utiliser en dernier ressort. Seul le triangle inférieur, diagonale incluse, du tableau de Burt a été édité ici. Rappelons que les blocs diagonaux  $B_{qq}$  croisant une variable avec elle-même sont eux-mêmes diagonaux et fournissent les effectifs dans chacune des modalités de la variable correspondante. Les blocs non diagonaux  $B_{qq'}$  avec  $q \neq q'$  sont les tableaux de contingence croisant les variables  $q$  et  $q'$ . Ici le premier bloc,  $B_{21}$ , croise les variables int et net. On vérifie que la somme des éléments est bien égale à l'effectif total  $n$ , ici égal à 15.

# 118 CHAPITRE 5. ANALYSE DES CORRESPONDANCES MULTIPLES

EDITION DU TABLEAU DE BURT

	net1	net2	net3	int1	int2	int3	sub1	sub2	sub3	lmt1	lmt2	lmt3	dct1	dct2	dct3	imm1	imm2	imm3	vr1	vr2	vr3
net1	5	0	0																		
net2	0	5	0																		
net3	0	0	5																		
int1	3	2	0	5	0	0															
int2	2	1	2	0	5	0															
int3	0	2	3	0	0	5															
sub1	1	0	4	1	1	3	5	0	0												
sub2	2	2	1	1	2	2	0	5	0												
sub3	2	3	0	3	2	0	0	0	5												
lmt1	2	2	1	1	2	2	1	1	3	5	0	0									
lmt2	0	2	3	1	2	2	2	2	1	0	5	0									
lmt3	3	1	1	3	1	1	2	2	1	0	0	5									
dct1	0	0	5	0	2	3	4	1	0	1	3	1	5	0	0						
dct2	2	3	0	1	2	2	0	3	2	2	1	2	0	5	0						
dct3	3	2	0	4	1	0	1	1	3	2	1	2	0	0	5						
imm1	2	3	0	2	2	1	0	1	4	4	1	0	0	2	3	5	0	0			
imm2	2	2	1	3	1	1	1	3	1	0	2	3	1	2	2	0	5	0			
imm3	1	0	4	0	2	3	4	1	0	1	2	2	4	1	0	0	0	5			
vr1	3	2	0	3	1	1	1	3	1	1	0	4	0	3	2	1	3	1	5	0	0
vr2	1	2	2	2	2	1	2	0	3	2	3	0	2	0	3	3	0	2	0	5	0
vr3	1	1	3	0	2	3	2	2	1	2	2	1	3	2	0	1	2	2	0	0	5
exp1	0	0	5	0	2	3	4	1	0	1	3	1	5	0	0	0	1	4	0	2	3
exp2	3	2	0	4	1	0	1	2	2	0	1	4	0	2	3	1	3	1	4	1	0
exp3	2	3	0	1	2	2	0	2	3	4	1	0	0	3	2	4	1	0	1	2	2
exp1	exp2	exp3																			
exp1	5	0	0																		
exp2	0	5	0																		
exp3	0	0	5																		
exp1	exp2	exp3																			

Rappelons qu'en ACM l'inertie totale du nuage après élimination du facteur trivial est égale à  $\frac{J}{Q} - 1$ . Ici chaque variable comporte trois modalités d'où  $I_G = 3 - 1 = 2$ .

EDITION DES VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 2.0000  
SOMME DES VALEURS PROPRES .... 2.0000

HISTOGRAMME DES 14 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE
1	.6272	31.36	31.36
2	.3993	19.97	51.33
3	.2834	14.17	65.50
4	.1824	9.12	74.61
5	.1485	7.42	82.04
6	.1159	5.80	87.83
7	.0870	4.35	92.19
8	.0509	2.54	94.73
9	.0363	1.81	96.54
10	.0322	1.61	98.15
11	.0207	1.04	99.19
12	.0126	.63	99.82
13	.0036	.18	100.00
14	.0000	.00	100.00



COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES MODALITES ACTIVES SUR LES AXES 1 A 5																	
MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
<b>1. Situation nette</b>																	
net1 - <13	4.17	2.00	- .73	-.39	.17	-.86	-.37	3.6	1.6	.4	17.0	3.8	.27	.08	.01	.37	.07
net2 - 13-15	4.17	2.00	-.64	.45	-.35	.85	.44	2.7	2.1	1.8	16.6	5.4	.20	.10	.06	.36	.10
net3 - >=15	4.17	2.00	1.37	-.06	.18	.01	-.07	12.4	.0	.5	.0	.1	.94	.00	.02	.00	.00
			CONTRIBUTION CUMULEE =					18.7	3.8	2.7	33.6	9.4					
<b>2. Interets</b>																	
int1 - <2.8	4.17	2.00	-.90	-.50	.75	.23	.26	5.3	2.6	8.2	1.3	2.0	.40	.13	.28	.03	.04
int2 - 2.8-3.5	4.17	2.00	.13	.35	-.13	-.34	-1.20	.1	1.2	3	2.6	40.4	.01	.06	.01	.06	.72
int3 - >=3.5	4.17	2.00	.76	.16	-.62	.10	.93	3.9	.3	5.6	.2	24.5	.29	.01	.19	.01	.44
			CONTRIBUTION CUMULEE =					9.3	4.1	14.0	4.1	66.9					
<b>3. Subventions</b>																	
sub1 - <2	4.17	2.00	1.02	-.34	.50	-.30	.46	7.0	1.2	3.6	2.0	6.0	.53	.06	.12	.04	.11
sub2 - 2-2.4	4.17	2.00	-.21	-.41	-.96	.27	-.39	.3	1.7	13.4	1.7	4.3	.02	.08	.46	.04	.08
sub3 - >=2.4	4.17	2.00	-.81	.74	.46	.02	-.07	4.4	5.8	3.1	.0	.1	.33	.28	.11	.00	.00
			CONTRIBUTION CUMULEE =					11.6	8.7	20.2	3.7	10.4					
<b>4. Endettement a long et moyen terme</b>																	
lmt1 - <7.8	4.17	2.00	-.27	1.03	-.17	-.69	-.29	.5	11.0	.4	10.9	2.4	.04	.53	.01	.24	.04
lmt2 - 7.8-9.8	4.17	2.00	.61	.16	.12	1.06	-.46	2.5	.3	.2	25.7	5.8	.19	.01	.01	.56	.10
lmt3 - >=9.8	4.17	2.00	-.34	-1.19	.04	-.37	.16	.8	14.7	.0	3.1	.7	.06	.71	.00	.07	.01
			CONTRIBUTION CUMULEE =					3.8	26.0	.7	39.7	9.0					
<b>5. Endettement a court terme</b>																	
dct1 - <22	4.17	2.00	1.37	-.06	.18	.01	-.07	12.4	.0	.5	.0	.1	.94	.00	.02	.00	.00
dct2 - 22-23.9	4.17	2.00	-.53	.02	-1.09	-.07	.01	1.9	.0	17.4	.1	.0	.14	.00	.59	.00	.00
dct3 - >=23.9	4.17	2.00	-.84	.04	.91	.07	.06	4.6	.0	12.1	.1	.1	.35	.00	.41	.00	.00
			CONTRIBUTION CUMULEE =					19.0	.1	30.0	.2	.3					
<b>6. Immobilisations</b>																	
imm1 - <19.5	4.17	2.00	-.72	1.10	.16	-.15	.08	3.5	12.6	.4	.5	.2	.26	.61	.01	.01	.00
imm2 - 19.5-24	4.17	2.00	-.36	-.85	-.32	.61	-.19	.9	7.5	1.5	8.5	1.0	.07	.36	.05	.19	.02
imm3 - >=24	4.17	2.00	1.09	-.25	.16	-.46	.11	7.8	.7	.4	4.9	.4	.59	.03	.01	.11	.01
			CONTRIBUTION CUMULEE =					12.2	20.8	2.2	13.9	1.5					
<b>7. Valeurs realisables et disponibles</b>																	
vr1 - <16.5	4.17	2.00	-.75	-.93	-.29	-.22	.22	3.7	9.1	1.							

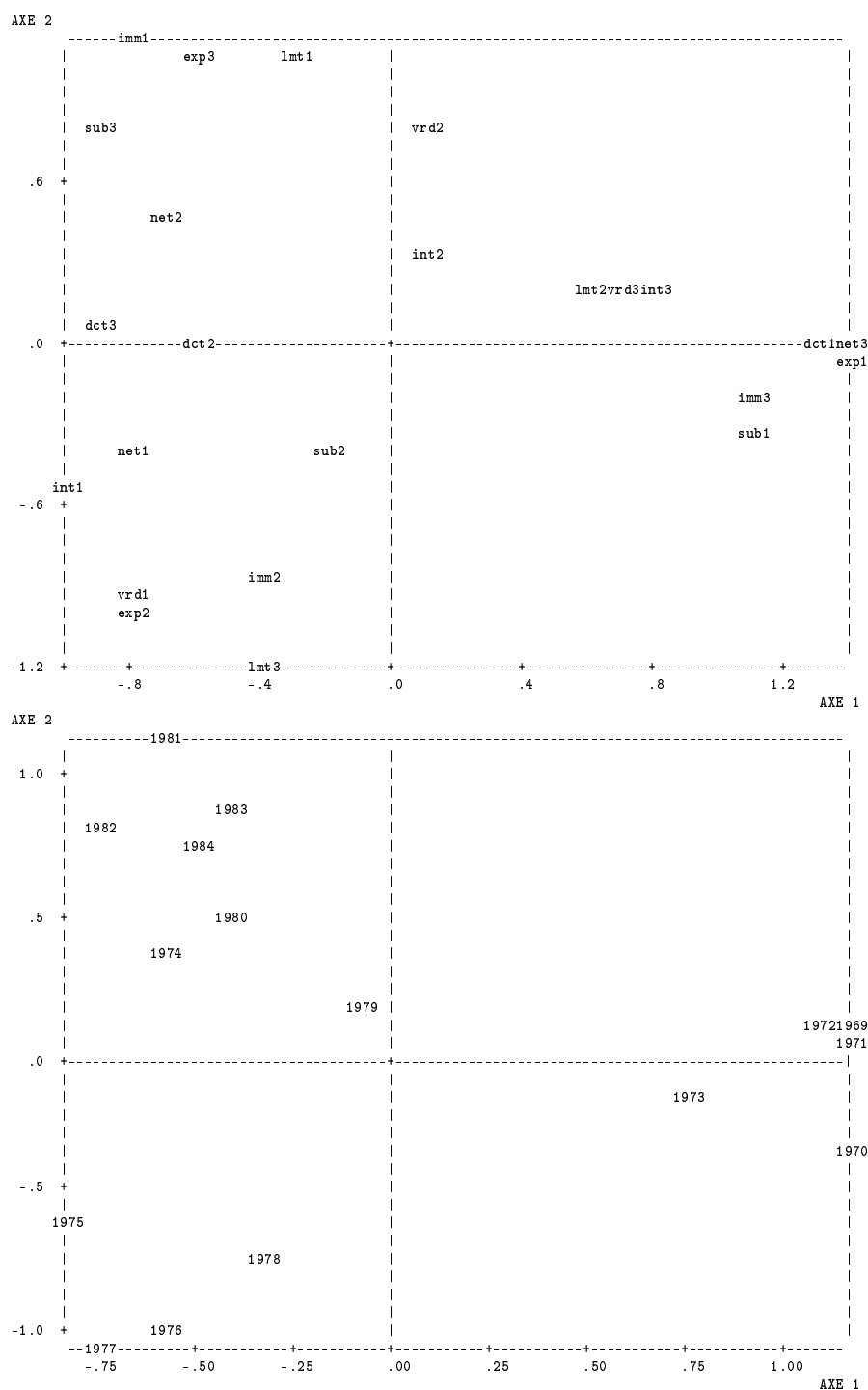
## 120 CHAPITRE 5. ANALYSE DES CORRESPONDANCES MULTIPLES

L'examen du nuage des individus indique les mêmes éléments qu'en ACP à savoir les années antérieures à 1973, qui contribuent pour près de 60 % à l'inertie conservée par le premier axe. La signification de ce premier axe est donc identique quelle que soit la méthode retenue, ACP ou AFC.

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS SUR LES AXES 1 A 5 - INDIVIDUS ACTIFS

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1969	6.67	2.00	1.16	.12	-.07	-.42	.48	14.3	.2	.1	6.4	10.2	.67	.01	.00	.09	.11
1970	6.67	2.00	1.15	-.32	-.02	-.32	.43	14.1	1.7	.0	3.8	8.4	.66	.05	.00	.05	.09
1971	6.67	2.00	1.21	.06	.40	.22	.24	15.6	.1	3.8	1.8	2.6	.74	.00	.08	.02	.03
1972	6.67	2.00	1.11	.09	.52	.09	-.45	13.2	.2	6.3	.3	9.2	.62	.00	.13	.00	.10
1973	6.67	2.00	.78	-.15	-.34	.45	-.83	6.4	.4	2.7	7.3	31.2	.30	.01	.06	.10	.35
1974	6.67	2.00	-.63	.35	.80	.75	.06	4.2	2.0	15.0	20.5	.2	.20	.06	.32	.28	.00
1975	6.67	2.00	-.81	-.64	-.10	.35	.26	6.9	6.9	.3	4.4	3.1	.33	.21	.01	.06	.03
1976	6.67	2.00	-.58	-1.02	.49	-.21	.19	3.6	17.2	5.8	1.6	1.6	.17	.52	.12	.02	.02
1977	6.67	2.00	-.78	-1.03	.15	-.04	-.09	6.4	17.7	.6	.1	.3	.30	.53	.01	.00	.00
1978	6.67	2.00	-.34	-.75	-.41	-.56	-.48	1.2	9.4	4.0	11.6	10.5	.06	.28	.08	.16	.12
1979	6.67	2.00	-.05	.15	-1.05	.75	.10	.0	.4	25.7	20.7	.5	.00	.01	.55	.28	.01
1980	6.67	2.00	-.46	.48	-.90	-.01	.55	2.3	3.9	19.1	.0	13.4	.11	.12	.41	.00	.15
1981	6.67	2.00	-.57	1.09	.32	-.01	-.13	3.5	19.7	2.4	.0	.7	.16	.59	.05	.00	.01
1982	6.67	2.00	-.75	.75	.65	-.35	.09	6.0	9.4	9.8	4.4	.3	.28	.28	.21	.06	.00
1983	6.67	2.00	-.45	.81	-.43	-.68	-.42	2.2	10.8	4.4	17.1	7.8	.10	.32	.09	.23	.09

L'interprétation du second axe est en revanche un peu différente. Se manifestent à la fois les années de l'entre deux chocs pétroliers, 1975 à 1978, marquées par un fort endettement à long et moyen terme, mais aussi les dernières années à partir de 1981. Du côté des variables on observe l'importance de LMT, mais qui se manifeste également sur le quatrième axe. L'appariement (2\*4) constitue peut-être le bon plan pour observer l'endettement à long et moyen terme. Sur ce second axe interviennent aussi les variables valeurs d'exploitation et immobilisations.



Sur cet exemple d'application de l'ACM, quelque peu artificiel, en raison d'une part de la nature des données et de la dimension réduite du nombre

d'observations, apparaissent toutefois quelques traits spécifiques de la méthode. En premier lieu une dilution plus grande des phénomènes qui rend souvent l'interprétation plus difficile. Cette dilution résulte à la fois de la dimension de l'espace dans lequel les données sont représentées, la substitution des modalités aux variables en étant la cause, et de la mise en classes des variables continues qui contribue à diminuer l'intensité de certains phénomènes. L'ACP utilise un critère d'ajustement quadratique ce qui contribue à amplifier le rôle joué par quelques éléments extrêmes<sup>1</sup>. A l'inverse le codage disjonctif complet atténue le rôle joué par les valeurs extrêmes, ce qui explique parfois le caractère plus diffus des phénomènes mis en évidence.

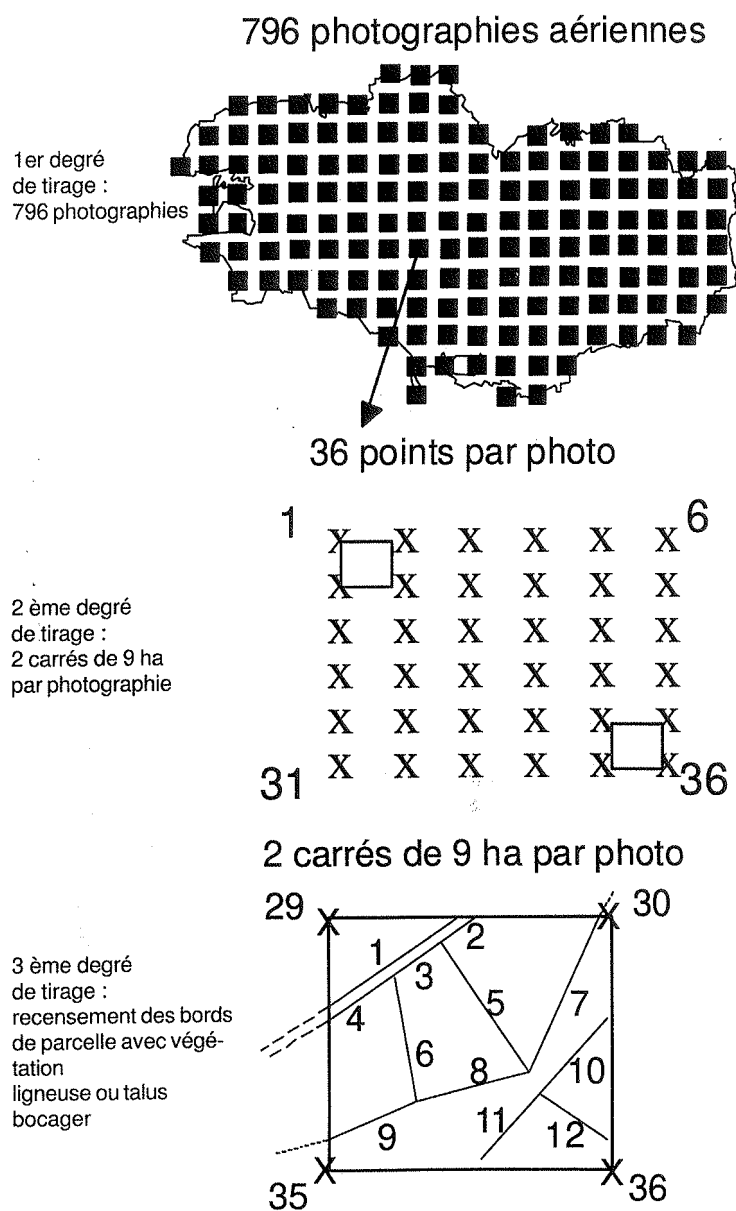
### 5.9.2 L'enquête sur les haies de Bretagne

Ce second exemple concerne l'enquête sur les haies de Bretagne réalisée en 1996 par le Ministère de l'Agriculture et l'Institut pour le Développement Forestier (IDF). L'objet de cette enquête était à partir du recueil sur le terrain d'une série d'indicateurs, de quantifier l'importance des haies en Bretagne en nombre et en linéaire, de dresser une typologie a posteriori, de préciser leurs fonctions essentielles et enfin de localiser les principaux types rencontrés. La méthode : L'enquête a été réalisée par sondage à partir de l'enquête sur l'utilisation du territoire (Ter-Uti) réalisée tous les ans par les services statistiques du Ministère de l'Agriculture. L'enquête « Teruti » est une enquête par sondage dans laquelle l'échantillon est constitué d'un ensemble de points du territoire. Cet échantillon résulte d'un tirage systématique à deux degrés de manière à assurer une répartition satisfaisante sur l'ensemble du territoire. Au premier degré, on tire un échantillon de photographies aériennes préparées par l'Institut Géographique National, réparties régulièrement sur l'ensemble du territoire métropolitain. Ces photographies, distantes de 6 km les unes des autres, couvrent un carré de 1 800 m de côté. En Bretagne l'échantillon comporte 796 photos.

---

1. Il s'agit d'ailleurs d'un problème posé par l'ACP qui conduit parfois à analyser non pas les données brutes mais leur rang. Cette méthode connue sous le nom « d'Analyse des rangs » est considérée comme plus robuste que l'ACP

## La méthode : une enquête dérivée de Ter-Uti



Au deuxième degré, on tire des points du territoire à l'intérieur des photos. Cette opération consiste à reporter sur chacune des photographies aériennes, une grille de 36 points à enquêter, alignés de 6 en 6. Les points sont distants d'environ 300 mètres sur le terrain. Pour l'enquête haies le second degré de tirage porte sur les carrés délimités par les 36 points de la photo. Parmi les 25 carrés potentiels (1 carré étant délimité par 4 points adjacents) 2 carrés ont été tirés. Le premier est formé des points 1-2-7-8 (situés en haut à gauche) et le second par les points 29-30-35-36 (en bas à droite). La collecte de l'information est ainsi réalisée en Bretagne sur un échantillon de 1 592 carrés de 9 hectares soit une surface totale correspondante de 14 328 hectares (environ un 200 ème du territoire breton). Au troisième et dernier degré de tirage les bords de parcelle présents à l'intérieur des carrés tirés sont recensés de manière exhaustive. De cette manière le sondage se ramène en fait à un sondage à deux degrés. Il importe de rappeler que l'unité statistique est constituée par le bord de parcelle avec végétation ligneuse ou talus bocager et non par la haie.

L'échantillon ainsi constitué contient 16 360 bords de parcelle (les individus) représentant environ 250 000 km de linéaire. Les variables analysées sont pour la plupart des variables de base mais comportent également des variables de synthèse telles le peuplement, l'efficacité filtre, les efficacités brise-vent rapprochée ou éloignée. La seule variable continue figurant en élément actif, la longueur du bord de parcelle, a fait l'objet d'un découpage en quatre modalités. Les variables supplémentaires sont soit des caractéristiques fonctionnelles, efficacité de la futaie et de la haie, soit le type de famille, soit géographiques (département et petite région agricole). Enfin la densité en arbres ainsi que la longueur « IFN » qui est une mesure de la longueur de la haie dont la définition diffère de celle du bord de parcelle recueillie ici. L'analyse comporte 17 variables actives totalisant 66 modalités.

Les résultats : La séquence des taux d'inertie révèle l'importance des premiers axes puis leur décroissance régulière à partir du huitième. On observe également que le taux d'inertie est relativement faible sur le premier axe avec 8.36. Signalons que le scree-test de Catell conduit ici à retenir les trois premiers axes ce qui est manifestement trop peu.

EDITION DES VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 2.8824  
 SOMME DES VALEURS PROPRES .... 2.8824

HISTOGRAMME DES 39 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.2409	8.36	8.36	*****
2	.1782	6.18	14.54	*****
3	.1634	5.67	20.21	*****
4	.1322	4.59	24.80	*****
5	.1137	3.95	28.74	*****
6	.0970	3.36	32.11	*****
7	.0920	3.19	35.30	*****
8	.0791	2.75	38.05	*****
9	.0775	2.69	40.73	*****
10	.0725	2.52	43.25	*****
11	.0710	2.46	45.71	*****
12	.0694	2.41	48.12	*****
13	.0659	2.29	50.40	*****
14	.0651	2.26	52.66	*****
15	.0621	2.15	54.81	*****
16	.0619	2.15	56.96	*****
17	.0610	2.12	59.08	*****
18	.0598	2.07	61.15	*****
19	.0591	2.05	63.20	*****
20	.0587	2.04	65.24	*****
21	.0582	2.02	67.26	*****
22	.0574	1.99	69.25	*****
23	.0571	1.98	71.23	*****
24	.0561	1.95	73.17	*****
25	.0555	1.93	75.10	*****
26	.0548	1.90	77.00	*****
27	.0537	1.86	78.87	*****
28	.0531	1.84	80.71	*****
29	.0518	1.80	82.51	*****
30	.0516	1.79	84.30	*****
31	.0493	1.71	86.01	*****
32	.0490	1.70	87.70	*****
33	.0468	1.62	89.33	*****
34	.0450	1.56	90.89	*****
35	.0427	1.48	92.37	*****
36	.0395	1.37	93.74	*****
37	.0354	1.23	94.97	*****
38	.0329	1.14	96.11	*****
39	.0287	.99	97.10	*****

Le premier axe peut s'interpréter comme un gradient d'efficacité brise-vent que celle-ci soit éloignée ou rapprochée (les contributions sont respectivement de 15.6 et 15.7 %). A gauche de l'axe figurent les haies à l'efficacité brise-vent nulle ou faible (modalités ebr1 et ebe1). Il s'agit soit de haies basses, soit de talus bocagers nus ou de haies reliques, soit encore de haies ornementales (modalité esso).

## 126 CHAPITRE 5. ANALYSE DES CORRESPONDANCES MULTIPLES

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRÉS DES MODALITES ACTIVES SUR LES AXES 1 A 5

MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
8 . ORIENTATION DE LA HAIE																	
or_1 - NO-SE	1.39	3.22	.02	-.10	.02	.04	-.02	.0	.1	.0	.0	.0	.00	.00	.00	.00	.00
or_2 - N-S	1.61	2.66	-.05	.14	-.02	.03	-.01	.0	.2	.0	.0	.0	.00	.01	.00	.00	.00
or_3 - NE-SO	1.40	3.20	.10	-.09	.06	.03	.00	.1	.1	.0	.0	.0	.00	.00	.00	.00	.00
or_4 - E-O	1.48	2.98	-.06	.03	-.05	-.09	.03	.0	.0	.0	.1	.0	.00	.00	.00	.00	.00
CONTRIBUTION CUMULEE =								.1	.3	.1	.1	.0					
9 . haie particuliere																	
esso - essence ornementa	.70	7.37	-1.37	.83	1.05	-.60	.33	5.5	2.7	4.7	1.9	.7	.26	.09	.15	.05	.01
h_J - haies jeunes cha	.11	51.12	-.92	.09	-.11	-.22	-.31	.4	.0	.0	.0	.1	.02	.00	.00	.00	.00
c_p - cepees	2.57	1.29	.45	-.37	.37	.50	.36	2.2	2.0	2.2	4.9	2.9	.16	.11	.11	.20	.10
hp_0 - pas de haies part.	2.50	1.35	-.04	.14	-.67	-.34	-.45	.0	.3	6.9	2.2	4.4	.00	.01	.33	.08	.15
CONTRIBUTION CUMULEE =								8.1	5.0	13.8	9.0	8.0					
10 . Largeur dominante																	
lar1 - <3 m	1.53	2.84	-1.21	-.07	.12	.05	-.31	9.2	.0	.1	.0	1.3	.51	.00	.00	.00	.03
lar2 - 3-10 m	3.44	.71	.32	-.02	-.12	.01	.25	1.5	.0	.3	.0	1.9	.15	.00	.02	.00	.09
lar3 - >=10 m	.91	5.44	.80	.21	.25	-.13	-.45	2.5	.2	.3	.1	1.6	.12	.01	.01	.00	.04
CONTRIBUTION CUMULEE =								13.2	.3	.8	.2	4.8					
11 . nature du talus																	
deni - denivele	.95	5.19	.03	-.45	-.18	-.37	-.19	.0	1.1	.2	1.0	.3	.00	.04	.01	.03	.01
boca - bocager	3.40	.73	.33	-.23	-.14	.34	-.12	1.6	1.0	.4	3.0	.4	.15	.07	.03	.16	.02
nat0 - pas de talus	1.53	2.85	-.76	.80	.42	-.54	.38	3.6	5.5	1.6	3.3	1.9	.20	.22	.06	.10	.05
CONTRIBUTION CUMULEE =								5.2	7.6	2.2	7.3	2.7					
12 . hauteur du talus																	
ht_1 - 0-0.5 m	.80	6.33	.11	-.06	-.43	.10	-.16	.0	.0	.9	.1	.2	.00	.00	.03	.00	.00
ht_2 - 0.5-1 m	1.86	2.15	.30	-.14	-.28	.27	-.10	.7	.2	.9	1.0	.2	.04	.01	.04	.03	.00
ht_3 - 1-1.5 m	1.20	3.89	.31	-.45	.13	.33	-.16	.5	1.4	.1	1.0	.3	.02	.05	.00	.03	.01
ht_4 - >=1.5 m	.48	11.15	.29	-.77	.13	-.32	-.14	.2	1.6	.1	.4	.1	.01	.05	.00	.01	.00
ht_0 - pas de talus	1.53	2.85	-.76	.80	.42	-.54	.38	3.6	5.5	1.6	3.3	1.9	.20	.22	.06	.10	.05
CONTRIBUTION CUMULEE =								5.0	8.6	3.6	5.8	2.7					
13 . pente amont																	
faib - faible - moderee	2.84	1.07	.04	-.63	.12	-.40	.02	.0	6.3	.3	3.4	.0	.00	.37	.01	.15	.00
fort - forte	.32	17.55	.43	-.64	-.05	-.79	.00	.2	.7	.0	1.5	.0	.01	.02	.00	.04	.00
pt_0 - pas de pente	2.72	1.16	-.09	.73	-.12	.51	-.02	.1	8.2	.2	5.3	.0	.01	.46	.01	.22	.00
CONTRIBUTION CUMULEE =								.4	15.2	.5	10.2	.0					
14 . situation de la haies sur la pente																	
perp - perpendiculaire	.89	5.64	.22	-1.08	.45	-1.03	.15	.2	5.8	1.1	7.2	.2	.01	.21	.04	.19	.00
parl - parallele	1.78	2.31	-.03	-.28	-.12	.04	-.05	.0	.8	.2	.0	.0	.00	.03	.01	.00	.00
biai - en biais	.49	10.89	.22	-1.08	.27	-1.08	.05	.1	3.2	.2	4.4	.0	.00	.11	.01	.11	.00
ph_0 - pas de pente	2.72	1.16	-.09	.73	-.12	.51	-.02	.1	8.2	.2	5.3	.0	.01	.46	.01	.22	.00
CONTRIBUTION CUMULEE =								.4	18.0	1.7	16.8	.2					
15 . entretien des haies																	
reco - recolte	.18	32.35	-.16	-.38	-1.05	-.07	.08	.0	.1	1.2	.0	.0	.00	.00	.03	.00	.00
perd - bois perdu	.44	12.32	-1.36	.60	1.03	-.51	.36	3.4	.9	2.9	.9	.5	.15	.03	.09	.02	.01
ent0 - pas d'entretien	5.26	.12	.12	-.04	-.05	.05	-.03	.3	.0	.1	.1	.0	.12	.01	.02	.02	.01
CONTRIBUTION CUMULEE =								3.7	1.1	4.2	1.0	.6					
19 . peuplement																	
ppl1 - ni taillis ni futaie	53	2.86	-1.24	-.43	.42	.31	-.45	9.8	1.6	1.7	1.1	2.7	.54	.06	.06	.03	.07
ppl2 - haies ajourees	1.04	4.66	-.12	-.40	-.82	.18	1.06	.1	.9	4.3	.3	10.3	.00	.03	.14	.01	.24
ppl3 - taillis tres dense	.47	11.59	.76	-.54	1.59	.84	.10	1.1	.8	7.2	2.5	.0	.05	.03	.22	.06	.00
ppl4 - tsf domi. taillis	.28	20.22	1.24	-.41	.85	.40	.04	1.8	.3	1.2	.3	.0	.08	.01	.04	.01	.00
ppl5 - futaie moy. dense	.47	11.48	-.06	.18	-1.09	-.33	.57	.0	.1	3.4	.4	1.4	.00	.00	.10	.01	.03
ppl6 - futaie tres dense	1.42	3.14	.40	.87	-.28	-.87	-.48	1.0	6.0	.7	8.2	2.9	.05	.24	.03	.24	.07
ppl7 - tsf domi. futaie	.68	7.63	1.14	.17	.21	.35	-.11	3.7	.1	.2	.6	.1	.17	.00	.01	.02	.00
CONTRIBUTION CUMULEE =								17.3	9.7	18.7	13.4	17.3					
21 . efficience filtre																	
fil1 - oui	1.38	3.26	.22	-1.08	.39	-1.05	.11	.3	9.0	1.3	11.5	.2	.02	.36	.05	.34	.00
fil2 - non	4.50	.31	-.07	.33	-.12	.32	-.03	.1	2.8	.4	3.5	.0	.02	.36	.05	.34	.00
CONTRIBUTION CUMULEE =								.4	11.8	1.7	15.0	.2					
22 . efficacite cloture																	
ec11 - oui	1.79	2.29	.50	.04	.52	.24	-.11	1.8	.0	3.0	.8	.2	.11	.00	.12	.03	.01
ec12 - non	4.09	.44	-.22	-.02	-.23	-.11	.05	.8	.0	1.3	.3	.1	.11	.00	.12	.03	.01
CONTRIBUTION CUMULEE =								2.6	.0	4.3	1.1	.3					
23 . Efficacite brise-vent rapproche																	
ebr1 - nulle	.95	5.18	-.93	-.66	-1.10	.19	-.76	3.4	2.3	7.1	.2	4.8	.17	.08	.23	.01	.11
ebr2 - faible	1.63	2.61	-.86	-.07	.32	.05	.24	5.0	.0	1.0	.0	.9	.28	.00	.04	.00	.02
ebr3 - moderee	1.14	4.16	.38	.10	-.55	-.03	1.21	.7	.1	2.1	.0	14.8	.04	.00	.07	.00	.35
ebr4 - forte	2.16	1.72	.86	.29	.53	-.11	-.49	6.6	1.0	3.8	.2	4.6	.42	.05	.17	.01	.14
CONTRIBUTION CUMULEE =								15.7	3.5	14.0	.5	25.1					
24 . Efficacite brise-vent eloigne																	
ebe1 - nulle	1.95	2.02	-1.06	-.49	-.01	.28	-.36	9.0	2.7	.0	1.2	2.2	.55	.12	.00	.04	.06
ebe2 - faible	.95	5.17	-.10	.03	-.66	-.13	1.48	.0	.0	2.5	.1	18.5	.00	.00	.08	.00	.43
ebe3 - moderee	1.45	3.05	.64	-.11	.56	.33	.31	2.5	.1	2.8	1.2	1.2	.14	.00	.10	.04	.03
ebe4 - forte	1.53	2.86	.80	.72	-.10	-.60	-.76	4.1	4.4	.1	4.1	7.8	.22	.18	.00	.12	.20
CONTRIBUTION CUMULEE =								15.6	7.2	5.4	6.6	29.7					



COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES MODALITES ACTIVES SUR LES AXES 1 A 5 (suite)																		
MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES					
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	
29 . dominante en arbre																		
dom1 - tetards	1.37	3.28	.39	.13	-.59	-.02	.16	.9	.1	2.9	.0	.3	.05	.01	.11	.00	.01	
dom2 - futaie	1.02	4.78	.44	.18	-.49	-.20	-.05	.8	.2	1.5	.3	.0	.04	.01	.05	.01	.00	
dom3 - tetards+futaie	.65	8.05	.89	.13	-.11	-.10	-.25	2.1	.1	.0	.1	.3	.10	.00	.00	.00	.01	
dom4 - resineux	.37	14.86	-.27	1.38	.50	-1.48	.26	.1	4.0	.6	6.1	.2	.00	.13	.02	.15	.00	
dom5 - autres dominantes	.26	21.69	.19	.21	-.51	-.42	.17	.0	.1	.4	.3	.1	.00	.00	.01	.01	.00	
dom6 - pas de dominantes	2.21	1.66	-.68	-.46	.60	.43	-.07	4.3	2.6	4.9	3.1	.1	.28	.13	.22	.11	.00	
CONTRIBUTION CUMULEE =								8.2	7.0	10.3	10.0	1.1						
31 . strate basse																		
stb1 - nulle	.47	11.41	-.79	-.30	-1.06	-.07	-.70	1.2	.2	3.2	.0	2.0	.05	.01	.10	.00	.04	
stb2 - faible <1/3	.96	5.12	-.50	-.34	-1.01	-.02	-.34	1.0	.6	6.0	.0	1.0	.05	.02	.20	.00	.02	
stb3 - moyenne 1/3-2/3	1.21	3.86	.08	-.03	-.38	-.04	.60	.0	.0	1.1	.0	3.9	.00	.00	.04	.00	.09	
stb4 - forte >2/3	3.24	.82	.23	.16	.60	.03	-.02	.7	.5	7.1	.0	.0	.07	.03	.44	.00	.00	
CONTRIBUTION CUMULEE =								2.9	1.3	17.4	.1	6.9						
32 . fosse utile																		
amon - fosse en amont	.17	32.84	.17	-.79	-.06	-.69	.07	.0	.6	.0	.6	.0	.00	.02	.00	.01	.00	
aval - fosse en aval	.34	16.11	.02	-.72	-.02	-.72	-.10	.0	1.0	.0	1.4	.0	.00	.03	.00	.03	.00	
inde - indetermine	.94	5.26	-.11	.01	-.25	.28	-.20	.0	.0	.4	.6	.3	.00	.00	.01	.02	.01	
fos0 - pas de fosse	4.42	.33	.02	.08	.06	.02	.05	.0	.2	.1	.0	.1	.00	.02	.01	.00	.01	
CONTRIBUTION CUMULEE =								.1	1.8	.5	2.6	.4						

A l'inverse, à droite de l'axe se profilent des haies « plus denses », caractérisées par des efficacités brise-vent éloignée ou rapprochée fortes (ebr4 et ebe4). Le deuxième axe oppose les bords de parcelle selon la présence ou non d'un talus, l'existence ou non d'une pente et peut donc s'interpréter comme un gradient de l'effet filtre. Un bord de parcelle possède un effet filtre s'il se trouve placé en biais ou perpendiculairement sur un terrain en pente. Le nuage des individus, ici les bords parcelle, représenté sur le premier plan, souligne sans doute mieux qu'un long discours, l'utilité de la classification pour segmenter les bords de parcelle en groupes homogènes.



IDENTIFICATION DES POINTS

\* : UN SEUL POINT

N : N POINTS SUPERPOSES

X : 10 POINTS SUPERPOSES OU PLUS



## 5.10 Compléments

### 5.10.1 Valeur-test

Le seul indicateur utilisable en ACP et en AFC pour apprécier le rôle joué par une variable supplémentaire est la qualité de représentation, c'est à dire le carré du cosinus de l'angle formé entre l'élément dans l'espace d'origine et sa projection sur le sous-espace considéré. Il n'existe pas de seuil objectif permettant de déterminer le caractère significatif ou non de cet indicateur. En revanche en ACM, l'expression de la coordonnée d'une modalité sur un axe permet de disposer d'un tel indicateur. Rappelons en effet que la coordonnée  $d_{jk}$  de la modalité  $j$  sur l'axe de rang  $k$  peut s'écrire :

$$d_{jk} = \frac{1}{\sqrt{\lambda_k}} \frac{1}{n_j} \sum_{i \in I_j} c_{ik}$$

où  $I_j$  désigne l'ensemble des individus possédant la modalité  $j$ .

Cette égalité peut encore s'écrire :

$$\sqrt{\lambda_k} d_{jk} = \frac{1}{n_j} \sum_{i \in I_j} c_{ik}$$

Le second membre de cette égalité correspond à la moyenne des coordonnées  $c_{ik}$  des  $n_j$  individus possédant la modalité  $j$ . D'où l'idée d'un test afin de déterminer si la position de cette modalité sur l'axe de rang  $k$  est « significative » ou non. Le test consiste en l'occurrence à examiner si la valeur observée peut être considérée comme la réalisation de la moyenne dans un tirage de  $n_j$  coordonnées parmi  $n$ . Si oui, alors la position de la modalité considérée n'est pas significative sur l'axe de rang  $k$  et inversement dans le cas contraire. D'après la théorie des sondages, on sait que dans un tirage à probabilités égales et sans remise (PESR) de  $n$  unités parmi  $N$ , la moyenne d'échantillon est un estimateur sans biais de la moyenne de la population, soit :

$$E(\hat{\bar{Y}}) = E(\bar{y}) = \bar{Y}$$

et que :

$$V(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right) \frac{S_Y^2}{n}$$

où :

$$S_Y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

L'expression de la variance de l'estimateur peut encore s'écrire :

$$V(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right) \frac{N}{N-1} \frac{V(Y)}{n}$$

soit :

$$V(\hat{\bar{Y}}) = \frac{N-n}{N-1} \frac{V(Y)}{n}$$

Par ailleurs, lorsque la taille de l'échantillon devient « grande », alors  $\bar{y}$  suit une loi normale  $N\left(\bar{Y}, \frac{N-n}{N-1} \frac{V(Y)}{n}\right)$ .

Ici,  $Y = C_k$ ,  $n = n_j$  et  $N = n$ . Par ailleurs  $\bar{C}_k = 0$  et  $V(C_k) = I_k = \lambda_k$ , d'où :

$$V(\bar{y}) = \frac{n - n_j}{n_j (n - 1)} \lambda_k$$

d'où la statistique du test :

$$\frac{\sqrt{\lambda_k} d_{jk} - \bar{C}_k}{\sqrt{\frac{n - n_j}{n_j (n - 1)} \lambda_k}} = \frac{d_{jk}}{\sqrt{\frac{n - n_j}{n_j (n - 1)}}} = d_{jk} \sqrt{\frac{n_j (n - 1)}{n - n_j}}$$

La quantité  $VT = d_{jk} \sqrt{\frac{n_j (n - 1)}{n - n_j}}$  est appelée « valeur-test ». Cette quantité

est comparée au fractile de la loi Normale centrée-réduite d'ordre  $\alpha$ .

Si  $|VT| > t_{1-\frac{\alpha}{2}}$  on rejette l'hypothèse selon laquelle la valeur observée  $\sqrt{\lambda_k} d_{jk}$  peut être considérée comme la réalisation de la moyenne dans un tirage PESR de  $n_j$  observations parmi  $n$ , auquel cas la position de la modalité  $j$  sur l'axe de rang  $k$  est jugée significative (en fait significativement différente de la moyenne). Au seuil usuel,  $\alpha = 5\%$ , la valeur critique est 1.96, d'où l'habitude de ne sélectionner que les modalités supplémentaires pour lesquelles la valeur-test est supérieure à 2 en valeur absolue.

Il convient de souligner qu'en toute rigueur, la notion de valeur-test ne peut être utilisée dans une optique décisionnelle que pour les seules modalités illustratives ou supplémentaires et ce, même si certains logiciels les éditent

pour l'ensemble des variables, y compris actives. Rappelons que pour les modalités actives, la quantité à examiner est la contribution (à l'inertie de l'axe). Enfin, cette utilisation suppose que l'effectif dans la modalité ne soit pas trop faible.

### 5.10.2 Taux d'inertie en ACM

A l'expérience, les taux observés en ACM sont faibles, voire très faibles, y compris sur les premiers axes. Cet état caractéristique de l'ACM a amené certains auteurs à proposer des taux jugés plus « réalistes », c'est à dire rendant mieux compte de la part d'information restituée.

Rappelons que les résultats d'une ACM peuvent être indifféremment acquis par l'AFC du tableau  $Z$  ou par celle du tableau  $B$ . Les valeurs propres issues des deux analyses vérifiant l'égalité suivante :

$$\lambda_k(B) = \lambda_k^2(Z)$$

Le tableau de BURT  $B$  peut, ainsi que nous l'avons vu antérieurement, se décomposer en blocs diagonaux  $B_{qq}$  et en blocs non-diagonaux  $B_{qq'}$ . Par conséquent, l'inertie totale  $I_B$  de  $B$  peut s'écrire comme somme de l'inertie des blocs diagonaux  $I_D$  et de l'inertie des blocs non-diagonaux  $I_E$ , soit :

$$I_B = I_D + I_E$$

Or, les blocs diagonaux ne présentent aucun intérêt puisque chacun d'entre eux croise une variable avec elle-même. Toutefois, c'est leur inertie qui va être maximale dans l'inertie totale. En effet, l'inertie étant égale au lien entre les deux caractères, il apparaît bien que l'information apportée par un caractère sur l'autre est maximale pour les blocs diagonaux. Elle est même parfaite dans la mesure où les deux caractères sont identiques. Les blocs diagonaux sont donc à inertie maximale mais à information nulle. L'inertie « utile » réside dans les blocs non-diagonaux qui constituent les tableaux de contingence croisant deux variables distinctes. D'où l'idée de certains auteurs de ne mesurer l'inertie conservée par les axes que par rapport à celle des seuls blocs non-diagonaux.

Comme  $I_B = I_D + I_E$  il est immédiat que  $I_E = I_B - I_D$ . D'autre part on peut montrer que :

$$I_D = \frac{J - Q}{Q^2}$$

Comme  $\lambda_k(B) = \lambda_k^2(Z)$ , alors :

$$I_B = \sum_{k=1}^{J-Q} \lambda_k(B) = \sum_{k=1}^{J-Q} \lambda_k^2(Z)$$

d'où :

$$I_E = \sum_{k=1}^{J-Q} \lambda_k^2(Z) - (J-Q) \frac{1}{Q^2}$$

Cette dernière expression peut encore s'écrire :

$$I_E = \sum_{k=1}^{J-Q} \left[ \lambda_k(Z) - \frac{1}{Q} \right]^2$$

En effet :

$$\begin{aligned} \sum_{k=1}^{J-Q} \left[ \lambda_k(Z) - \frac{1}{Q} \right]^2 &= \sum_{k=1}^{J-Q} \left[ \lambda_k^2(Z) + \frac{1}{Q^2} - 2\frac{1}{Q}\lambda_k^2(Z) \right] \\ &= \sum_{k=1}^{J-Q} \lambda_k^2(Z) + \frac{J-Q}{Q^2} - \frac{2}{Q} \sum_{k=1}^{J-Q} \lambda_k(Z) \end{aligned}$$

Comme

$$\sum_{k=1}^{J-Q} \lambda_k(Z) = \frac{J}{Q} - 1 = \frac{J-Q}{Q}$$

alors

$$\begin{aligned} \sum_{k=1}^{J-Q} \left[ \lambda_k(Z) - \frac{1}{Q} \right]^2 &= \sum_{k=1}^{J-Q} \lambda_k^2(Z) + \frac{J-Q}{Q^2} - \frac{2}{Q} \frac{J-Q}{Q} \\ &= \sum_{k=1}^{J-Q} \lambda_k^2(Z) - \frac{J-Q}{Q^2} \end{aligned}$$

d'où le résultat, donc :

$$I_E = \sum_{k=1}^{J-Q} \left[ \lambda_k(Z) - \frac{1}{Q} \right]^2$$

Les taux d'inertie modifiés proposés par GREENACRE consistent donc à ne considérer que la seule inertie portée par les blocs non-diagonaux, soit :

$$\tau_k^{(G)} = \frac{\left[ \lambda_k(Z) - \frac{1}{Q} \right]^2}{\sum_{k=1}^{J-Q} \left[ \lambda_k(Z) - \frac{1}{Q} \right]^2}$$

Ces taux étant calculés pour  $\lambda_k(Z) > \frac{1}{Q}$

On peut par ailleurs remarquer que :

$$\bar{\lambda}(Z) = \frac{1}{J-Q} \sum_{k=1}^{J-Q} \lambda_k(Z) = \frac{1}{J-Q} \frac{J-Q}{Q} = \frac{1}{Q}$$

Dans ces conditions, les taux d'inertie modifiés peuvent encore s'écrire :

$$\tau_k^{(G)} = \frac{[\lambda_k(Z) - \bar{\lambda}(Z)]^2}{\sum_{k=1}^{J-Q} [\lambda_k(Z) - \bar{\lambda}(Z)]^2}$$

Signalons que BENZECRI avait proposé les taux modifiés suivants, qui vont être « optimistes » dans la mesure où l'on ne considère pas la totalité des axes mais seuls ceux dont l'inertie  $\lambda_k$  est supérieure à la valeur moyenne  $\frac{1}{Q}$  :

$$\tau_k^{(B)} = \frac{\left[\lambda_k(Z) - \frac{1}{Q}\right]^2}{\sum_{k > \frac{1}{Q}} \left[\lambda_k(Z) - \frac{1}{Q}\right]^2}$$



L'application au second exemple présenté concernant les haies de Bretagne montre bien l'intérêt de ces taux modifiés. Le premier axe ne restitue en effet que 8.36 % de l'inertie totale alors qu'il s'interprète comme un gradient d'efficacité brise-vent et correspond à l'équivalent d'un effet taille en ACP. Le taux modifié fournit un pourcentage de 33.06 % se rapprochant ainsi de ce que l'on observe en ACP pour un même phénomène.

L'inertie totale est égale à 0.2698 ( $I_B = \sum_{k=1}^{J-Q} \lambda_k^2(Z)$ ) et l'on observe que l'inertie des blocs non-diagonaux est égale à 0.1003 soit seulement 37.2 % de l'inertie totale. L'essentiel de l'inertie provient bien des blocs diagonaux, qui sont à inertie maximale mais à information nulle.

Enquête HAIES 1996 - Taux d'inertie modifiés (GREENACRE)  
(Q=17 , J=66)

Axe $k$	$I_T(k)$	$I_E(k)$	$I_E(k)/I_E$	Taux modifiés	Taux initiaux	$\lambda_k(Z) > 1/Q$
1	0,2409	0,0332	0,3306	33,06	8,36	VRAI
2	0,1782	0,0143	0,1421	14,21	6,18	VRAI
3	0,1634	0,0109	0,1090	10,90	5,67	VRAI
4	0,1322	0,0054	0,0537	5,37	4,59	VRAI
5	0,1137	0,0030	0,0300	3,00	3,94	VRAI
6	0,0970	0,0015	0,0145	1,45	3,37	VRAI
7	0,0920	0,0011	0,0110	1,10	3,19	VRAI
8	0,0791	0,0004	0,0041	0,41	2,74	VRAI
9	0,0775	0,0003	0,0035	0,35	2,69	VRAI
10	0,0725	0,0002	0,0019	0,19	2,52	VRAI
11	0,0710	0,0001	0,0015	0,15	2,46	VRAI
12	0,0694	0,0001	0,0011	0,11	2,41	VRAI
13	0,0659	0,0001	0,0005	0,05	2,29	VRAI
14	0,0651	0,0000	0,0004	0,04	2,26	VRAI
15	0,0621	0,0000	0,0001	0,01	2,15	VRAI
16	0,0619	0,0000	0,0001	0,01	2,15	VRAI
17	0,0610	0,0000	0,0000	0,00	2,12	VRAI
18	0,0598	0,0000	0,0000	0,00	2,07	VRAI
19	0,0591	0,0000	0,0000	0,00	2,05	VRAI
20	0,0587	0,0000	0,0000	0,00	2,04	FAUX
...						
Total	2,8824	0,1003		100,00	100,00	



# Chapitre 6

## Méthodes de Classification

### 6.1 Introduction

Les méthodes de classification constituent un champ d'étude très vaste ainsi qu'en témoigne la nombreuse littérature existant sur le sujet. L'objectif visé ici est double. D'une part présenter les méthodes les plus usuelles et non certaines méthodes très originales dues à l'imagination fertile des chercheurs mais peu usitées en pratique. En second lieu l'appréhension des méthodes de classification est ici essentiellement conçu dans le prolongement et en complémentarité avec les techniques factorielles.

L'objectif des méthodes de classification, quelles que puissent être par ailleurs leurs différences, demeure toujours de découper une population en groupes homogènes.

Après avoir rappelé dans une première partie quelques notions élémentaires indispensables, nous aborderons le problème de la construction d'une suite de partitions à partir d'un exemple. La présentation du déroulement d'un algorithme de classification ascendante hiérarchique effectué, le problème fondamental du choix de la distance entre parties (stratégie d'agrégation) sera exposé. Ce qui nous amènera à présenter les stratégies usuelles et notamment la plus fréquente, en l'occurrence la méthode de Ward.

Une fois la construction d'une classification ascendante présentée, l'intérêt se portera sur l'utilisation complémentaire de la classification et des méthodes factorielles sur le plan pratique.

Enfin ce chapitre se termine par la présentation des techniques de partitionnement, centres mobiles et nuées dynamiques, plus frustes que les méthodes hiérarchiques mais plus simples et plus rapides et qui sont encore parfois utilisées.

## 6.2 Définitions

### 6.2.1 Distance

Une distance  $d$  définie sur un ensemble  $E$  est une application de  $E \times E$  dans  $\mathbf{R}^+$  vérifiant les propriétés suivantes :

$$(1) \forall x \in E, \quad d(x, x) = 0$$

$$(2) \forall x \in E, \forall y \in E, \quad d(x, y) = d(y, x)$$

$$(3) \forall x \in E, \forall y \in E, \quad d(x, y) = 0 \Rightarrow x = y$$

$$(4) \forall x \in E, \forall y \in E, \forall z \in E, \quad d(x, y) \leq d(x, z) + d(y, z)$$

Lorsque la propriété (4) n'est pas vérifiée l'application  $d$  est appelée indice de distance.

### 6.2.2 Distances usuelles

Soient  $x$  et  $y$  deux vecteurs de  $\mathbf{R}^p$ ,  $x = (x_j)$  et  $y = (y_j)$ , la distance euclidienne usuelle entre  $x$  et  $y$  s'écrit :

$$d^2(x, y) = \sum_{j=1}^p (x_j - y_j)^2$$

La distance inverse des variances s'écrit :

$$d^2(x, y) = \sum_{j=1}^p \frac{1}{\sigma_j^2} (x_j - y_j)^2$$

Il s'agit de la distance utilisée en ACP normée.

En AFC la distance utilisée entre les profils  $i_1$  et  $i_2$  s'écrit :

$$d^2(i_1, i_2) = \sum_{j=1}^p \frac{1}{f_{\cdot j}^2} \left( \frac{f_{i_1 j}}{f_{\cdot j}} - \frac{f_{i_2 j}}{f_{\cdot j}} \right)^2$$

La distance entre deux éléments peut également se calculer sur leurs coordonnées factorielles :

$$d^2(i_1, i_1) = \sum_{j=1}^p (c_{i_1j} - c_{i_2j})^2$$

où  $c_{ij}$  désigne la coordonnée de l'élément  $i$  sur l'axe factoriel de rang  $j$ .

Dans le calcul de cette distance chaque coordonnée est implicitement pondérée par le « poids » de chaque facteur (la variance d'un facteur n'est autre que l'inertie de l'axe correspondant égale à la valeur propre associée).

Il est également possible, ce qui sera le cas en pratique, de limiter le calcul des distances aux premiers facteurs, soit :

$$d^2(i_1, i_1) = \sum_{j=1}^{q < p} (c_{i_1j} - c_{i_2j})^2$$

Il convient d'observer qu'ici, il ne faut surtout pas « normer » les données, dans la mesure où la variance d'un facteur n'est autre que l'inertie correspondante. Dans ces conditions, normer les données reviendrait à accorder à chaque facteur, quel que soit son rang, un « poids » identique dans le calcul des distances, ce qui risquerait de conduire à des résultats peu probants.

Ces distances constituent toutes un cas particulier de la distance suivante, appelée distance de Minkowski :

$$d(x, y) = \left( \sum_{j=1}^p |x_j - y_j|^k \right)^{\frac{1}{k}}$$

Les distances précédentes correspondant au cas où  $k = 2$ . Ces distances sont dites euclidiennes car liées à un produit scalaire.

Remarque :

Dans la définition précédente lorsque  $k = 1$  on obtient la distance suivante appelée « city-block » :

$$d(x, y) = \sum_{j=1}^p |x_j - y_j|$$

### 6.2.3 Distance ultramétrique

On appelle distance ultramétrique toute distance possédant la propriété suivante, plus forte que l'inégalité triangulaire :

$$(5) \quad \forall x \in E, \forall y \in E, \forall z \in E, \quad d(x, y) \leq \max(d(x, z), d(y, z))$$

Toute distance ultramétrique vérifie la propriété (4).

#### 6.2.4 Hiérarchie

Soit  $E$  un ensemble fini. On appelle partition de  $E$  toute famille de sous-ensembles  $E_i$  de  $E$  vérifiant les relations :

$$\bigcup_i E_i = E$$

et

$$E_i \cap E_j = \emptyset \quad \text{si} \quad i \neq j$$

$H$  est une hiérarchie de parties de  $E$  si  $H$  est un sous-ensemble de parties de  $E$ , tel que :

$$\forall h_1 \in H, \forall h_2 \in H, \quad h_1 \cap h_2 \in (h_1, h_2, \emptyset)$$

En d'autres termes on a soit  $h_1 \subset h_2$ , soit  $h_2 \subset h_1$ , soit  $h_1 \cap h_2 = \emptyset$ .

Deux éléments d'une partition sont, soit contenus l'un dans l'autre, soit leur intersection est vide.

Exemple : soit  $E = \{a, b, c, d, e\}$ , examinons si  $P(E)$  l'ensemble des parties de  $E$ , est une hiérarchie de parties de  $E$ . Il est immédiat que  $\{a, b\}$  et  $\{a, c\}$  sont des éléments de  $P(E)$  et la propriété n'est pas vérifiée, en effet  $\{a, b\} \cap \{a, c\} = \{a\} \neq \emptyset$  et aucun des deux ensembles n'est inclus dans l'autre.

Par contre, l'ensemble suivant  $H = \{\{a\}, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$  est bien une hiérarchie de parties de  $E$ .

Une hiérarchie est dite fine si chacun des éléments de base, les singletons, appartient à la hiérarchie.

Elle est dite totale si elle est fine et si  $E$  lui-même appartient à la hiérarchie. Dans la suite, seules les hiérarchies totales seront considérées.

Une hiérarchie est un arbre, c'est à dire satisfait à la propriété suivante :

$$\forall h, h_1, h_2 \in H, \quad h \in h_1 \quad \text{et} \quad h \in h_2 \quad \Rightarrow \quad h_1 \subset h_2 \quad \text{ou} \quad h_2 \subset h_1$$

Démonstration :

Soient  $h, h_1$  et  $h_2 \in H$  avec  $h \in h_1$  et  $h \in h_2$ . Alors  $h \in h_1 \cap h_2$ . Comme  $H$  est une hiérarchie,  $h_1 \cap h_2 \in \{h_1, h_2, \emptyset\}$  et  $h_1 \cap h_2 = h \neq \emptyset$  alors on a soit  $h_1 \subset h_2$  soit  $h_2 \subset h_1$ .

En conséquence on peut représenter toute hiérarchie par un arbre.

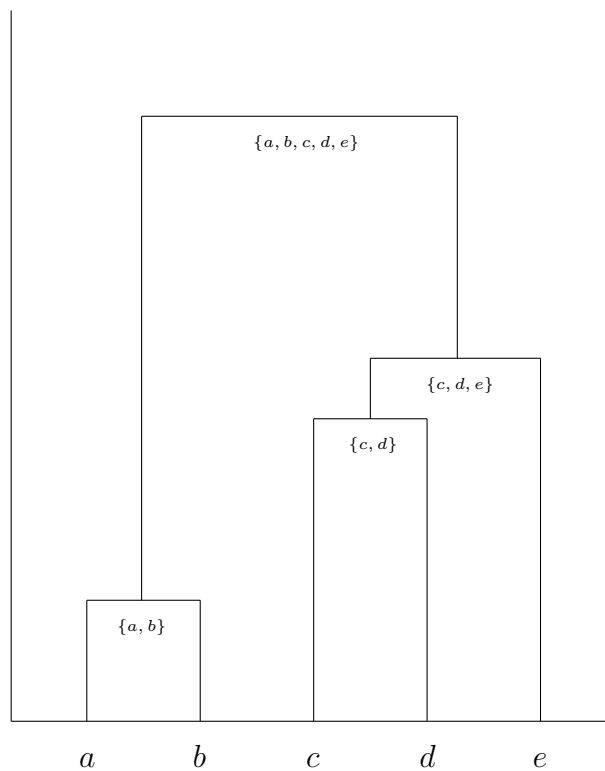
Exemple :

Soit  $E$  un ensemble et  $H_E$  une hiérarchie de  $E$  avec :

$$E = \{a, b, c, d, e\}$$

$$H_E = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{c, d\}, \{c, d, e\}, \{a, b, c, d, e\}\}$$

Alors  $H_E$  peut être représentée graphiquement sous la forme suivante :



Tout élément de  $H$  est représenté par un noeud et les noeuds sont reliés par des branches.

### 6.2.5 Définition d'une partition compatible avec une hiérarchie

Une partition de  $E$  est dite compatible avec la hiérarchie  $H$  si ses éléments appartiennent à  $H$ .

Exemples de partitions compatibles avec  $H$  :

$$P_1 = \{\{a, b\}, \{c, d, e\}\}$$

$$P_2 = \{\{a, b\}, \{c, d\}, \{e\}\}$$

Ici  $P_2$  est comprise dans  $P_1$ . Elle est dite « plus fine » que  $P_1$  ou « emboîtée » dans  $P_1$ .

### 6.2.6 Hiérarchie indicée

Une hiérarchie indicée est une hiérarchie munie d'une application  $d^*$  définie de  $H$  dans  $\mathbf{R}^+$  vérifiant :

$$\forall h_1, h_2 \in H, \quad h_1 \subset h_2 \Rightarrow d^*(h_1) \leq d^*(h_2)$$

$d^*(h)$  s'appelle l'indice de  $h$  ou le niveau d'agrégation ou encore le diamètre de la partie  $h$ .

Il s'agit en effet du niveau auquel la partie ou classe  $h$  est constituée, c'est à dire le niveau auquel sont réunis pour la première fois les éléments composant la classe  $h$ . Une hiérarchie indicée peut être représentée par un arbre indicé.

Remarque :

$d^*$  peut être définie de  $H$  dans  $\mathbf{R}^+$  ou dans une partie de  $\mathbf{R}^+$ , par exemple l'intervalle  $[0, 1]$ .

Par ailleurs, pour tous les éléments terminaux  $i$ ,  $d^*(i) = 0$ .

Enfin, la condition suivante est parfois imposée  $d^*(E) = 1$ .

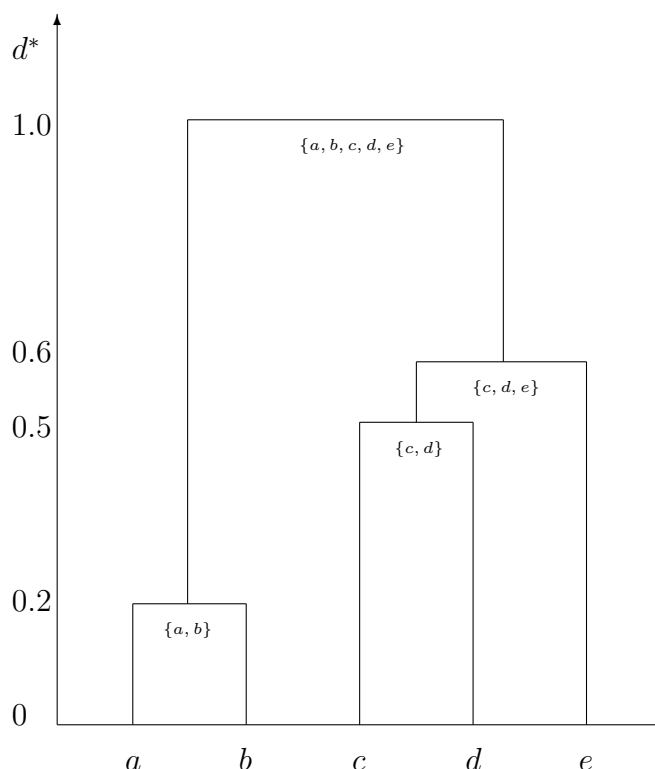
Exemple :

Considérons la hiérarchie indicée suivante :

$$H = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{c, d\}, \{c, d, e\}, \{a, b, c, d, e\}\}$$

avec  $d(\{a, b\}) = 0.2$ ,  $d(\{c, d\}) = 0.5$ ,  $d(\{c, d, e\}) = 0.6$  et  $d(E) = 1$ .





### 6.2.7 Propriété importante

Il est équivalent de se donner sur un ensemble  $E$  une distance ultramétrique ou une hiérarchie indicée. En d'autres termes, disposant d'une hiérarchie indicée sur un ensemble, il est possible d'en déduire une distance ultramétrique et réciproquement, à partir d'une distance ultramétrique entre éléments terminaux il est possible de construire une hiérarchie indicée.

## 6.3 Classification ascendante hiérarchique

### 6.3.1 Principe

Les méthodes de classification ascendante hiérarchique (CAH) consistent à construire sur un ensemble  $E$  muni d'une distance  $d$ , une hiérarchie indicée. Elles constituent en quelque sorte la traduction algorithmique de l'adage selon lequel « Qui se ressemble s'assemble ».

Le principe est le suivant :

A l'étape initiale, les  $n$  éléments de base de  $E$  (singletons), appelés éléments

terminaux, constituent des classes à eux seuls. Dans un premier temps l'algorithme calcule l'ensemble des distances entre les éléments terminaux. Les deux éléments les plus proches, au sens de cette distance, sont alors réunis en une classe. La « distance » entre cette classe ou ce nouvel élément et les  $n - 2$  éléments terminaux est ensuite calculée. Il est ainsi nécessaire de définir une distance entre un élément terminal et une classe puis plus tard entre deux classes. Le processus est ensuite réitéré, les deux éléments les plus proches parmi les  $n - 1$ , soit  $n - 2$  éléments terminaux et une classe, sont réunis. A la fin du déroulement de l'algorithme tous les éléments figurent dans une seule classe.

Les méthodes, très nombreuses en CAH, se différencient par le choix de la distance entre classes que l'on appelle une stratégie d'agrégation. Toutefois nous verrons qu'en pratique lorsque la CAH est utilisée en complément des méthodes factorielles, une distance entre parties s'impose de manière naturelle.

### 6.3.2 Exemple

Considérons l'ensemble suivant  $E$  à 5 éléments muni de la distance  $d$  :

$d$	$a$	$b$	$c$	$d$	$e$
$a$	0	0.2	1	0.7	1
$b$		0	1.05	0.75	0.8
$c$			0	0.3	1.5
$d$				0	1.3
$e$					0

La distance  $d^*$  entre classes que nous allons choisir est la suivante :

$$d^*(h_1, h_2) = \inf\{d(a_i, b_j)\} \text{ avec } a_i \in h_1 \text{ et } b_j \in h_2$$

La distance entre deux parties  $h_1$  et  $h_2$  est définie comme la plus petite distance entre deux éléments de chacun des deux ensembles.

On remarque que si  $i$  et  $j$  sont deux éléments terminaux alors on a bien  $d^*({i}, {j}) = d(i, j)$ .

Etape 1 :

Les deux éléments les plus proches sont  $a$  et  $b$  car  $d(a, b) = d^*({a}, {b}) = 0.2$ . Les éléments  $a$  et  $b$  vont donc être réunis en une classe que l'on peut noter  $\{a, b\}$  ou 6.

Etape	Classe	Eléments réunis	Niveau d'agrégation
1	6	a - b	0.2

Il nous faut maintenant calculer la distance entre cette classe composée des deux éléments  $a$  et  $b$  et les trois éléments terminaux  $c$ ,  $d$  et  $e$ . Remarquons que les distances entre ces trois éléments ne sont pas modifiées.

$$d^*({a, b}, c) = \inf\{d(a, c), d(b, c)\} = \inf\{1, 1.05\} = 1$$

$$d^*({a, b}, d) = \inf\{d(a, d), d(b, d)\} = 0.7$$

$$d^*({a, b}, e) = \inf\{d(a, e), d(b, e)\} = 0.8$$

Le tableau des distances est alors mis à jour :

$d^*$	$\{a, b\}$	$c$	$d$	$e$
$\{a, b\}$	0	<i>1.0</i>	<i>0.7</i>	<i>0.8</i>
$c$		0	0.3	1.5
$d$			0	1.3
$e$				0

Les distances calculées apparaissent en *italique*.

Etape 2 :

Les deux éléments les plus proches sont maintenant  $c$  et  $d$  ( $d^*(c, d) = 0.3$ ).  $c$  et  $d$  vont donc être agrégés pour former la seconde classe numérotée 7, le niveau d'agrégation étant de 0.3.

Le tableau des distances est mis à jour :

$d^*$	$\{a, b\}$	$\{c, d\}$	$e$
$\{a, b\}$	0	<i>0.7</i>	0.8
$\{c, d\}$		0	<i>1.3</i>
$e$			0

Calculons la distance entre  $\{c, d\}$  et les autres éléments soit  $\{a, b\}$  et  $e$ .

Si l'on revient à la définition alors :

$$d^*({a, b}, {c, d}) = \inf\{d(a, c), d(a, d), d(b, c), d(b, d)\} = 0.7$$

Mais cette distance peut être calculée à partir des distances précédentes  $d^*$ , sans revenir aux distances entre éléments terminaux. En effet :

$$d^*(\{a, b\}, \{c, d\}) = \inf\{d^*(\{a, b\}, \{c\}), d^*(\{a, b\}, \{d\})\} = \inf\{1, 0.7\} = 0.7$$

De même :

$$d^*(\{c, d\}, \{e\}) = \inf\{d^*(\{c\}, \{e\}), d^*(\{d\}, \{e\})\} = \inf\{1.5, 1.3\} = 1.3$$

Etape 3 :

La distance minimale observée conduit à agréger les classes  $\{a, b\}$  et  $\{c, d\}$  car  $\min\{d^*\} = d^*(\{a, b\}, \{c, d\}) = 0.7$

Calculons la distance entre cette nouvelle classe et l'élément  $e$  :

$$d^*(\{a, b, c, d\}, e) = \inf(d^*(\{a, b\}, e), d^*(\{c, d\}, e)) = d^*(\{a, b\}, e) = 0.7$$

d'où le tableau des distances :

$d^*$	$\{a, b, c, d\}$	$e$
$\{a, b, c, d\}$	0	0.8
$e$		0

Etape 4 :

$e$  est réuni à  $\{a, b, c, d\}$  pour former  $E$ .

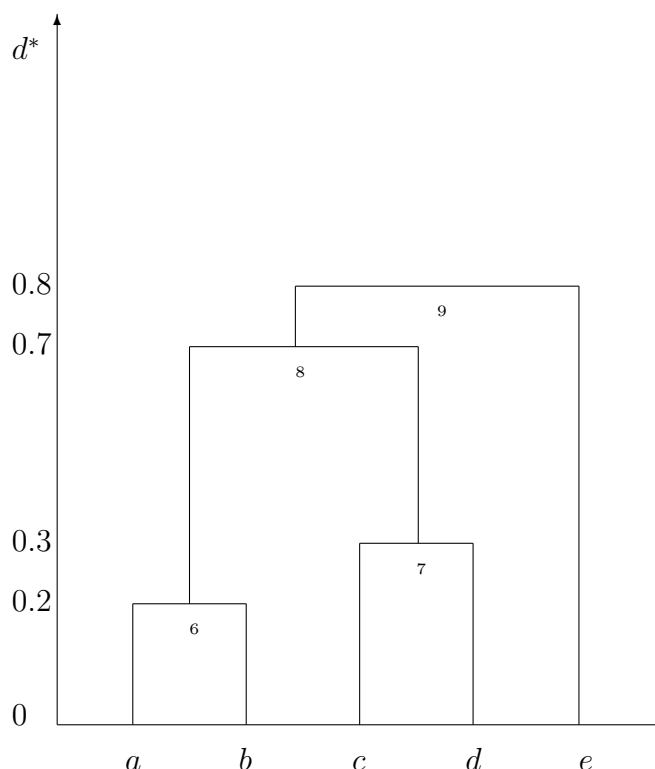
Le déroulement de l'algorithme peut être résumé de la manière suivante :

Etape	Classe	Eléments réunis	Niveau d'agrégation
1	6	$a - b$	0.2
2	7	$c - d$	0.3
3	8	$\{a, b\} - \{c, d\}$	0.7
4	9	$\{a, b, c, d\} - e$	0.8

On obtient ainsi une hiérarchie  $H$  de  $E$  :

$$H = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{c, d\}, \{a, b, c, d\}, \{a, b, c, d, e\}\}$$

Cette hiérarchie indicée peut être représentée par un arbre :



Remarques :

1- Les différentes partitions compatibles avec la hiérarchie sont obtenues en coupant l'arbre par une droite horizontale.

2- La distance utilisée ici entre classes conduit à l'ultramétrie inférieure maxima ou sous-dominante (single linkage). Cette distance est peu utilisée en pratique car elle conduit à des classes allongées (effet de chaînage). Son choix ici résulte de considérations pédagogiques. Cette stratégie d'agrégation conduit en effet à des calculs peu nombreux et simples.

Exemple 2 :

Considérons le même ensemble et le même tableau des distances mais choisissons cette fois une autre stratégie d'agrégation :

$$d^*(h_1, h_2) = \sup\{d(a_i, b_j)\} \quad \text{avec} \quad a_i \in h_1 \quad \text{et} \quad b_j \in h_2$$

La distance entre deux parties est ici la distance maximale entre éléments de chacun des deux ensembles.

Etape 1 :

Comme précédemment les deux éléments les plus proches sont  $a$  et  $b$ . Calculons la distance entre la classe  $\{a, b\}$  et les éléments terminaux restants :

$$d^*(\{a, b\}, c) = \sup\{d(a, c), d(b, c)\} = 1.05$$

$$d^*(\{a, b\}, d) = \sup\{d(a, d), d(b, d)\} = 0.75$$

$$d^*(\{a, b\}, e) = \sup\{d(a, e), d(b, e)\} = 1$$

$d^*$	$\{a, b\}$	$c$	$d$	$e$
$\{a, b\}$	0	1.05	0.75	1
$c$		0	0.3	1.5
$d$			0	1.3
$e$				0

Etape 2 :

On agrège  $c$  et  $d$  au niveau 0.3

Il faut effectuer le calcul des distances entre  $\{c, d\}$  d'une part et la classe  $\{a, b\}$  et l'élément  $e$  d'autre part :

$$d^*(\{c, d\}, \{a, b\}) = \sup\{d^*(c, \{a, b\}), d^*(d, \{a, b\})\} = 1.05$$

$$d^*(\{c, d\}, e) = \sup\{d^*(c, e), d^*(d, e)\} = 1.5$$

$d^*$	$\{a, b\}$	$\{c, d\}$	$e$
$\{a, b\}$	0	1.05	1
$\{c, d\}$		0	1.5
$e$			0

Etape 3 :

L'élément  $e$  est agrégé à la classe  $\{a, b\}$

$d^*$	$\{a, b, e\}$	$c, d$
$\{a, b, e\}$	0	1.5
$c, d$		0

$$d^*(\{a, b, e\}, \{c, d\}) = \sup\{d^*(\{c, d\}, \{a, b\}), d^*(\{c, d\}, e)\} = 1.5$$

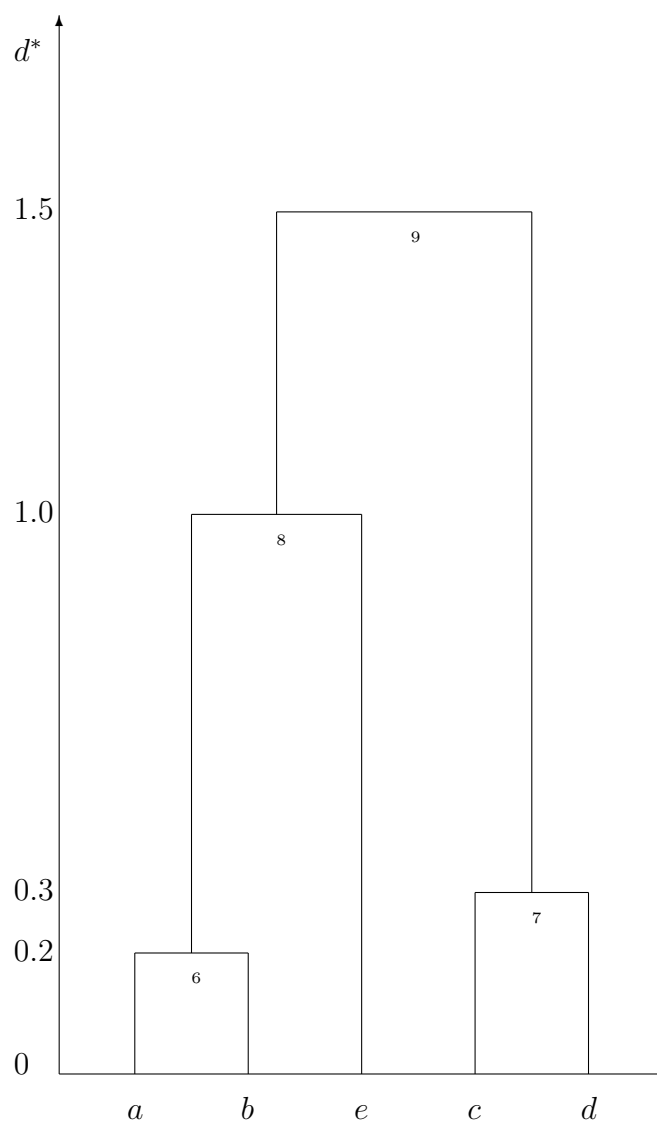
Etape 4 :

La classe  $\{a, b, e\}$  est réunie à la classe  $\{c, d\}$

Résumé :

Etape	Classe	Eléments réunis	Niveau d'agrégation
1	6	$a - b$	0.2
2	7	$c - d$	0.3
3	8	$\{a, b\} - e$	1.0
4	9	$\{a, b, e\} - \{c, d\}$	1.5

L'arbre correspondant à cette classification est le suivant :



On remarque que les deux classifications ne conduisent pas aux mêmes classes. Ainsi selon la stratégie d'agrégation utilisée la hiérarchie obtenue n'est pas la même, ce qui peut laisser penser que les résultats obtenus sont subjectifs. En fait nous verrons que lorsque la CAH est associée à une analyse factorielle quelle qu'elle soit, bien que les distances diffèrent, une stratégie d'agrégation s'impose de manière naturelle.

## 6.4 Stratégie d'agrégation selon la variance. La méthode de Ward

Lorsque l'on dispose de données dans un espace métrique et qu'il est possible de donner un sens à la notion d'inertie, la stratégie d'agrégation la plus utilisée est alors celle de Ward.

Soit un ensemble  $N(I)$  de  $n$  points à classer. On suppose que ces éléments appartiennent à  $\mathbf{R}^p$ .

### 6.4.1 Notation

Soient :

$X[n, p] = (x_{ij})$  les coordonnées des  $n$  éléments dans  $\mathbf{R}^p$ .

$d$  la distance dont est muni l'espace  $\mathbf{R}^p$ .

$p_i$  la masse associée à chaque observation  $i$  avec  $\sum_{i=1}^n p_i = 1$ .

$g = (g_j)$  le barycentre du nuage avec  $g_j = \sum_{i=1}^n p_i x_{ij}$ .

Soit par ailleurs  $P_q$  une partition de  $N(I)$  en  $q$  classes  $C_q$  ( $q = 1, Q$ ).

Notons  $g_q = (g_{jq})$  le barycentre de la classe  $q$  avec  $g_{jq} = \frac{\sum_{i \in C_q} p_i x_{ij}}{p_q}$

où  $p_q = \sum_{i \in C_q} p_i$ .



### 6.4.2 Décomposition de l'inertie

L'inertie totale du nuage par rapport à son centre de gravité peut alors d'après la formule de Huygens se décomposer de la manière suivante :

$$I_G = \sum_{q=1}^Q \sum_{i \in C_q} p_i \|x_i - g_q\|^2 + \sum_{q=1}^Q p_q \|g_q - g\|^2$$

La première quantité est l'inertie à l'intérieur des classes ou inertie intra-classes. La seconde est l'inertie entre les classes ou inertie inter-classes, c'est à dire mesurée à partir des centres de gravité des classes.

Démonstration :

$$I_G = \sum_{i=1}^n p_i \|x_i - g\|^2 = \sum_{q=1}^Q \sum_{i \in C_q} p_i \|x_i - g\|^2$$

soit encore :

$$I_G = \sum_{q=1}^Q \sum_{i \in C_q} p_i \|x_i - g_q + g_q - g\|^2$$

d'où :

$$I_G = \sum_{q=1}^Q \sum_{i \in C_q} p_i [\|x_i - g_q\|^2 + \|g_q - g\|^2 + 2^t (x_i - g_q) (g_q - g)]$$

finalement :

$$I_G = \sum_{q=1}^Q \sum_{i \in C_q} p_i \|x_i - g_q\|^2 + \sum_{q=1}^Q \sum_{i \in C_q} p_i \|g_q - g\|^2 + 2 \sum_{q=1}^Q \sum_{i \in C_q} p_i^t (x_i - g_q) (g_q - g)$$

Examinons chacun de ces termes :

Le premier revient à sommer sur l'ensemble des  $Q$  classes la quantité  $\sum_{i \in C_q} p_i \|x_i - g_q\|^2$

qui correspond à la l'inertie de la classe  $q$ , c'est à dire mesurée par rapport à  $g_q$  et sur les seuls éléments de la classe  $q$ . Cette quantité est appelée inertie interne ou intra.

Le second terme,  $\sum_{q=1}^Q \sum_{i \in C_q} p_i \|g_q - g\|^2$ , peut encore s'écrire  $\sum_{q=1}^Q \|g_q - g\|^2 \sum_{i \in C_q} p_i$ ,

soit  $\sum_{q=1}^Q p_q \|g_q - g\|^2$ . Il correspond à l'inertie calculée à partir des barycentres

des  $Q$  classes induites par la partition. Cette quantité est appelée inertie entre les classes ou inertie inter.

Montrons que le troisième terme est nul

$$\sum_{q=1}^Q \sum_{i \in C_q} p_i^t (x_i - g_q) (g_q - g) = \sum_{q=1}^Q {}^t (g_q - g) \sum_{i \in C_q} p_i (x_i - g_q)$$

or

$$\sum_{i \in C_q} p_i (x_i - g_q) = \sum_{i \in C_q} p_i x_i - g_q \sum_{i \in C_q} p_i = p_q \frac{1}{p_q} \sum_{i \in C_q} p_i x_i - p_q g_q$$

comme

$$\frac{1}{p_q} \sum_{i \in C_q} p_i x_i = g_q$$

il vient

$$\sum_{i \in C_q} p_i (x_i - g_q) = p_q g_q - p_q g_q = 0$$

d'où le résultat.

### 6.4.3 Critère de Ward

Examinons maintenant l'incidence de cette propriété sur le déroulement de l'algorithme de classification ascendante hiérarchique.

Au départ chaque classe étant réduite à un élément, l'inertie intra-classes est nulle et l'inertie totale se confond avec l'inertie inter-classes. A l'arrivée, lorsque tous les éléments du nuage sont réunis en une seule classe, l'inertie inter-classes est nulle et l'inertie totale égale à l'inertie intra-classes.

Ainsi, au cours du déroulement de l'algorithme, lors du passage d'une partition  $q$  classes à une partition en  $q - 1$  classes, l'inertie intra-classes augmente et l'inertie inter-classes diminue de la même quantité en raison de la propriété vue plus haut.

L'inertie intra constitue un indicateur de la qualité d'une partition. En effet, plus elle est importante et plus cela indique que les éléments à l'intérieur de chacune des classes sont hétérogènes. En conséquence on décidera de regrouper les deux classes, parmi les  $q$ , qui entraînent le plus faible gain d'inertie intra ou, ce qui revient au même, la plus faible perte d'inertie inter.

Considérons la partition  $P_{q-1}$  obtenue en fusionnant deux éléments  $a$  et  $b$  de la partition  $P_q$  précédente. Cherchons la perte d'inertie inter-classes créée

par la fusion des deux classes  $a$  et  $b$ .

Nous allons montrer que cette perte d'inertie est égale à :

$$\Delta_{(a,b)} = \frac{p_a p_b}{p_a + p_b} d^2(g_a, g_b)$$

Démonstration :

Dans le passage de  $q$  classes à  $q - 1$  du à la réunion des classes  $a$  et  $b$ , la perte d'inertie inter correspond à l'inertie inter de chacune des deux classes et le gain à l'inertie inter de la classe résultant de la fusion, soit :

$$\Delta_{(a,b)} = p_a \|g_a - g\|^2 + p_b \|g_b - g\|^2 - p_{a \cup b} \|g_{a \cup b} - g\|^2 \quad (6.1)$$

où  $g_{a \cup b}$  désigne le barycentre de la classe résultant de la fusion des classes  $a$  et  $b$ , soit :

$$g_{a \cup b} = \frac{p_a g_a + p_b g_b}{p_a + p_b}$$

par ailleurs :

$$\|g_{a \cup b} - g\|^2 = \left\| \frac{p_a g_a + p_b g_b}{p_a + p_b} - g \right\|^2 = \frac{1}{(p_a + p_b)^2} \|p_a g_a + p_b g_b - (p_a + p_b)g\|^2$$

soit encore :

$$\|g_{a \cup b} - g\|^2 = \frac{1}{(p_a + p_b)^2} \|p_a(g_a - g) + p_b(g_b - g)\|^2$$

soit :

$$\begin{aligned} & \|g_{a \cup b} - g\|^2 \\ &= \frac{1}{(p_a + p_b)^2} [p_a^2 \|g_a - g\|^2 + p_b^2 \|g_b - g\|^2 + 2p_a p_b {}^t(g_a - g)(g_b - g)] \end{aligned}$$

d'où :

$$\begin{aligned} & (p_a + p_b) \|g_{a \cup b} - g\|^2 \\ &= \frac{1}{(p_a + p_b)} [p_a^2 \|g_a - g\|^2 + p_b^2 \|g_b - g\|^2 + 2p_a p_b {}^t(g_a - g)(g_b - g)] \end{aligned}$$

par conséquent, la perte d'inertie inter-classes est :

$$\begin{aligned}
\Delta_{(a,b)} &= p_a \|g_a - g\|^2 + p_b \|g_b - g\|^2 - p_{a \cup b} \|g_{a \cup b} - g\|^2 \\
&= p_a \|g_a - g\|^2 + p_b \|g_b - g\|^2 \\
&\quad - \frac{1}{p_a + p_b} [p_a^2 \|g_a - g\|^2 + p_b^2 \|g_b - g\|^2 + 2p_a p_b^t (g_a - g)(g_b - g)] \\
&= \frac{p_a p_b}{p_a + p_b} [\|g_a - g\|^2 + \|g_b - g\|^2 - 2^t (g_a - g)(g_b - g)] \\
&= \frac{p_a p_b}{p_a + p_b} \|g_a - g_b\|^2 \\
&= \frac{p_a p_b}{p_a + p_b} d^2(g_a, g_b)
\end{aligned}$$

d'où le résultat :

$$\Delta_{(a,b)} = \frac{p_a p_b}{p_a + p_b} d^2(g_a, g_b)$$

La méthode de Ward peut donc s'interpréter comme la construction d'une hiérarchie indicée en prenant comme distance  $d^*$  entre parties la perte d'inertie inter-classes ou ce qui est équivalent le gain d'inertie intra-classes, soit :

$$d^*(a, b) = \frac{p_a p_b}{p_a + p_b} d^2(g_a, g_b)$$

La poursuite du déroulement de l'algorithme exige la mise à jour des distances entre la classe  $a \cup b$  et les éléments restants.

Nous allons montrer que la distance entre la classe  $a \cup b$  et un élément quelconque  $c$  s'écrit :

$$d^*(a \cup b, c) = \frac{1}{p_a + p_b + p_c} [(p_a + p_c) d^*(a, c) + (p_b + p_c) d^*(b, c) - p_c d^*(a, b)]$$

En appliquant la formule précédente à  $a \cup b$  d'une part et à  $c$  de l'autre, alors :

$$d^*(a \cup b, c) = \frac{(p_{a \cup b}) p_c}{p_{a \cup b} + p_c} d^2(g_{a \cup b}, g_c)$$

De plus, la formule précédente (6.1) peut être appliquée à l'élément  $c$ , soit :

$$\Delta_{(a,b)} = d^*(a, b) = p_a d^2(g_a, g_c) + p_b d^2(g_b, g_c) - p_{a \cup b} d^2(g_{a \cup b}, g_c)$$

d'où :

$$p_{a \cup b} d^2(g_{a \cup b}, g_c) = p_a d^2(g_a, g_c) + p_b d^2(g_b, g_c) - d^*(a, b)$$

Multiplions les termes des deux côtés par  $\frac{p_c}{p_{a \cup b}} = \frac{p_c}{p_a + p_b + p_c}$  :

$$\begin{aligned} \frac{p_{a \cup b} p_c}{p_a + p_b + p_c} d^2(g_{a \cup b}, g_c) &= d^*(a \cup b, c) \\ &= \frac{p_{a \cup b} p_c}{p_a + p_b + p_c} [p_a d^2(g_a, g_c) + p_b d^2(g_b, g_c) - d^*(a, b)] \end{aligned}$$

soit :

$$d^*(a \cup b, c) = \frac{(p_a + p_b) p_c}{p_a + p_b + p_c} [p_a d^2(g_a, g_c) + p_b d^2(g_b, g_c) - d^*(a, b)]$$

or :

$$p_a p_c d^2(g_a, g_c) = (p_a + p_c) \frac{p_a p_c}{p_a + p_c} d^2(g_a, g_c) = (p_a + p_c) d^*(a, c)$$

de même :

$$p_b p_c d^2(g_b, g_c) = (p_b + p_c) d^*(b, c)$$

par conséquent :

$$d^*(a \cup b, c) = \frac{1}{p_a + p_b + p_c} [(p_a + p_c) d^*(a, c) + (p_b + p_c) d^*(b, c) - p_c d^*(a, b)]$$

Remarquons que la distance entre la classe formée de la réunion de  $a$  et de  $b$  d'une part et de  $c$  de l'autre peut être calculée à partir des seules distances  $d^*$  entre ces trois éléments.

Le critère de Ward va s'appliquer de manière naturelle sur des données traitées en Analyse Factorielle, quelle que soit la méthode ACP, AFC ou ACM. Seules les données et la métrique initiale vont changer, métrique usuelle en ACP normée et distance du chi-2 en AFC et ACM. Il est par ailleurs toujours possible d'utiliser la métrique usuelle directement sur les coordonnées factorielles.

Remarques :

1- Dans ce cas il ne faut surtout pas utiliser la métrique inverse des variances et travailler sur des coordonnées factorielles réduites. En effet si l'on se souvient que la variance de chaque facteur n'est autre que l'inertie conservée par l'axe correspondant, l'utilisation de la métrique usuelle,  $M = I$ , conserve à

chaque facteur son importance. A l'inverse de l'ACP dans laquelle il faut souvent réduire les variables, la méthode de Ward sur coordonnées factorielles implique le choix de la distance euclidienne usuelle.

2- Dans ce dernier cas attention aussi à ne pas sélectionner un nombre d'axes faible. La CAH sur coordonnées factorielles n'est équivalente à celle réalisée sur les données de base que dans le cas où la totalité des facteurs est prise en compte. Souvent cette dernière solution, surtout en ACM, n'est guère réaliste. Toutefois la prise en compte d'un nombre trop restreint de facteurs aura tendance à améliorer de manière artificielle la qualité des partitions obtenues. En effet en éliminant les facteurs résiduels on renforce le poids des premiers facteurs ce qui ne peut qu'améliorer les classes.

#### 6.4.4 Exemple

Appliquons la stratégie d'agrégation de Ward à l'exemple introductif. Il est d'abord nécessaire de calculer les distances  $d^*$  entre éléments terminaux, soit :

$$d^*(a, b) = \frac{p_a p_b}{p_a + p_b} d^2(g_a, g_b)$$

Ici, les masses sont égales, soit  $p_i = \frac{1}{5}$ , par conséquent :

$$\frac{p_a p_b}{p_a + p_b} = \frac{\frac{1}{5} \cdot \frac{1}{5}}{\frac{1}{5} + \frac{1}{5}} = \frac{1}{10}$$

Les distances initiales deviennent donc :

$d^*$	$a$	$b$	$c$	$d$	$e$
$a$	0	0.004	0.1	0.049	0.1
$b$		0	0.11	0.056	0.064
$c$			0	0.009	0.225
$d$				0	0.169
$e$					0

Les deux éléments les plus proches demeurent toujours les éléments  $a$  et  $b$  qui vont être réunis dans la classe numérotée 6.

Il reste maintenant à mettre à jour les distances entre cette classe et les autres éléments, ici terminaux, restants. A titre d'exemple, calculons la distance entre la classe  $\{a, b\}$  et  $c$  :

$$d^*(\{a, b\}, c) = \frac{1}{m_a + m_b + m_c} [(p_a + p_c)d^*(a, c) + (p_b + p_c)d^*(b, c) - p_c d^*(a, b)]$$

d'où :

$$d^* (\{a, b\}, c) = \frac{5}{3} \left[ \frac{2}{5}(0.1) + \frac{2}{5}(0.11) - \frac{1}{5}(0.004) \right] = 0.139$$

A l'issue de cette première étape le tableau des distances entres parties devient :

$d^*$	$\{a, b\}$	$c$	$d$	$e$
$\{a, b\}$	0	0.139	0.069	0.108
$c$		0	0.009	0.225
$d$			0	0.169
$e$				0

A l'étape 2 les éléments  $c$  et  $d$  vont être réunis pour former la classe 7 et les distances deviennent :

$d^*$	$\{a, b\}$	$\{c, d\}$	$e$
$\{a, b\}$	0	0.152	0.108
$\{c, d\}$		0	0.260
$e$			0

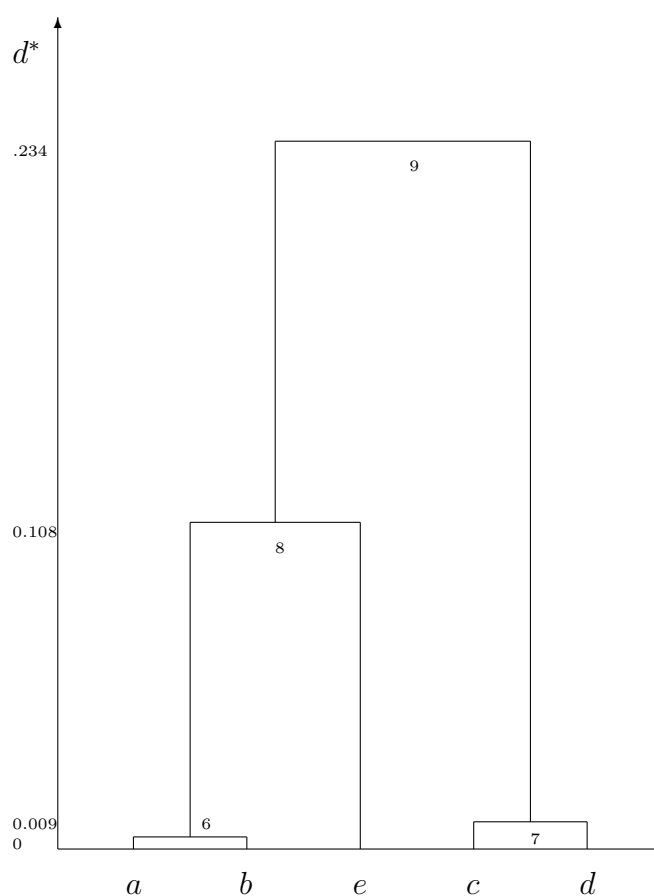
L'étape suivante conduit à la fusion des éléments  $\{a, b\}$  et  $e$  et la mise à jour des distances donne le tableau suivant :

$d^*$	$\{a, b, e\}$	$\{c, d\}$
$\{a, b, e\}$	0	0.234
$\{c, d\}$		0

La dernière étape conduit à la réunion des classes  $\{a, b, e\}$  et  $\{c, d\}$ . Le résumé du déroulement de l'algorithme figure dans le tableau suivant :

Etape	Classe	Eléments réunis	Niveau d'agrégation
1	6	$a - b$	0.04
2	7	$c - d$	0.09
3	8	$\{a, b\} - e$	0.108
4	9	$\{a, b, e\} - \{c, d\}$	0.234

L'arbre de classification :



Illustrons sur cet exemple l'équivalence entre la construction d'une hiérarchie indicée de  $E$  et la définition d'une distance ultramétrique sur  $E$ . Définissons la distance suivante entre éléments de  $E$  :

$$d(x, y) = d^*(\{x\}, \{y\})$$

En d'autres termes, la distance entre deux éléments est l'indice de diamètre de la plus petite partie contenant les éléments  $x$  et  $y$ . Les distance entre éléments de  $E$  figurent dans le tableau suivant :

$d^*$	$a$	$b$	$c$	$d$	$e$
$a$	0	0.004	0.234	0.234	0.108
$b$		0	0.234	0.234	0.108
$c$			0	0.09	0.234
$d$				0	0.234
$e$					0



## 6.5 Autres stratégies d'agrégation

D'autres stratégies d'agrégation peuvent être utilisées, bien que celle de Ward s'impose assez logiquement en complément d'une Analyse Factorielle. Citons de manière non-exhaustive :

La stratégie « barycentrique » :

$$d^*(a, b) = d^2(g_a, g_b)$$

Cette distance entre parties est calculée sur les seuls barycentres des deux parties, donc sans tenir compte de leurs poids respectifs.

La stratégie « moment-classe » :

$$d^*(a, b) = I(a \cup b / g_{a \cup b}) = \sum_{i \in a \cup b} p_i d^2(i, g_{a \cup b})$$

La stratégie variance :

$$d^*(a, b) = \frac{1}{p_{a \cup b}} \sum_{i \in a \cup b} p_i d^2(i, g_{a \cup b})$$

...

## 6.6 Application

Afin d'illustrer le déroulement d'un algorithme de classification nous allons reprendre l'exemple traité dans le premier chapitre.

L'algorithme commence par calculer les distances entre les 15 éléments (années) deux à deux. La distance utilisée ici est la distance euclidienne usuelle sur coordonnées factorielles, la totalité des axes ayant été utilisée.

Les deux éléments les plus proches sont 3 et 4, soient les années 1971 et 1972.

CLASSIFICATION HIERARCHIQUE : DESCRIPTION DES NOEUDS

NUM.	AINÉ	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
16	3	4	2	2.00	.02175	*
17	9	8	2	2.00	.03050	*
18	12	13	2	2.00	.03606	*
19	14	6	2	2.00	.07679	**
20	7	10	2	2.00	.10058	**
21	2	1	2	2.00	.10912	***
22	17	20	4	4.00	.14132	***
23	16	5	3	3.00	.15747	****
24	18	11	3	3.00	.18964	****
25	19	15	3	3.00	.23953	*****
26	23	21	5	5.00	.41629	*****
27	25	24	6	6.00	.49056	*****
28	27	22	10	10.00	1.88712	*****
29	28	26	15	15.00	4.10326	*****
SOMME DES INDICES DE NIVEAU =						8.00000

La fusion de ces deux éléments se traduit par la constitution d'un nouvel élément numéroté 16, correspondant à leur barycentre. Cette fusion entraîne une perte d'inertie inter-classes égale à 0.02175 figurant dans la colonne INDICE. Les distances entre cette classe et les 13 éléments terminaux sont calculées et le processus est réitéré.

A l'étape 2 ce sont les éléments terminaux 8 et 9 (1976 et 1977) qui sont réunis pour former la classe 17.

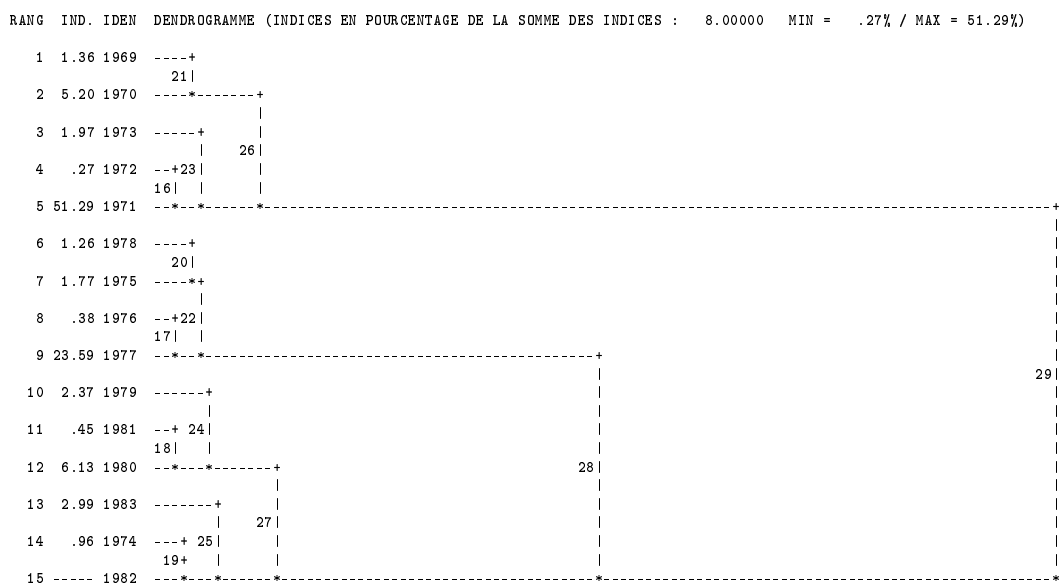
On aboutit finalement à une seule classe (29) rassemblant la totalité des éléments à classer.

On observe également la croissance de l'inertie intra-classes au fur et à mesure du déroulement de l'algorithme, indicatrice de l'hétérogénéité croissante des classes ainsi constituées.

Remarque :

La somme des augmentations d'inertie intra-classes est égale à l'inertie totale du nuage, soit ici 8 (en ACP normée l'inertie est égale au nombre de variables actives).

La représentation sous forme d'arbre :



Le résultat de la classification figure sur le graphique ci-dessus. Les longueurs des branches sont proportionnelles à la perte d'inertie. Le numéro des classes a été rajouté sur l'arbre alors que dans la plupart des cas il n'y figure pas.

Utilisation de la classification :

Pour obtenir une partition il suffit de couper l'arbre de classification par

une droite, ici verticale. Le problème concret qui se pose est celui du choix du niveau de coupure. Si l'on se souvient que la longueur des branches représente l'augmentation de l'inertie intra-classes résultant de l'agrégation de deux classes, on voit que l'on aura intérêt à couper l'arbre avant un saut important. En effet, un saut important dénote une forte perte d'inertie inter-classes donc une forte augmentation de l'inertie intra-classes, indicateur d'une grande hétérogénéité des classes réunies à cette étape. Cette démarche est résumée de manière imagée par la formule suivante due à Michel VOLLE : « Il faut couper les branches de l'arbre lorsqu'elles sont longues ».

Ici l'examen du déroulement de l'algorithme fait ressortir 3 niveaux de coupure possibles, en l'occurrence les classes 28, 27 et 25.

La coupure sur la classe 28 entraîne une partition en deux groupes, 26 et 28. Cette coupure correspond à l'opposition entre les années antérieures au premier choc pétrolier, 1969 à 1973, et les autres années.

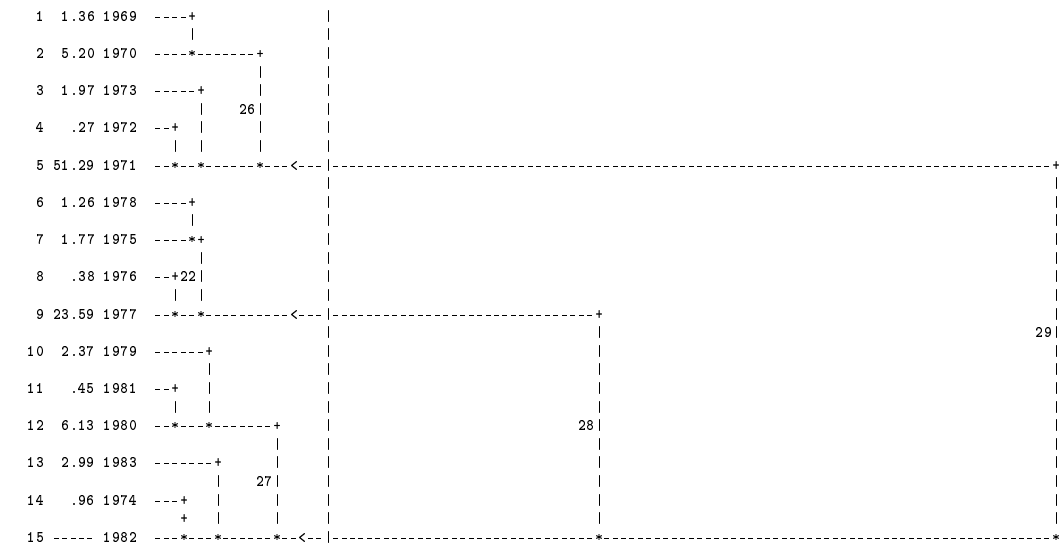
La coupure sur la classe 27 se traduit par une partition en 3 classes, 27, 22 et 26.

Le procédé pratique pour déterminer les classes composant la partition est le suivant : on trace une ligne horizontale après la classe sur laquelle on coupe l'arbre et on retient toutes les classes dont le numéro est inférieur à celui de cette classe. En effet si le numéro est inférieur cela indique que les classes ont été formées antérieurement et donc qu'elles demeurent dans la partition.

#### CLASSIFICATION HIERARCHIQUE : DESCRIPTION DES NOEUDS

NUM.	AINÉ	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
16	3	4	2	2.00	.02175	*
17	9	8	2	2.00	.03050	*
18	12	13	2	2.00	.03606	*
19	14	6	2	2.00	.07679	**
20	7	10	2	2.00	.10058	**
21	2	1	2	2.00	.10912	***
22	17	20	4	4.00	.14132	***
23	16	5	3	3.00	.15747	****
24	18	11	3	3.00	.18964	****
25	19	15	3	3.00	.23953	*****
26	23	21	5	5.00	.41629	*****
27	25	24	6	6.00	.49056	*****
-----						
28	27	22	10	10.00	1.88712	*****
29	28	26	15	15.00	4.10326	*****

La partition résultante peut également être obtenue à partir de l'arbre de classification mais le procédé est moins précis en raison d'une part des problèmes de superpositions et de l'absence des numéros de classe d'autre part.



La composition des trois classes est la suivante :  
COMPOSITION DE : COUPURE 'a' DE L'ARBRE EN 3 CLASSES

```

----- CLASSE 1 / 3 -----
1969 1970 1971 1972 1973
----- CLASSE 2 / 3 -----
1975 1976 1977 1978
----- CLASSE 3 / 3 -----
1974 1979 1980 1981 1982 1983

```

Remarques :

1- C'est le choix du niveau de coupure qui va déterminer le nombre de classes. Ce dernier ne peut être fixé *a priori*. Toutefois, certains logiciels demandent le nombre de classes et non le niveau de coupure

2- La coupure en deux classes correspond très souvent au phénomène mis en évidence sur le premier axe factoriel.

3- Bien que le nombre de classes puisse être très élevé, la plupart du temps il reste limité, en général inférieur à la dizaine. La multiplication des classes rend les différences entre celles-ci de plus en plus ténues et donc de plus en plus difficiles à interpréter.

4- Ici la coupure sur la classe 24, bien que ce niveau n'ait pas été retenu, entrainerait la création d'une classe réduite à un élément (15) soit l'année 1983. La présence de classes réduites à un élément dans le haut de l'arbre dénote souvent un élément particulier probablement déjà mis en évidence par

l'Analyse Factorielle.

5- La CAH a été réalisée ici sur les seuls éléments actifs de l'analyse. L'inclusion d'éléments supplémentaires est toujours possible, la partition choisie fournissant un groupe d'affectation pour ces éléments. Il existe toutefois des méthodes spécifiques pour ce genre de problème. Il est aussi possible une fois la partition réalisée à partir des seuls éléments actifs de calculer la distance entre chaque classe et l'élément supplémentaire et de l'affecter à la classe la plus proche.

## 6.7 Les aides à l'interprétation en classification

En règle générale les éléments suivants sont fournis :

Le quotient entre inertie inter-classes et inertie totale, fournit un indicateur de la qualité de la partition retenue. Ici ce rapport est égal à 0.75. Signalons toutefois que ce critère n'est pas adapté au choix du niveau de coupure. En effet, compte tenu du déroulement de l'algorithme de Ward, il reviendra toujours à sélectionner la partition conduisant au plus grand nombre de classes. Le critère à utiliser est celui de l'évolution de l'inertie intra-classes.

Les aides à l'interprétation visent à préciser quelles variables caractérisent le plus les classes obtenues. Plusieurs indicateurs peuvent être utilisés.

### 6.7.1 La notion de valeur-test

Cette notion adéjà été vue lors du chapitre précédent consacré à l'Analyse Factorielle des Correspondances Multiples. L'idée de base est de comparer l'écart observé pour chaque variable entre la valeur sur la classe et la valeur prise sur la totalité. L'écart constaté est-il ou non significatif?

Application au cas des variables continues :

Soit  $m_k$  la moyenne de la variable  $X$  dans la classe  $k$  
$$m_k = \frac{\sum_{i \in I_k} p_i x_i}{\sum_{i \in I_k} p_i}.$$

Si la variable  $X$  n'est pas caractéristique de la classe  $k$  on peut alors considérer que les valeurs observées dans cette classe, résumées par  $m_k$ , ne se différencient pas de celles que l'on observerait en effectuant un tirage aléatoire sans remise de  $n_k$  éléments parmi  $n$ .

On sait (voir votre cours de théorie des sondages) que dans un sondage aléatoire simple, la moyenne d'échantillon est un estimateur sans biais de la moyenne  $m$  dans la population et que la variance de cet estimateur est :

$$V(m_k) = \left(1 - \frac{n_k}{n}\right) \frac{S_x^2}{n_k}$$

où  $S_X^2$  désigne la dispersion de  $X$ , soit :

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2$$

Il est immédiat que :

$$V(m_k) = \left(1 - \frac{n_k}{n}\right) \frac{n}{n-1} \frac{V(X)}{n_k}$$

soit encore :

$$V(m_k) = \frac{n - n_k}{n - 1} \frac{V(X)}{n_k}$$

Ici à l'inverse de ce qui se passe lors d'un sondage,  $V(X)$  étant connu il n'y a pas lieu de l'estimer.

La statistique du test est donc ici  $VT = \frac{m_k - m}{\sqrt{\frac{n - n_k}{n - 1} \frac{V(X)}{n_k}}}$  qui est comparée

au fractile de la loi normale centrée-réduite. Plus la valeur-test est importante et plus on est amené à rejeter l'hypothèse que la valeur observée  $m_k$  puisse résulter d'un tirage aléatoire donc ne pas se différencier de  $m$ . Soit la valeur-test est fournie, soit le niveau de signification du test.

L'interprétation probabiliste n'est licite en toute rigueur que pour les variables supplémentaires. Il n'est pas possible en effet de tester l'indépendance entre les classes d'une partition et les variables actives utilisées pour réaliser cette partition. Toutefois pour les variables actives le classement des valeurs-test par valeurs décroissantes facilite la caractérisation de la classe.

Il est à noter que l'observation des valeurs absolues des valeurs-tests prises dans chaque classe permet aussi d'apprécier le degré de caractérisation de chacune des classes dans la partition retenue.

Application :

Cherchons les variables actives les plus caractéristiques de la classe 1 (1969-1973).

Cherchons par exemple la valeur-test pour la variable NET :

$m_k = 17.55$ ,  $m = 13.91$  et  $V(X) = (2.89)^2$ , d'où :

$$V(m_k) = \frac{n - n_k}{n_k} \frac{V(X)}{n - 1} = \frac{15 - 5}{5} \frac{2.89^2}{15 - 1} = 1.19 \Rightarrow \sqrt{V(m_k)} = 1.09$$

d'où :

$$\frac{m_k - m}{\sqrt{V(m_k)}} = \frac{17.55 - 13.91}{1.09} = 3.33$$

CARACTERISATION  
PAR LES CONTINUES  
DES CLASSES OU MODALITES DE : COUPURE 'a' DE L'ARBRE EN 3 CLASSES

V.TEST	PROBA	MOYENNES		ECARTS TYPES		VARIABLES CARACTERISTIQUES	IDEN
		CLASSE GENERALE	CLASSE GENERAL			NUM.LIBELLE	
		CLASSE 1 / 3	( POIDS = 5.00	EFFECTIF = 5 )			aa1a
3.33	.000	17.55	13.91	1.03	2.89	1.Situation nette	NET
2.96	.002	25.59	22.11	.90	3.11	6.Immobilisations	IMM
-2.49	.006	1.52	2.19	.52	.71	3.Subventions	SUB
-3.33	.000	5.40	10.21	.42	3.82	7.Valeurs d'exploitation	EXP
-3.47	.000	19.21	22.26	.86	2.33	5.Dettes É court terme	DCT
		CLASSE 2 / 3	( POIDS = 4.00	EFFECTIF = 4 )			aa2a
3.01	.001	11.20	9.00	.45	1.66	4.Dettes É long et moyen terme	LMT
-2.34	.010	10.91	13.91	1.55	2.89	1.Situation nette	NET
-2.42	.008	15.94	17.18	.33	1.15	8.Valeurs rCalisables et disponibles	VRD
-2.57	.005	2.54	3.14	.24	.53	2.IntCrets	INT
		CLASSE 3 / 3	( POIDS = 6.00	EFFECTIF = 6 )			aa3a
2.88	.002	13.81	10.21	1.90	3.82	7.Valeurs d'exploitation	EXP
2.38	.009	2.75	2.19	.57	.71	3.Subventions	SUB
-2.94	.002	7.40	9.00	.49	1.66	4.Dettes É long et moyen terme	LMT
-3.39	.000	18.66	22.11	.76	3.11	6.Immobilisations	IMM

Le classement des valeurs-test permet de déterminer les variables les plus caractéristiques de chacune des classes. Il vient ici confirmer l'interprétation de l'ACP réalisée auparavant. Le signe de la valeur-test permet également de voir directement si la moyenne dans la classe est supérieure ou inférieure à celle de l'ensemble.

La notion de valeur-test peut également être utilisée afin de déterminer sur quel(s) axe(s) les différentes classes d'une partition sont le mieux représentées. On sait en effet que les coordonnées des individus sur un axe sont centrées. Soit  $c_{kq}$  la coordonnée du barycentre de la classe  $q$  sur l'axe de rang  $k$ . On examine si  $c_{kq}$  peut être considéré comme résultant du tirage aléatoire de  $n_q$  individus parmi  $n$ . Ici comme dans le cas précédent la valeur-test possède une valeur indicatrice et leur classement par ordre décroissant permet de détecter

les axes sur lesquels la classe est la mieux représentée. On examine ainsi la quantité :

$$\frac{c_{kq} - 0}{\sqrt{V_q}}$$

où :

$$V_q = \left(1 - \frac{n_q}{n}\right) \frac{1}{n_q} \frac{n}{n-1} \lambda_k = \frac{n - n_q}{n-1} \frac{\lambda_k}{n_q}$$

dans laquelle  $\lambda_k$  désigne la variance de  $C_k$ , soit l'inertie  $I_k$  portée par l'axe  $k$  égale à la valeur propre d'ordre  $k$  de la matrice d'inertie.

Application :

Cherchons la valeur-test sur l'axe 1 pour la classe 1 de la partition en trois groupes.

Ici  $c_{kq} = 2.86$ ,  $n_q = 5$ ,  $n = 15$  et  $\lambda_k = 4.4904$  donc  $\frac{c_{kq}}{\sqrt{V_q}} = \frac{2.86}{\sqrt{0.64}} = 3.6$

COORDONNEES ET VALEURS-TEST SUR LES AXES 1 A 5															
CLASSES						VALEURS-TEST					COORDONNEES				
IDEN	LIBELLE	EFF.	P. ABS	1	2	3	4	5	1	2	3	4	5	DISTO.	
COUPURE 'a' DE L'ARBRE EN 3 CLASSES															
aa1a	- CLASSE 1 / 3	5	5.00	3.6	-.2	-.3	-.4	.4	2.86	-.11	-.08	-.11	.06	8.21	
aa2a	- CLASSE 2 / 3	4	4.00	-1.2	3.2	.1	.4	-.5	-1.12	2.09	.03	.14	-.09	5.65	
aa3a	- CLASSE 3 / 3	6	6.00	-2.4	-2.7	.2	.0	.0	-1.63	-1.30	.05	.00	.01	4.37	

## 6.7.2 Caractérisation de la partition par les variables

On peut chercher à apprécier non plus classe par classe mais globalement le rôle de chaque variable dans la partition. Pour cela on se fonde sur un test d'égalité des moyennes dans chaque classe. Sous l'hypothèse d'égalité des moyennes, la quantité  $\frac{n-q}{q-1} \frac{V_{Inter}}{V_{Intra}}$  suit une loi de Fisher à  $q-1$  et  $n-q$  degrés de liberté,  $q$  désignant le nombre de classes dans la partition et  $n$  le nombre total d'individus.

Application :

Cherchons cette quantité pour la variable NET :

$j$	$n_j$	$\mu_j$	$\sigma_j$
1	5	17.552	1.030
2	4	10.913	1.551
3	6	12.875	2.893
	15	13.911	0.573



$$V_{Inter} = \sum_{j=1}^3 \frac{n_j}{n} (\mu_j - \mu)^2 = 7.2451$$

$$V_{Intra} = \sum_{j=1}^3 \frac{n_j}{n} \frac{1}{n_j} \sum_{k \in C_j} (x_{jk} - \mu_k)^2 = 1.1265$$

d'où  $\frac{n-q}{q-1} \frac{V_{Inter}}{V_{Intra}} = 38.59$  (38.61 sur le listage plus bas).

Les valeurs afférentes aux différentes variables apparaissent dans le tableau suivant :

1 . NET - Situation nette							
CLASSES	EFFECTIF	POIDS	MOYENNE	ECART TYPE	MINIMUM	MAXIMUM	
aa1a - CLASSE 1 / 3	5	5.00	17.552	1.030	16.210	19.010	
aa2a - CLASSE 2 / 3	4	4.00	10.913	1.551	9.460	13.430	
aa3a - CLASSE 3 / 3	6	6.00	12.875	.573	11.750	13.430	
ENSEMBLE	15	15.00	13.911	2.893	9.460	19.010	
FISHER = 38.61 / 12 DEGRES DE LIBERTE AU DENOMINATEUR							
PROBA ( FISHER > 38.61 ) = .000 / VALEUR-TEST = 4.39							
4 . LMT - Dettes È long et moyen terme							
CLASSES	EFFECTIF	POIDS	MOYENNE	ECART TYPE	MINIMUM	MAXIMUM	
aa1a - CLASSE 1 / 3	5	5.00	9.148	.899	7.380	9.820	
aa2a - CLASSE 2 / 3	4	4.00	11.205	.449	10.720	11.810	
aa3a - CLASSE 3 / 3	6	6.00	7.405	.487	6.760	8.100	
ENSEMBLE	15	15.00	8.999	1.655	6.760	11.810	
FISHER = 33.31 / 12 DEGRES DE LIBERTE AU DENOMINATEUR							
PROBA ( FISHER > 33.31 ) = .000 / VALEUR-TEST = 4.22							

### 6.7.3 Les aides à l'interprétation pour les variables qualitatives (ACM)

On retrouve les mêmes éléments, à savoir les valeurs-tests pour les modalités, pour les barycentres des classes issues de la partition et le classement des variables dans la caractérisation globale de la partition. Ces différents éléments seront illustrés sur l'étude des bords de parcelle présentée dans le chapitre consacré à l'ACM.

#### Les valeurs-tests pour les modalités

La démarche est identique à celle utilisée pour les variables continues, il s'agit ici d'une proportion et non d'une moyenne. Considérons la modalité 1 de la variable ebve (efficacité brise-vent éloignée ou nulle). Dans l'ensemble de la population 33.13 % des bords de parcelle étudiées possèdent cette modalité. Dans la classe 1 issue de la partition en huit groupes, 97.42 % sont dans ce

cas.

Soit  $n_j$  l'effectif dans la modalité  $j$  et  $n$  l'effectif total. La proportion de la modalité  $j$  dans la population est donc de  $\frac{n_j}{n}$ .

Soit  $n_{jk}$  l'effectif possédant la modalité  $j$  dans la classe  $k$  d'effectif  $n_k$ . La proportion observée dans la classe  $k$  est donc  $\frac{n_{jk}}{n_k}$ .

Si la proportion observée dans la classe  $k$  ne se distingue pas de la proportion dans la population, alors elle peut alors être considérée comme résultant d'un tirage aléatoire sans remise de  $n_k$  éléments parmi  $n$ . La variance de l'estimateur d'une proportion dans un sondage aléatoire simple est égale à :

$$V(\hat{p}) = \frac{n - n_k}{n - 1} \frac{1}{n_k} \frac{n_j}{n} p(1 - p) = \frac{n - n_k}{n - 1} \frac{1}{n_k} \frac{n_j}{n} \left(1 - \frac{n_j}{n}\right)$$

soit ici :

$$V(\hat{p}) = \frac{n - n_k}{n - 1} \frac{1}{n_k} \frac{n_j}{n} \left(1 - \frac{n_j}{n}\right)$$

La valeur-test est alors égale à :

$$VT = \frac{\frac{n_{jk}}{n_k} - \frac{n_j}{n}}{\sqrt{V(\hat{p})}}$$

Sous l'hypothèse  $H_0$  d'égalité de la proportion dans la population et de la proportion observée dans la classe considérée, cette quantité suit une loi normale centrée-réduite et on examine la probabilité correspondante.

Comme pour les variables continues ce test ne peut en toute rigueur s'effectuer que pour les seules variables supplémentaires. Toutefois l'extension des valeurs-tests aux variables actives permet leur classement, ce qui facilite la caractérisation des classes.

## 6.7. LES AIDES À L'INTERPRÉTATION EN CLASSIFICATION

169

## CARACTERISATION

## PAR LES MODALITES

DES CLASSES OU MODALITES DE : COUPURE 'a' DE L'ARBRE EN 8 CLASSES

V.TEST	PROBA	CLA/	MOD/	GLOBAL	MODALITES		IDEN	POIDS
		MOD	CLA		CARACTERISTIQUES	DES VARIABLES		
				12.29	CLASSE 1 / 8		aa1a	397533
*****	.000	43.24	97.47	27.71	faible	Efficacite brise-vent rapproche	ebr2	896142
937.26	.000	36.15	97.42	33.13	nulle	Efficacite brise-vent eloigne	ebe1	*****
899.67	.000	41.83	88.21	25.93	ni taillis ni futaie	peuplement	pp11	838372
710.80	.000	28.87	88.21	37.57	densite nulle	densite en arbre	den1	*****
710.80	.000	28.87	88.21	37.57	pas de dominantes	dominante en arbre	dom6	*****
695.85	.000	21.79	99.70	56.26	faible efficience	efficience de la futaie	eff1	*****
695.85	.000	21.79	99.70	56.26	nulle	longueur IFN	ifn0	*****
537.26	.000	32.70	57.94	21.79	nulle	continue brise-vent	cbv1	704437
348.89	.000	27.29	39.48	17.79	faible <1/3	continue brise-vent	cbv2	575096
291.22	.000	21.72	45.96	26.01	<3 m	Largeur dominante	lar1	841026
268.38	.000	22.56	37.78	20.59	moyenne 1/3-2/3	strate basse	stb3	665687
250.24	.000	19.27	49.13	31.34	Finistère	departement	d_29	*****
246.45	.000	28.07	21.15	9.26	PRA=29361	Petite Region Agricole	PR12	299468
191.55	.000	15.90	62.48	48.31	faible - moderee	pente amont	faib	*****
176.16	.000	15.97	56.68	43.64	cepees	haie particuliere	c_p	*****
153.32	.000	17.53	33.48	23.48	oui	efficience filtre	fill	759225
110.30	.000	13.79	71.39	63.63	oui	efficience cloture	clo1	*****
107.50	.000	15.67	33.32	26.15	ROUTE	famille	ROUT	845455
105.04	.000	17.03	20.87	15.07	perpendiculaire	situation de la haies sur la pente	perp	487272
96.23	.000	18.42	12.60	8.41	en biais	situation de la haies sur la pente	bia1	271992
95.52	.000	13.86	62.03	55.01	forte >=2/3	strate basse	stb4	*****
89.64	.000	16.17	21.02	15.98	indetermine	fosse utile	inde	516676
82.11	.000	17.48	11.96	8.41	PLEIN CHAMP	famille	PLCH	271973
80.15	.000	15.72	20.68	16.17	denivele	nature du talus	deni	522766
75.89	.000	15.09	25.09	20.44	1-1.5 m	hauteur du talus	ht_3	660923
72.78	.000	13.42	63.14	57.84	bocager	nature du talus	boca	*****
70.22	.000	19.54	5.50	3.46	PRA=29358	Petite Region Agricole	PR10	111834
70.22	.000	16.76	11.22	8.23	>=1.5 m	hauteur du talus	ht_4	266107
66.84	.000	12.70	92.42	89.49	pas d'entretien	entretien des haies	ent0	*****
53.90	.000	13.80	33.93	30.22	parallèle	situation de la haies sur la pente	par1	977122
38.82	.000	13.58	26.15	23.68	NO-SE	ORIENTATION DE LA HAIE	or_1	765545
36.16	.000	13.27	34.21	31.70	0.5-1 m	hauteur du talus	ht_2	*****
34.69	.000	13.44	26.05	23.84	NE-SO	ORIENTATION DE LA HAIE	or_3	770746
30.51	.000	13.00	40.40	38.20	40-80 m	longeur du bord de parcelle	lon2	*****
29.63	.000	13.39	21.64	19.87	80-120 m	longeur du bord de parcelle	lon3	642523
28.47	.000	15.31	3.82	3.07	PRA=56364	Petite Region Agricole	PR25	99232
27.91	.000	15.82	2.83	2.20	PRA=35358	Petite Region Agricole	PR19	71111
26.75	.000	13.70	12.37	11.10	INTERFACE PRE-CHAMP	famille	PRCH	359033
22.71	.000	13.73	8.79	7.87	PRA=22358	Petite Region Agricole	PR_1	254402
21.86	.000	13.64	9.16	8.26	PRA=29363	Petite Region Agricole	PR14	266983
20.03	.000	12.85	31.80	30.43	oui	efficacite cloture	ec11	983862
15.56	.000	13.45	6.39	5.85	fosse en aval	fosse utile	aval	189021
6.79	.000	13.01	3.13	2.96	fosse en amont	fosse utile	amon	95560
-4.22	.000	11.96	4.95	5.09	PRA=29362	Petite Region Agricole	PR13	164600
-6.49	.000	12.00	13.31	13.64	0-0.5 m	hauteur du talus	ht_1	440935
-9.60	.000	11.85	13.15	13.64	INTERMEDIAIRE DIVERS	famille	INTD	441113
-11.07	.000	11.49	5.41	5.79	PRA=56362	Petite Region Agricole	PR23	187220
-14.01	.000	11.24	4.93	5.39	forte	pente amont	fort	174316
-19.87	.000	12.05	68.20	69.56	non	efficacite cloture	ec12	*****
-21.83	.000	11.61	23.76	25.16	E-0	ORIENTATION DE LA HAIE	or_4	813438
-30.59	.000	11.14	16.82	18.57	Morbihan	departement	d_56	600512
-37.87	.000	8.17	1.63	2.46	PRA=29360	Petite Region Agricole	PR11	79544
-39.68	.000	9.82	6.00	7.51	bois perdu	entretien des haies	perd	242784
-50.10	.000	10.82	24.04	27.32	N-S	ORIENTATION DE LA HAIE	or_2	883457
-52.28	.000	9.15	5.86	7.87	PRA=56363	Petite Region Agricole	PR24	254493
-60.46	.000	6.50	1.59	3.00	recolte	entretien des haies	reco	96950
-62.25	.000	9.51	10.53	13.61	PLEIN PRE	famille	PRE	439963
-65.62	.000	6.13	1.53	3.07	PRA=22362	Petite Region Agricole	PR_5	99327
-65.86	.000	10.16	19.41	23.48	<40 m	longeur du bord de parcelle	lon1	759237
-69.29	.000	8.99	8.74	11.95	essence ornementale	haie particuliere	esso	386455
-79.36	.000	4.16	.81	2.39	PRA=35357	Petite Region Agricole	PR18	77339
-81.66	.000	8.10	6.77	10.27	PRA=22359	Petite Region Agricole	PR_2	331945
-83.13	.000	7.46	4.87	8.02	INTERMEDIAIRE LACHE	famille	INTL	259370
-87.89	.000	11.35	69.46	75.21	pas de fosse	fosse utile	fos0	*****
-92.59	.000	5.88	2.63	5.49	PRA=35359	Petite Region Agricole	PR20	177578
-93.37	.000	8.90	13.80	19.06	BATIMENT	famille	BATI	616377
*****	.000	4.30	1.54	4.41	autres dominantes	dominante en arbre	dom5	142535
*****	.000	5.30	2.63	6.09	PRA=35097	Petite Region Agricole	PR15	197062
*****	.000	9.67	28.61	36.36	non	efficience cloture	clo2	*****
*****	.000	7.99	11.49	17.68	haies ajourees	peuplement	pp12	571713
*****	.000	4.99	2.72	6.69	PRA=22360	Petite Region Agricole	PR_3	216320
*****	.000	10.39	49.43	58.46	3-10 m	Largeur dominante	lar2	*****
*****	.000	7.99	14.00	21.54	Ille-&-Vilaine	departement	d_35	696639

*****	.000	8.64	20.05	28.53	Côtes d'Armor	departement	d_22	922649
*****	.000	10.69	66.52	76.51	non	efficience filtre	fil2	*****
*****	.000	7.65	16.18	25.99	pas de talus	hauteur du talus	ht_0	840277
*****	.000	7.65	16.18	25.99	pas de talus	nature du talus	nat0	840277
*****	.000	8.99	31.06	42.48	pas de haies part.	haie particuliere	hp_0	*****
*****	.000	8.66	32.59	46.29	pas de pente	situation de la haies sur la pente	ph_0	*****
*****	.000	8.66	32.59	46.29	pas de pente	pente amont	pt_0	*****
*****	.000	.96	.49	6.30	resineux	dominante en arbre	dom4	203829
*****	.000	.00	.00	4.71	tsf domi. taillis	peuplement	pp14	152375
*****	.000	3.66	4.61	15.52	>=10 m	Largeur dominante	lar3	501722
*****	.000	.46	.30	8.02	futaie moy. dense	peuplement	pp15	259188
*****	.000	3.18	4.48	17.31	futaie	dominante en arbre	dom2	559880
*****	.000	.08	.05	8.06	nulle	strate basse	stb1	260526
*****	.000	.00	.00	7.94	taillis tres dense	peuplement	pp13	256756
*****	.000	.89	.80	11.05	tetards+futaie	dominante en arbre	dom3	357215
*****	.000	4.55	11.79	31.87	densite moyenne	densite en arbre	den2	*****
*****	.000	.00	.00	11.58	tsf domi. futaie	peuplement	pp17	374606
*****	.000	.28	.30	13.18	efficience moyenne	efficience de la futaie	ef12	426080
*****	.000	2.35	4.47	23.35	tetards	dominante en arbre	dom1	754995
*****	.000	.19	.25	16.21	faible	Efficacite brise-vent eloigne	ebe2	524263
*****	.000	.11	.15	16.34	faible <1/3	strate basse	stb2	528345
*****	.000	.04	.05	16.17	nulle	Efficacite brise-vent rapproche	ebr1	523009
*****	.000	1.16	2.33	24.72	moderee	Efficacite brise-vent eloigne	ebe3	799227
*****	.000	.13	.20	19.38	moderee	Efficacite brise-vent rapproche	ebr3	626584
*****	.000	.15	.25	19.79	moyenne 1/3-2/3	continuite brise-vent	cbv3	639978
*****	.000	.06	.10	20.72	>=50 m	longueur IFN	ifn2	670152
*****	.000	.11	.20	23.00	0-50 m	longueur IFN	ifn1	743868
*****	.000	.00	.00	24.13	futaie tres dense	peuplement	pp16	780249
*****	.000	.00	.00	25.93	forte	Efficacite brise-vent eloigne	ebe4	838347
*****	.000	.00	.00	30.55	forte efficience	efficience de la futaie	ef13	987957
*****	.000	.00	.00	30.55	rideau continu	densite en arbre	den3	987957
*****	.000	.76	2.28	36.72	forte	Efficacite brise-vent rapproche	ebr4	*****
*****	.000	.71	2.33	40.63	forte >=2/3	continuite brise-vent	cbv4	*****

### 6.7.4 Les valeurs-tests pour les barycentres

Le calcul des valeurs-tests pour les coordonnées des barycentres des classes sur les axes factoriels est identique à celui des variables continues. Par exemple on voit que la classe 5 est bien représentée sur l'axe 3.

COORDONNEES ET VALEURS-TEST SUR LES AXES 1 A 5													
CLASSES				VALEURS-TEST					COORDONNEES				
IDEN - LIBELLE	EFF.	P.ABS		1	2	3	4	5	1	2	3	4	5
COUPURE 'a' DE L'ARBRE EN 8 CLASSES													
aa1a - CLASSE 1 / 8	2014	397532.90		-38.1	-35.0	24.5	18.0	-6.7	-.80	-.73	.51	.38	-.14
aa2a - CLASSE 2 / 8	1668	329462.20		-66.7	44.0	50.0	-23.9	18.1	-1.55	1.02	1.16	-.56	.42
aa3a - CLASSE 3 / 8	2595	513376.60		-50.9	-38.2	-62.0	10.2	-43.0	-.92	-.69	-1.12	.18	-.77
aa4a - CLASSE 4 / 8	773	152379.30		35.3	-11.6	24.3	11.3	1.2	1.24	-.41	.85	.40	.04
aa5a - CLASSE 5 / 8	1313	258723.90		28.4	-20.3	60.0	31.6	3.7	.75	-.54	1.59	.84	.10
aa6a - CLASSE 6 / 8	1841	363513.50		52.4	8.4	10.6	15.8	-7.5	1.15	.18	.23	.35	-.16
aa7a - CLASSE 7 / 8	2905	574403.80		10.3	-7.1	-42.1	7.4	85.1	.17	-.12	-.71	.12	1.43
aa8a - CLASSE 8 / 8	3251	643928.10		39.2	50.4	-22.9	-52.9	-47.4	.62	.79	-.36	-.83	-.74





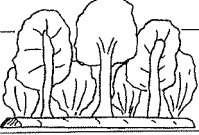
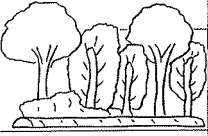
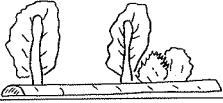

### 6.7.5 Caractérisation de la partition par les variables

La caractérisation globale de la partition par les variables s'opère par un test du  $\chi^2$  sur les tableaux croisant les classes de la partition et les modalités de chaque variable.

26 . departement										
EFFECTIF % EN LIGNE % EN COLONNE	aa1a	aa2a	aa3a	aa4a	aa5a	aa6a	aa7a	aa8a	MARG	
d_22 - Côtes d'Armor	79697 8.64 20.05	104234 11.30 31.64	159590 17.30 31.09	31015 3.36 20.35	86960 9.43 33.61	128771 13.96 35.42	159590 17.30 27.78	172742 18.72 26.83	1922649 100.00 28.53	
d_29 - Finistère	195315 19.27 49.13	102186 10.08 31.02	141958 14.01 27.65	94310 9.31 61.89	125419 12.37 48.48	112227 11.07 30.87	162828 16.07 28.35	79346 7.83 12.32	***** 100.00 31.34	
d_35 - Ille-&-Vilaine	55643 7.99 14.00	62674 9.00 19.02	140616 20.19 27.39	12454 1.79 8.17	21092 3.03 8.15	66290 9.52 18.24	136197 19.55 23.71	201682 28.95 31.32	696638 100.00 21.54	
d_56 - Morbihan	66878 11.14 16.82	60367 10.05 18.32	71218 11.86 13.87	14598 2.43 9.58	25251 4.21 9.76	56224 9.36 15.47	115803 19.28 20.16	190178 31.67 29.53	600511 100.00 18.57	
ENSEMBLE	397532 12.29 100.00	329462 10.19 100.00	513376 15.88 100.00	152379 4.71 100.00	258723 8.00 100.00	363513 11.24 100.00	574403 17.76 100.00	643928 19.91 100.00	***** 100.00 100.00	

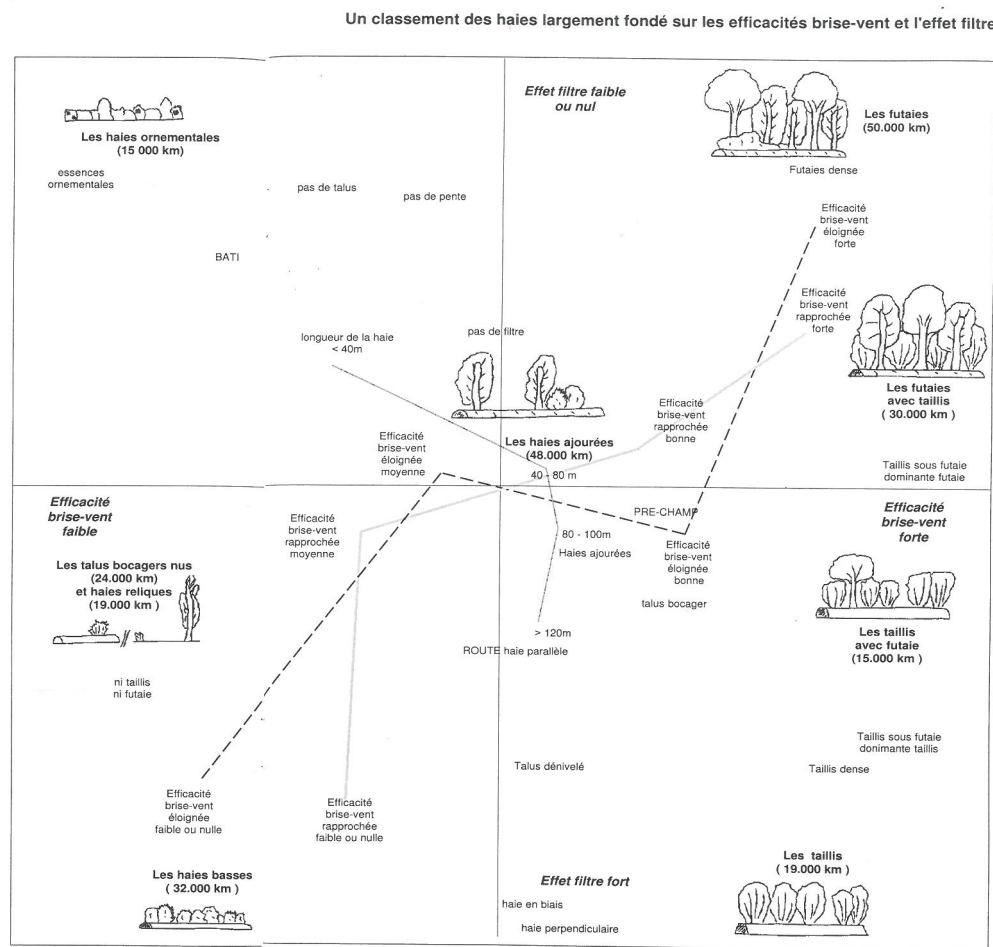
CHI2 =\*\*\*\*\* / 21 DEGRES DE LIBERTE / 0 EFFECTIFS THEORIQUES INFERIEURS A 5  
 PROBA ( CHI2 >\*\*\*\*\* ) = .000 / V.TEST = 605.67

L'utilisation de ces diverses aides à l'interprétation a permis de préciser les huit classes retenues dans l'étude. Ces différentes classes sont illustrées ci-après.

	<b>LES HAIES BASSES</b> (32.000 km)
	<b>LES TALUS BOCAGERS NUS (24.000 km)</b> et <b>HAIES RELIQUES</b> (19.000 km)
	<b>LES TAILLIS</b> (19.000 km)
	<b>LES TAILLIS avec FUTAIE</b> (15.000 km)
	<b>LES FUTAIES AVEC TAILLIS</b> (30.000 km)
	<b>LES FUTAIES</b> (50.000 km)
	<b>LES HAIES AJOUREES</b> (48.000 km)
	<b>LES HAIES ORNEMENTALES</b> (15.000 km)

## 6.7. LES AIDES À L'INTERPRÉTATION EN CLASSIFICATION 173

Sur le plan factoriel suivant formé par les deux premiers axes, la caractérisation de chaque élément de la typologie retenue figure au barycentre de la classe correspondante :



Sur le graphique suivant on observe la répartition géographique des différentes classes. La hauteur des haies épouse un gradient Ouest-Est assez remarquable. Les haies situées à l'Ouest en bordure de la côte finistérienne sont en générale basses et comprennent peu d'arbres de haurt jet. En revanche plus on progresse vers l'Est de la région et plus la hauteur semble augmenter, comme l'indique la sur-représentation en Ile-et-Vilaine des haies de futaies.

#### Les haies arborées

Elles se distinguent selon la nature des strates et comprennent les haies *avec taillis dominant* et les *haies avec futaie dominante*. (L' arbre de futaies est un arbre de haut jet, le plus souvent d'émonde. Le taillis s'en distingue par la pousse à partir de la même souche, de plusieurs rejets qui donnent à terme une cèpée).

#### Les haies avec taillis dominant

Les haies de taillis avec 34 000 km forment 13 % du linéaire total. Il s'agit d'une classe homogène dont la strate arborée est complétée d'une strate basse présente sur plus des 2/3 des bords de parcelle. Cette classe peut être scindée en deux selon la présence éventuelle d'une futaie éparse. Plus de 80 % de ces haies sont associées à un talus, avec forte présence des talus bocagers hauts. Ces haies se rencontrent plutôt sur un terrain en pente, avec une plus grande fréquence des haies en travers de pente. Les taillis s'associent au maillage agricole, plus particulièrement aux pâtures.

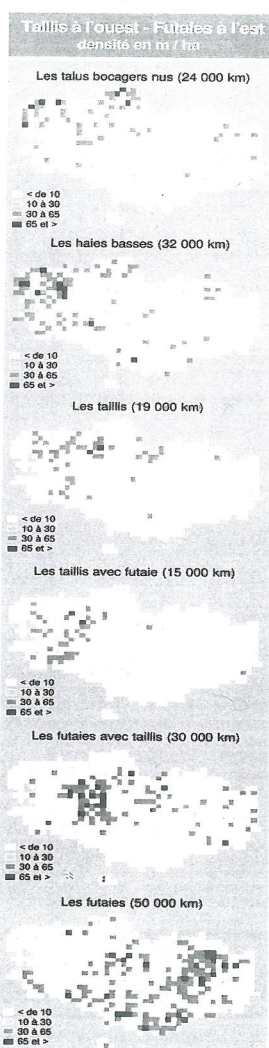
#### Les haies avec futaie dominante

Elles représentent 80 000 km de bords de parcelle, soit près du tiers du linéaire total. Ces haies sont constituées d'arbres de futaie en rideau quasi continu sur 50 000 km et par un mélange futaie-taillis sur 30 000 km. Elles sont souvent associées à un talus qui est plutôt bocager bas. Elles sont plutôt situées en terrain plat ou parallèle à la pente. Ces haies s'associent au maillage agricole, notamment aux pâtures.

#### Les haies ajourées

Elles totalisent 48 000 km, soit 19 % du linéaire total. Un trait dominant de ces haies tient à la discontinuité de la bande de feuillage des 3 à 5 m due à une strate de futaie ou taillis ajourée. L'observation de la strate basse permet de distinguer dans cette classe deux sous-types : les haies ajourées avec strate basse continue (19 000 km) et les haies ajourées avec strate basse absente ou discontinue (29 000 km).

Les haies ajourées sont très souvent associées à un talus, avec prédominance des talus bocagers bas. Elles sont souvent en terrain plat ou parallèle à la pente.



Les haies ajourées se rencontrent en espace agricole et le long des routes.

Si ce type de haie existe en Bretagne depuis longtemps, il est aussi actuellement engendré par des prélèvements non reconstitués dans des rideaux boisés. Ceci explique l'omniprésence de ces haies sur le territoire breton, notamment dans les secteurs où le maillage bocager arboré est dense.

#### Les haies ornementales

Elles constituent un segment bien particulier qui représente, avec 15 000 km, 6 % du linéaire total. Ce groupe se distingue par la présence d'essences ornementales. La strate basse est très souvent continue et l'écran de feuillage entre 3 et 5 m est très faible. Sa position dans le paysage la rend également très caractéristique ainsi que l'absence d'éléments associés, talus et fossé. Dans la majorité des cas elles sont situées en terrain plat et à proximité du bâti. Les haies ornementales se rencontrent en bordure littorale à l'ouest de la région. Elles sont cependant fréquentes dans le bassin de Rennes.

#### Taillis à l'ouest - Futaies à l'est

La répartition géographique des principaux types d'éléments bocagers bretons met en évidence un très net gradient d'ouest en est. Les talus nus et haies basses de l'extrême nord-ouest laissent place aux taillis dominants dans le centre Finistère, puis les mélanges futaie-taillis dans le centre ouest Bretagne et les futaies dans l'ensemble de la Haute Bretagne. Cette opposition entre les haies de taillis à l'ouest et les haies de futaies à l'est recouvre la distribution des peuplements forestiers feuillus : les peuplements forestiers de taillis dominant dans le Finistère et la futaie feuillue est surtout présente en Ile-et-Vilaine.

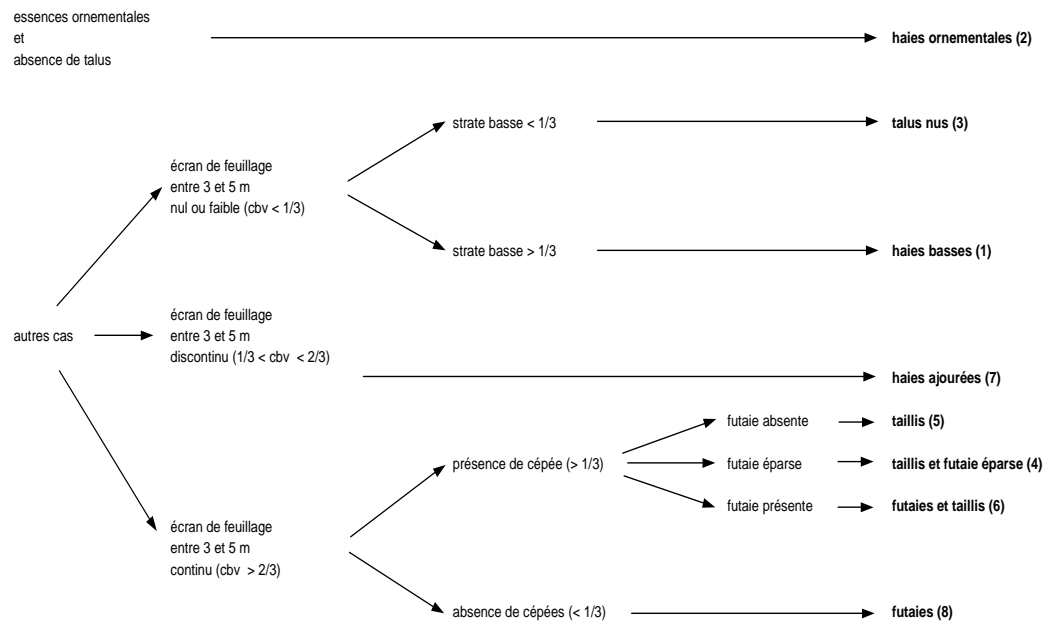


Cette typologie a également permis, outre la réalisation d'une cartographie, la constitution d'une « clé » de détermination des éléments bocagers. Il s'agit là d'un des aspects sans doute les plus intéressants de la CAH. L'une des originalités de l'enquête était de ne pas se fonder sur un classement des haies à dire d'expert mais de recueillir un certain nombre d'indicateurs permettant l'élaboration a posteriori d'une typologie.

Une clé de détermination simplifiée, basée sur des observations rapides de terrain, a été élaborée à partir de la typologie précédente. Cette clé de détermination s'appuie sur cinq variables ou observations :

- prédominance ou non d'essences ornementales,
- continuité brise-vent (remplissage de la bande entre 3 et 5 m),
- présence d'une strate basse sur plus ou moins un tiers du bord de parcelle,
- présence ou non de cépées sur plus d'un tiers du bord de parcelle,
- présence ou non d'arbres de futaie, selon trois modalités :  
absence de futaie,  
futaie éparse (moins de 3 arbres ou arbres à plus de 10 m) futaie présente.

La clé de détermination est présentée dans le tableau ci-dessous.



Afin d'apprécier la pertinence d'un classement des bords de parcelle à partir de cette clé, chacun des 16 360 bords de parcelle a été rangé dans l'une des 8 classes obtenues par la clé. Chaque bord de parcelle est alors affecté d'un numéro de classe « typologie » et un numéro de classe « clé de détermination ». Une parfaite adéquation entre ces deux modes de classement supposerait que le tableau ci-dessous, qui dénombre les 16 360 bords de parcelles selon leurs appartenances, soit parfaitement diagonal.

On observe en réalité que la clé de détermination classe correctement 14 096 bords de parcelle, soit 86 % de la population. On observe également que dans les deux classements, la classe « haie ajourée », intermédiaire entre les haies basses et les haies arborées demeure la classe la moins « étanche », avec des transferts possibles de bords de parcelles vers les haies de futaies, de taillis sous futaie, ou les haies basses. Malgré ces réserves, la clé de détermination simplifiée peut être utile pour la caractérisation de la nature des éléments bocagers dans un territoire breton donné.

Répartition croisée des bords de parcelles selon la typologie et la clé de détermination :

	typ 1	typ 2	typ 3	typ 4	typ 5	typ 6	typ 7	typ 8	total
clé 1	1954	286					663	23	2926
clé 2	7	1291	23				14	53	1388
clé 3	1	6	2545				26	3	2581
clé 4				772		1			773
clé 5					1303				1303
clé 6						1371			1371
clé 7	5	31	23		1303	468	2131	443	3101
clé 8	47	54	4	1	10	1	71	2729	2917
total	2014	1668	2595	773	1313	1841	2905	3251	16360
% capté	97	77	98	100	99	74	73	84	86

## 6.8 Une autre approche dans la caractérisation des classes

Les aides à l'interprétation d'une partition réalisée à partir d'une classification qui viennent d'être présentées sont très liées au logiciel SPAD, notamment la notion de valeur test.

Il existe d'autres éléments qui permettent d'éclairer les partitions, faisant appel à des considérations géométriques.

Afin de déterminer les variables les plus caractéristiques d'une classe on peut rechercher la direction privilégiée dans laquelle une classe s'éloigne du barycentre du nuage. Soient  $g_{jk}$  les coordonnées du barycentre de la classe  $k$  dans  $\mathbf{R}^p$  et  $g_j$  celles du barycentre de l'ensemble.

La quantité  $\|g_j - g\|^2 = \sum_{j=1}^p (g_{jk} - g_j)^2$  constitue un indicateur de l'excentricité de la classe  $k$ . La part prise par la variable  $j$  dans cet éloignement s'écrit :

$$\frac{(g_{jk} - g_j)^2}{\|g_j - g\|^2}$$

Cette quantité permet de détecter quelles sont les variables responsables de l'éloignement de la classe  $k$  du barycentre. Il est également possible de les signer en les affectant du signe de l'écart  $g_{jk} - g_j$ .

Appliquons cette démarche à la classe 28 issue de la classification opérée sur l'exemple traité par l'ACP. Les données doivent être centrées et réduites ou

alors il faut calculer la distance entre  $g_k$  et  $g$  avec la métrique inverse des variances de terme générique  $\frac{1}{\sigma_j^2}$ .

$j$	$g_{jk}$	$g_j$	$\sigma_j$	$Q_{jk} = \frac{(g_{jk} - g_j)^2}{\sigma_k^2}$	$\text{signe}(g_{jk} - g_j) \frac{Q_{jk}}{\rho_k^2}$
NET	17.55	13.91	2.89	1.59	+0.19
INT	3.58	3.14	0.53	0.69	+0.08
SUB	1.52	2.19	0.71	0.89	-0.11
DCT	19.21	22.26	2.33	1.71	-0.21
LMT	9.15	9.00	1.66	0.01	+0.00
IMM	25.59	22.11	3.11	1.25	+0.15
EXP	5.40	10.21	3.82	1.59	-0.19
VRD	18.00	17.18	1.15	0.51	+0.06
				$\Sigma = 8.23 = \rho_k^2$	

Les principales variables responsables de l'éloignement de la classe  $k$  sont NET et IMM d'une part, DCT et EXP de l'autre. Ces variables coïncident avec celles dégagées par l'examen des valeurs-tests.

Ces quantités calculées ici pour des données traitées par l'ACP peuvent l'être en AFC et en ACM.

De la même manière il est possible de déterminer les axes factoriels sur lesquels les différentes classes d'une partition sont le mieux représentées.

Soit  $c_{jk}$  la coordonnée du barycentre de la classe  $k$  sur l'axe factoriel de rang  $j$ . La qualité de la représentation de  $g_k$  sur l'axe  $j$  s'écrit :

$$QLT_j(k) = \frac{c_{jk}^2}{\|g_k - g\|^2}$$

Application :

Examinons sur quel(s) axe(s) la classe 28 issue du premier exemple présenté est le mieux représentée :

$j$	$c_{jk}$	$\frac{c_{jk}^2}{\ g_k - g\ ^2} * 1000$
1	2.86	996
2	-0.11	1
3	-0.08	1
4	-0.11	1
5	0.06	0

## 6.9 Complémentarité entre analyse factorielle et classification

Schématiquement la classification permet de segmenter la population étudiée en groupes homogènes que l'Analyse Factorielle permet d'interpréter et de représenter. En positionnant les groupes à leur barycentre sur les plans factoriels on peut aider à l'interprétation des résultats, notamment en déterminant les axes sur lesquels les classes sont le mieux représentées. De même l'édition des plans factoriels sur lesquels les individus sont remplacés par le numéro de la classe à laquelle ils appartiennent dans la partition permet d'apprécier l'homogénéité des classes constituées.

## 6.10 Compléments

### 6.10.1 Optimisation par réallocation

La partition obtenue après coupure de l'arbre hiérarchique peut être améliorée. On peut calculer pour chaque élément de base la distance au barycentre de chacune des classes et l'affecter au groupe dont il est le plus proche. Celui-ci ne coïncide pas nécessairement avec le groupe d'affectation dans la partition.

CONSOLIDATION DE LA PARTITION AUTOUR DES 8 CENTRES DE CLASSES,  
REALISEE PAR 20 ITERATIONS A CENTRES MOBILES

PROGRESSION DE L'INERTIE INTER-CLASSES

ITERATION	I.TOTALE	I.INTER	QUOTIENT
0	1.880440	.591489	.3145
1	1.880447	.657856	.3498
2	1.880447	.663706	.3530
3	1.880446	.664572	.3534
4	1.880447	.664823	.3535
5	1.880447	.664958	.3536

ARRET APRES L'ITERATION 5 : L'ACCROISSEMENT DE L'INERTIE INTER-CLASSES  
PAR RAPPORT A L'ITERATION PRECEDENTE N'EST QUE DE .020 %.

DECOMPOSITION DE L'INERTIE CALCULEE SUR 20 AXES

		INERTIES		EFFECTIFS		POIDS		DISTANCES	
		AVANT	APRES	AVANT	APRES	AVANT	APRES	AVANT	APRES
INERTIE	INTER-CLASSES	.5915	.6650						
	INERTIES INTRA-CLASSE								
	CLASSE 1 / 8	.1180	.1331	1795	2014	354098.00397532.90		.5022	.4733
	CLASSE 2 / 8	.1128	.1058	1719	1668	339679.50329462.20		.9410	1.0971
	CLASSE 3 / 8	.2428	.2372	2496	2595	493612.20513376.60		.5822	.6767
	CLASSE 4 / 8	.0480	.0484	772	773	152187.50152379.30		1.4979	1.5027
	CLASSE 5 / 8	.0665	.0668	1310	1313	258136.80258723.90		1.1504	1.1475
	CLASSE 6 / 8	.1051	.1246	1683	1841	332321.00363513.50		.7245	.7036
	CLASSE 7 / 8	.2797	.2521	3000	2905	593632.40574403.80		.2682	.3495
	CLASSE 8 / 8	.3161	.2475	3585	3251	709706.70643928.10		.2858	.4205
INERTIE	TOTALE	1.8804	1.8804						
QUOTIENT (INERTIE INTER / INERTIE TOTALE) : AVANT ... .3145									
APRES ... .3536									

Cette procédure a été utilisée ici et elle améliore la partition initiale, le rapport de l'inertie inter-classe à l'inertie totale passant de 0.3145 à 0.3536, soit un gain de 12.75 %. L'inconvénient de cette procédure est que désormais les différentes partitions ne sont plus emboîtées, ainsi la partition en 9 classes ne se décline plus à partir de la partition en 8 classes par décontraction de l'une des classes.

Par ailleurs, les modifications intervenant dans la constitution des classes constituent un indicateur de leur stabilité.

Signalons enfin qu'à l'issue d'une CAH la partition en  $k$  classes obtenue ne constitue pas la meilleure partition possible d'un ensemble à  $n$  éléments en  $k$  classes. Il est toutefois impossible en raison du nombre de partitions en  $k$  classes de déterminer la partition optimale.

### 6.10.2 L'algorithme des voisins réciproques

L'utilisation de la classification ascendante hiérarchique est restée longtemps restreinte à des ensembles modestes de données en raison à la fois de la lourdeur des calculs et de l'encombrement requis. Cette méthode s'est affranchie des ces contraintes à la fois par l'évolution de la puissance des ordinateurs mais aussi par le développement de nouveaux algorithmes, notamment celui dit des « voisins réciproques ». A chaque étape on détermine pour chaque élément du nuage son plus proche voisin, c'est à dire celui dont il est le plus proche. Si le plus proche voisin de  $i$  est  $j$  et si  $i$  est également le plus proche voisin de  $j$ , on dit que  $i$  et  $j$  sont voisins réciproques et ils sont alors directement agrégés. Cette méthode accélère considérablement l'algorithme de Ward dans lequel à chaque étape seulement 2 éléments sont réunis. L'équivalence entre les deux approches a été montrée sous certaines conditions. La convergence de l'algorithme des voisins réciproques s'opère en  $n^2$  alors que

celle de Ward est en  $n^3$ <sup>1</sup>.

## 6.11 Méthodes de partitionnement

### 6.11.1 Introduction

Ces techniques plus frustes que les méthodes hiérarchiques se sont développées tant que les méthodes hiérarchiques compte tenu de la masse des calculs nécessaires restaient confinés aux jeux de données de dimension modeste.

Bien que désormais la puissance de calcul et le développement de nouveaux algorithmes rendent possibles l'utilisation de la CAH sur de grands ensembles de données, les méthodes de partitionnement sont toujours utilisées, notamment en complément des techniques hiérarchiques.

### 6.11.2 Les centres mobiles

Cette méthode constitue l'une des techniques de base. De nombreuses variantes ont été développées. Soient un ensemble de  $n$  éléments sur lesquels  $p$  variables ont été relevées et que l'on désire séparer en  $q$  groupes.

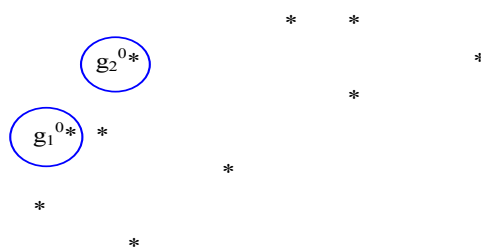
Observons déjà une différence importante avec les méthodes hiérarchiques. Alors qu'avec ces dernières le nombre de groupes n'est pas déterminé *a priori* mais résulte du choix du niveau de coupure dicté par l'examen des niveaux d'agrégation successifs, les techniques de partitionnement requièrent de fixer *a priori* le nombre de classes.

Dans un premier temps  $q$  éléments parmi les  $n$  sont tirés de manière aléatoire ou alors choisis en raison de leur représentativité. Illustrons la démarche de manière graphique à partir de 10 éléments que l'on désire séparer en deux groupes.

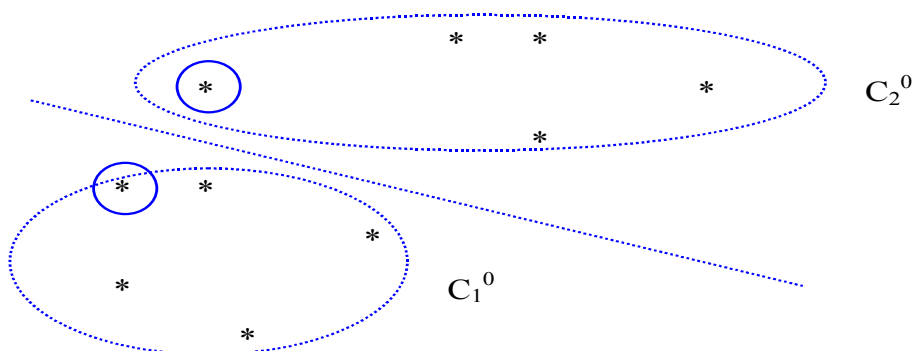
A l'étape 0 deux éléments sont tirés aléatoirement :

---

1. voir par exemple « Classification automatique » p 176 - dans CELEUX - DIDAY - GOVAERT - LECHEVALLIER - RALAMBONDRAINY - Dunod 1989

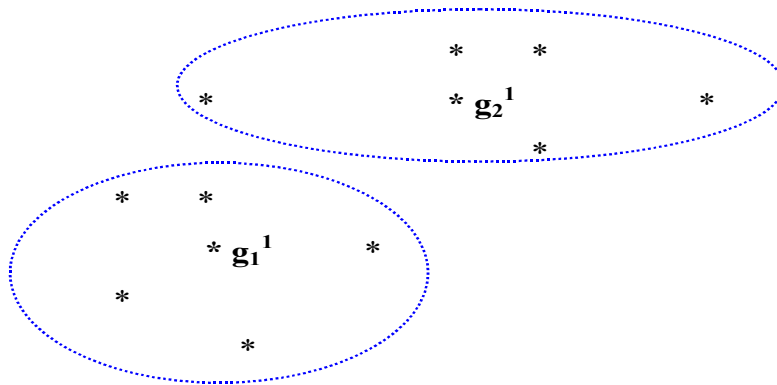


Les distances entre chaque élément et les deux centres initiaux sont calculés puis chaque élément est affecté au centre dont il est le plus proche, ce qui fournit une première partition en deux classes  $C_1^0$  et  $C_2^0$ .

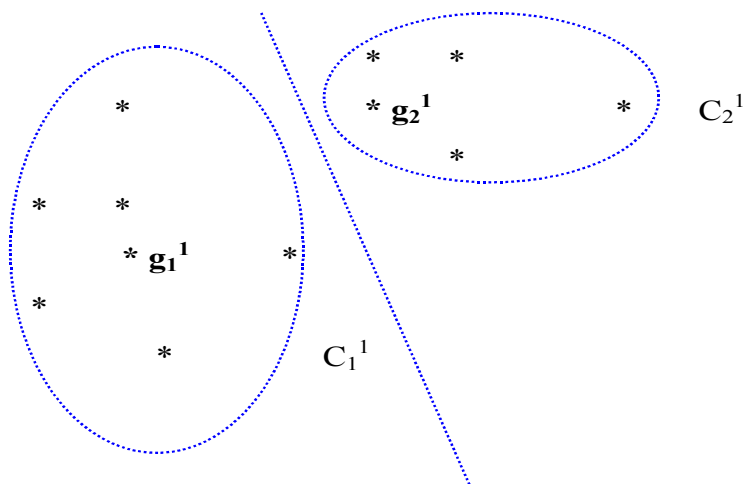


Les centres de gravité des deux classes  $g_1^{(1)}$  et  $g_2^{(1)}$  sont alors calculés et le processus est réitéré (d'où le nom de « centres mobiles », les barycentres se modifiant à chaque itération).

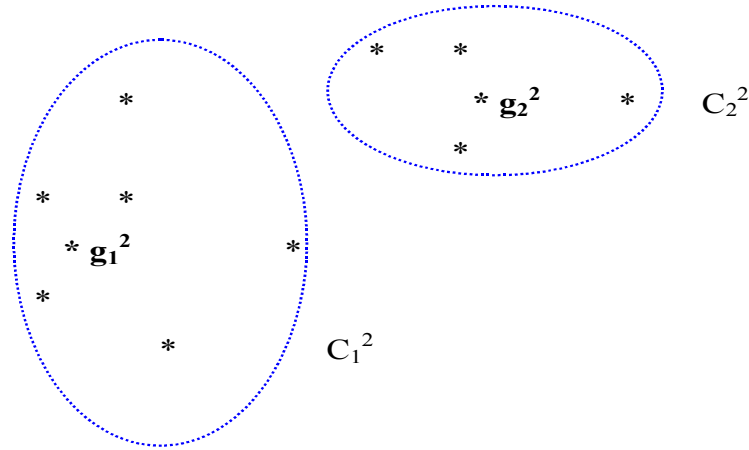




En affectant chaque élément au centre dont il est le plus proche on obtient une seconde partition en deux classes  $C_1^1$  et  $C_2^1$ .



Etape 2 :



Le critère d'arrêt de l'algorithme réside :

- soit dans le nombre d'itérations fixé a priori (10 par exemple).
- soit dans l'identité successive de deux partitions.
- soit dans l'évolution de l'inertie intra-classes dont la diminution devient inférieure à un seuil spécifié a priori entre deux partitions successives.

Ici par exemple, après deux itérations la partition obtenue ne va plus évoluer. Le problème de fond posé par les centres mobiles est leur convergence vers des optima locaux. En d'autres termes on obtient la meilleure partition finale compte tenu de la partition initiale, donc du tirage des centres.

Cet algorithme connaît de nombreuses variantes, par exemple celle des « k-means » de Mac Queen. Dans cette dernière les  $k$  premiers individus fournissent les  $k$  premiers centres. L'élément suivant ( $k + 1$ ) est affecté au centre le plus proche mais on n'attend pas d'avoir procédé à l'allocation de tous les éléments pour calculer les nouveaux centres de classe. On remplace directement le centre le plus proche par le barycentre constitué et on réitère le procédé. La partition est ainsi obtenue en une seule fois.

Démontrons qu'avec la méthode des centres mobiles l'inertie intra-classes ne peut que diminuer entre deux itérations successives.

Considérons un ensemble de  $n$  éléments à classer en  $q$  groupes.

Soit  $C_k^j$  la classe  $k$  à l'itération  $j$  et  $g_{kj}$  le barycentre de cette classe.

A l'étape  $j$  considérons la quantité suivante :

$$Q_j = \sum_{k=1}^q \sum_{i \in C_k^j} m_i d^2(i, g_k^{(j)})$$

A l'étape  $j + 1$  cette quantité s'écrit :

$$Q_{j+1} = \sum_{k=1}^q \sum_{i \in C_k^{j+1}} m_i d^2(i, g_k^{(j+1)})$$

L'inertie intra-classes à l'étape  $j$  s'écrit :

$$I_j = \sum_{k=1}^q \sum_{i \in C_k^j} m_i d^2(i, g_k^{(j+1)})$$

Les  $g_k^{(j+1)}$  sont en effet calculés comme barycentres à partir des classes  $C_k^j$ . D'après le théorème de Huygens la quantité  $Q_j$  peut s'écrire :

$$Q_j = I_j + \sum_{k=1}^q m_k d^2(g_k^j, g_k^{(j+1)})$$

où  $m_k$  désigne la masse de la classe  $k$ .

Par conséquent  $I_j \leq Q_j$ .

Par ailleurs, on sait que les distances  $d^2(i, g_k^{(j+1)})$  sont minimales pour  $i$  dans  $C_k^{j+1}$ , donc :

$$Q_{j+1} \leq I_j$$

d'où :

$$Q_{j+1} \leq I_j \leq Q_j$$

### 6.11.3 Les fortes fortes

Cette approche permet de remédier en partie au problème posé par les centres mobiles dont elle constitue en quelque sorte une généralisation. Elle consiste à lancer plusieurs fois la méthode des centres mobiles en générant de nouveaux centres et à examiner ensuite la « partition produit ». On appelle « formes fortes » les éléments qui dans les partitions successives ont toujours été classées ensemble.

Considérons par exemple un ensemble à  $n$  éléments. L'utilisation des centres mobiles fournit une première partition en 5 classes,  $C_1^1$ ,  $C_2^1$ ,  $C_3^1$ ,  $C_4^1$  et  $C_5^1$ .

L'utilisation de nouveau de la méthode avec tirage d'autres centres initiaux conduit à une nouvelle partition en 5 classes :  $C_1^2$ ,  $C_2^2$ ,  $C_3^2$ ,  $C_4^2$  et  $C_5^2$ .

La partition produit contient alors  $5^2 = 25$  classes.

Le processus peut être poursuivi afin d'obtenir une troisième partition. La partition produit contient alors  $5^3 = 125$  classes dont de nombreuses seront probablement vides.

Les formes fortes seront constituées des éléments qui dans chacune des trois partitions ont été classés ensemble. Seules les classes les plus nombreuses sont ensuite conservées. Les classes vides sont bien entendu éliminées et les classes d'effectif faible peuvent être supprimées en utilisant une procédure de réallocation. De cette manière le nombre de classes n'est plus égal à celui fixé dans chacune des procédures de base (centres mobiles).

#### 6.11.4 La classification mixte

La technique précédente peut être couplée avec une méthode de classification ascendante hiérarchique afin d'accélérer cette dernière.

Dans un premier temps une méthode de partitionnement est utilisée afin de dégager des formes fortes puis la CAH est ensuite utilisée sur les formes fortes. L'idée étant que les méthodes de partitionnement sont plus grossières que les méthodes hiérarchiques mais qu'il est inutile d'utiliser ces dernières sur de grands ensembles de données et qu'il vaut mieux au contraire les réserver sur des ensembles plus stables. La segmentation en huit classes réalisée sur l'enquête Haies utilise cette procédure mixte.

# Bibliographie

- [1] Analyses Factorielles Simples et Multiples - Brigitte ESCOFIER Jérôme PAGES - 4ème édition - Dunod 2008
- [2] Statistique Exploratoire Multidimensionnelle - Ludovic LEBART Alain MORINEAU Marie PIRON - Dunod 1995
- [3] Techniques de la Description Statistique - L. LEBART A. MORINEAU N. TABARD - Dunod 1977
- [4] Analyse des Données - Michel VOLLE - 4ème édition - Economica 1980
- [5] Analyse des Données Evolutives - GERI - Technip 1996



# Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Analyse d'un nuage de points quelconque</b>	<b>7</b>
2.1	Introduction . . . . .	7
2.2	Analyse directe . . . . .	7
2.2.1	Position générale du problème . . . . .	7
2.2.2	La démarche . . . . .	8
2.2.3	Inertie d'un nuage de points . . . . .	8
2.2.4	Recherche d'un axe $\Delta u_1$ conservant au mieux le nuage	9
2.3	Formalisation . . . . .	10
2.3.1	Notations . . . . .	10
2.3.2	Position du problème . . . . .	10
2.3.3	Recherche du second axe . . . . .	11
2.4	Propriétés . . . . .	12
2.4.1	Inertie d'un sous-espace . . . . .	12
2.4.2	Recherche du meilleur sous-espace de dimension $q$ pour représenter le nuage . . . . .	13
2.5	Analyse dans $\mathbf{R}^n$ . . . . .	13
2.5.1	Introduction . . . . .	13
2.5.2	Formalisation . . . . .	14
2.6	Reconstitution des données initiales . . . . .	15
2.7	Méthodologie de l'interprétation . . . . .	17
2.7.1	Sélection du nombre d'axes à analyser . . . . .	17
2.7.2	Examen d'un nuage . . . . .	20
2.8	Qualité de la représentation . . . . .	22
2.9	Élément supplémentaire . . . . .	25
2.10	Analyse d'un nuage de points. Cas général . . . . .	26
2.10.1	Notation . . . . .	26
2.10.2	Définitions . . . . .	26
2.10.3	Analyse directe . . . . .	27
2.10.4	Analyse duale . . . . .	28

2.10.5	Formules de reconstitution des données . . . . .	29
2.10.6	Retour au cas initial . . . . .	31
2.10.7	Formulaire . . . . .	32
2.11	Exemple . . . . .	32
2.11.1	La sélection du nombre d'axes . . . . .	34
2.11.2	Interprétation des deux premiers axes . . . . .	35
2.11.3	Les représentations graphiques . . . . .	38
2.11.4	Conclusion . . . . .	39
<b>3</b>	<b>Analyse en Composantes Principales</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.2	Le choix de la métrique . . . . .	43
3.3	Analyse directe - Formalisation . . . . .	45
3.4	Propriétés . . . . .	46
3.4.1	Matrice d'inertie en ACP normée . . . . .	46
3.4.2	Inertie totale en ACP normée . . . . .	46
3.5	Le nuage des individus . . . . .	46
3.5.1	Composante principale . . . . .	46
3.5.2	Propriétés des composantes principales . . . . .	47
3.5.3	Qualité de la représentation . . . . .	49
3.5.4	Contributions . . . . .	49
3.6	Le nuage des variables . . . . .	49
3.6.1	Coordonnée d'une variable $X_j$ sur l'axe $k$ . . . . .	49
3.6.2	Qualité de la représentation . . . . .	51
3.6.3	Contributions . . . . .	51
3.6.4	Corrélations entre variables . . . . .	51
3.6.5	Une propriété spécifique de l'ACP normée . . . . .	52
3.7	Éléments supplémentaires . . . . .	53
3.7.1	Individus supplémentaires . . . . .	54
3.7.2	Variables supplémentaires . . . . .	54
3.8	Exemple . . . . .	54
3.8.1	Les axes . . . . .	54
3.8.2	Le nuage des individus . . . . .	58
3.8.3	Le nuage des variables . . . . .	58
3.8.4	Les représentations graphiques . . . . .	60
3.9	Compléments . . . . .	63
3.9.1	Variables qualitatives en ACP . . . . .	63
3.9.2	Représentations simultanées en ACP . . . . .	64



<b>4</b>	<b>Analyse Factorielle des Correspondances</b>	<b>65</b>
4.1	Introduction . . . . .	65
4.2	Notation . . . . .	66
4.3	Une métrique spécifique en AFC . . . . .	68
4.4	Résolution . . . . .	71
4.4.1	Analyse dans $\mathbf{R}^p$ . Le nuage des profils-lignes . . . . .	71
4.4.2	Analyse dans $\mathbf{R}^n$ . Le nuage des profils-colonnes . . . . .	72
4.4.3	Interprétation des résultats . . . . .	74
4.5	Propriété fondamentale . . . . .	78
4.5.1	Les relations de transition en AFC . . . . .	78
4.5.2	La représentation simultanée des lignes et des colonnes en AFC . . . . .	80
4.6	Formes classiques des nuages . . . . .	81
4.7	Compléments . . . . .	82
4.8	Lien avec le test du $\chi^2$ . . . . .	87
4.9	AFC et analyse d'un nuage de points quelconque . . . . .	91
4.10	Sélection des axes . . . . .	93
4.11	Utilisation des résultats . . . . .	96
4.12	Extensions . . . . .	98
4.12.1	L'analyse factorielle des correspondances multiples . . . . .	98
4.12.2	L'AFC sur tableaux de notes . . . . .	99
4.12.3	L'AFC appliquée à certains tableaux chronologiques . . . . .	99
<b>5</b>	<b>Analyse des Correspondances Multiples</b>	<b>101</b>
5.1	Introduction . . . . .	101
5.2	L'ACM simple prolongement de l'AFC . . . . .	102
5.2.1	Le traitement des variables en ACM . . . . .	102
5.2.2	Notation . . . . .	102
5.2.3	Tableau de codage disjonctif complet . . . . .	103
5.2.4	Tableau de codage condensé . . . . .	103
5.2.5	Propriétés du tableau de codage disjonctif complet . . . . .	103
5.2.6	Tableau de Burt . . . . .	104
5.2.7	AFC du tableau de codage disjonctif complet . . . . .	104
5.2.8	Résumé . . . . .	106
5.3	Propriétés . . . . .	106
5.3.1	Inertie totale du nuage . . . . .	106
5.3.2	Inertie d'une modalité . . . . .	107
5.3.3	Inertie d'une variable . . . . .	108
5.3.4	Distance entre individus . . . . .	109
5.3.5	Distance entre modalités . . . . .	109
5.3.6	Propriétés des modalités . . . . .	110

5.4	Le codage des variables . . . . .	111
5.4.1	Les variables continues . . . . .	111
5.4.2	Le codage des variables qualitatives . . . . .	111
5.5	Equivalence pour l'acquisition des facteurs . . . . .	112
5.5.1	AFC du tableau $Z$ (Rappel) . . . . .	112
5.5.2	AFC du tableau de Burt $B(= {}^tZZ)$ . . . . .	113
5.5.3	Relation entre les coordonnées des modalités dans l'analyse de $Z$ et de $B$ . . . . .	113
5.6	Un cas particulier . . . . .	114
5.7	Le cas binaire . . . . .	114
5.7.1	AFC du tableau disjonctif complet $Z$ . . . . .	114
5.7.2	AFC du tableau de contingence $C$ . . . . .	115
5.8	Conclusion . . . . .	117
5.9	Exemples . . . . .	117
5.9.1	Premier exemple . . . . .	117
5.9.2	L'enquête sur les haies de Bretagne . . . . .	122
5.10	Compléments . . . . .	130
5.10.1	Valeur-test . . . . .	130
5.10.2	Taux d'inertie en ACM . . . . .	132
<b>6</b>	<b>Méthodes de Classification</b>	<b>137</b>
6.1	Introduction . . . . .	137
6.2	Définitions . . . . .	138
6.2.1	Distance . . . . .	138
6.2.2	Distances usuelles . . . . .	138
6.2.3	Distance ultramétrique . . . . .	139
6.2.4	Hiérarchie . . . . .	140
6.2.5	Définition d'une partition compatible avec une hiérarchie	142
6.2.6	Hiérarchie indicée . . . . .	142
6.2.7	Propriété importante . . . . .	143
6.3	Classification ascendante hiérarchique . . . . .	143
6.3.1	Principe . . . . .	143
6.3.2	Exemple . . . . .	144
6.4	Stratégie d'agrégation selon la variance. La méthode de Ward	150
6.4.1	Notation . . . . .	150
6.4.2	Décomposition de l'inertie . . . . .	151
6.4.3	Critère de Ward . . . . .	152
6.4.4	Exemple . . . . .	156
6.5	Autres stratégies d'agrégation . . . . .	159
6.6	Application . . . . .	159
6.7	Les aides à l'interprétation en classification . . . . .	163

6.7.1	La notion de valeur-test . . . . .	163
6.7.2	Caractérisation de la partition par les variables . . . . .	166
6.7.3	Les aides à l'interprétation pour les variables qualita- tives (ACM) . . . . .	167
6.7.4	Les valeurs-tests pour les barycentres . . . . .	171
6.7.5	Caractérisation de la partition par les variables . . . . .	171
6.8	Une autre approche dans la caractérisation des classes . . . . .	177
6.9	Complémentarité entre analyse factorielle et classification . . . . .	179
6.10	Compléments . . . . .	179
6.10.1	Optimisation par réallocation . . . . .	179
6.10.2	L'algorithme des voisins réciproques . . . . .	180
6.11	Méthodes de partitionnement . . . . .	181
6.11.1	Introduction . . . . .	181
6.11.2	Les centres mobiles . . . . .	181
6.11.3	Les fortes fortes . . . . .	185
6.11.4	La classification mixte . . . . .	186