L'analyse de données

Polycopié de cours ENSIETA - Réf. : 1463

Arnaud MARTIN

Septembre 2004

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

Les statistiques peuvent être vues en fonction de l'objectif fixé ; classiquement les méthodes statistiques sont employées soit pour explorer les données (nommée statistique exploratoire) soit pour prédire un comportement (nommée statistique prédictive ou décisionnelle [Goa03] ou encore inférentielle [Sap90]). La statistique exploratoire s'appuie sur des techniques descriptives et graphiques. Elle est généralement décrite par la statistique descriptive qui regroupe des méthodes exploratoires simples, uni- ou bidimensionnelle (moyenne, moments, quantiles, variance, corrélation, ...) et la statistique exploratoire multidimensionnelle. L'analyse de données s'inscrit dans ce cadre de la statistique exploratoire multidimensionnelle. Nous verrons que des méthodes issues de l'analyse de données peuvent également servir la statistique prédictive (*cf.* chapitre 6).

Les méthodes d'analyse de données ont commencées à être développées dans les années 50 poussées par le développement de l'informatique et du stockage des données qui depuis n'a cessé de croître. L'analyse de données a surtout été développée en France par J.P. Benzécri [Ben80a], [Ben80b] qui a su par l'analyse des correspondances représenter les données de manière simple et interprétable. Il décrit l'analyse de données selon cinq principes, un peu désuets aujourd'hui :

- 1^{er} principe : Statistique n'est pas probabilité.
- $2^{ème}$ principe : Le modèle doit suivre les données et non l'inverse.
- $3^{ème}$ principe : Il convient de traiter simultanément des informations concernant le plus grand nombre possible de dimensions.
- $4^{ème}$ principe : Pour l'analyse des faits complexes et notamment des faits sociaux, l'ordinateur est indispensable.
- $5^{ème}$ principe : Utiliser un ordinateur implique d'abandonner toutes techniques conçues avant l'avènement du calcul automatique.

Ces cinq principes montrent bien l'approche d'une part de la statistique à la différence des probabilités - les modèles doivent coller aux données - et d'autre part de l'analyse de données - il faut traiter le plus grand nombre de données simultanément ce qui implique l'utilisation de l'ordinateur et ainsi l'utilisation de nouvelles techniques adaptées.

L'analyse de données fait toujours l'objet de recherche pour s'adapter à tout type de données et faire face à des considérations de traitements en temps réel en dépit de la

quantité de données toujours plus importante. Les méthodes développées (et l'analyse de données) sont maintenant souvent intégrées avec des méthodes issues de l'informatique et de l'intelligence artificielle (apprentissage numérique et symbolique) dans le *data mining* traduit en français par "fouille de données" ou encore extraction de connaissance à partir de données [HL03].

1.1 Domaines d'application

Aujourd'hui les méthodes d'analyse de données sont employées dans un grand nombre de domaines qu'il est impossible d'énumérer. Actuellement ces méthodes sont beaucoup utilisées en marketing par exemple pour la gestion de la clientèle (pour proposer de nouvelles offres ciblées par exemple). Elles permettent également l'analyse d'enquêtes par exemple par l'interprétation de sondages (où de nombreuses données qualitatives doivent être prises en compte). Nous pouvons également citer la recherche documentaire qui est de plus en plus utile notamment avec internet (la difficulté porte ici sur le type de données textuelles ou autres). Le grand nombre de données en météorologie a été une des première motivation pour le développement des méthodes d'analyse de données. En fait, tout domaine scientifique qui doit gérer de grande quantité de données de type varié ont recours à ces approches (écologie, linguistique, économie, *etc*) ainsi que tout domaine industriel (assurance, banque, téléphonie, *etc*). Ces approches ont également été mis à profit en traitement du signal et des images, où elles sont souvent employées comme prétraitements (qui peuvent être vus comme des filtres). En ingénierie mécanique, elles peuvent aussi permettre d'extraire des informations intéressantes sans avoir recours à des modèles parfois alourdis pour tenir compte de toutes les données.

1.2 Les données

Nous considérons tout d'abord que la *population*[1] peut être décrite par des données de deux types de *caractères* : qualitatif ou quantitatif. Les caractères qualitatifs peuvent être purs (*variables nominales*) *i.e.* que les *modalités* ne possèdent pas de structure d'ordre ou ordonnés (*variables ordinales*) *i.e.* que les *modalités* qualitatives sont ordonnées. Il est aisé de comprendre que les données à caractère qualitatif doivent être adaptées pour les méthodes numériques.

Les méthodes d'analyse de données supposent souvent une organisation des données particulière, naturelle, mais parfois difficile à réaliser selon l'application et les données. Le choix d'un tableau permet une organisation dans le plan de toutes les données et ainsi de traiter simultanément toute l'information. Ainsi la plupart des méthodes nécessitent une organisation des données présentée par le tableau 1.1. Nous verrons au Chapitre 4 que selon les données ce tableau est quelque peu modifié, mais l'idée de tableau reste présente dans toutes les méthodes d'analyse de données.

[1] Les mots en italique sont définis dans le glossaire page 103.

		Variables			
		1 ..... k ..... K			
Individus	1				
	i	.....	x_{ik}	.....	
	I				

Tab. 1.1 – Représentation des données.

Ainsi les observations ou *individus* ou encore *unités statistiques* sont représentés en ligne et sont chacun décrits par des *variables* ou *caractères*. Nous conserverons les notations du tableau 1.1 dans la suite du document. x_{ik} est donc la valeur de la variable k pour l'individu i avec $k = 1, ..., K$ et $i = 1, ..., I$. Par abus de notations, pour des considérations de simplification de celles-ci, I représente à la fois le nombre d'individus et l'ensemble des individus $\{1, ..., i, ..., I\}$, de même K représente le nombre de variables et l'ensemble des variables $\{1, ..., k, ..., K\}$.

Cette représentation des données peut faciliter la lecture de petits tableau, *i.e.* lorsqu'il y a peu de données. Cependant, dès lors que la taille du tableau est grand, ou que nous recherchons des relations entre plus de deux individus ou plus de deux variables, cette représentation et les techniques simples de la statistique descriptive ne suffisent plus.

1.3 Les objectifs

Les objectifs que se sont fixés les chercheurs en analyse de données sont donc de répondre aux problèmes posés par des tableaux de grandes dimensions. Les objectifs sont souvent présentés en fonction du type de méthodes, ainsi deux objectifs ressortent : la visualisation des données dans le meilleur espace réduit et le regroupement dans tout l'espace.

Les méthodes de l'analyse de données doivent donc permettre de représenter synthétiquement de vastes ensembles numériques pour faciliter l'opérateur dans ses décisions. En fait d'ensembles numériques, les méthodes d'analyse de données se proposent également de traiter des données qualitatives, ce qui en fait des méthodes capables de considérer un grand nombre de problèmes. Les représentations recherchées sont bien souvent des représentations graphiques, comme il est difficile de visualiser des points dans des espaces de dimensions supérieures à deux, nous chercherons à représenter ces points dans des plans. Ces méthodes ne se limitent pas à une représentation des données, ou du moins pour la rendre plus aisée, elles cherchent les *ressemblances* entre les individus et les *liaisons* entre les variables. Ces proximités entre individus et variables vont permettre à l'opérateur de

déterminer une *typologie* des individus et des variables, et ainsi il pourra interpréter ses données et fournir une synthèse des résultats des analyses. Nous voyons donc que les deux objectifs précédemment cités sont très liés voir indissociables, ce qui entraîne souvent l'utilisation conjointe de plusieurs méthodes d'analyse de données.

1.4 Les méthodes

L'analyse de données regroupe deux familles de méthodes suivant les deux objectifs cités précédemment :

- Une partie des méthodes cherche à représenter de grands ensembles de données par peu de variables *i.e.* recherche les dimensions pertinentes de ces données. Les variables ainsi déterminées permettent une représentation synthétique recherchée. Parmi ces méthodes de nombreuses analyses sont issues de l'analyse factorielle, telles que l'analyse en composantes principales, l'analyse factorielle des correspondances, l'analyse factorielle des correspondances multiples, ou encore l'analyse canonique.
  L'*analyse en composantes principales* est l'une des méthodes les plus employées. Elle est particulièrement adaptée aux variables quantitatives, continues, *a priori* corrélées entre elles. Une fois les données projetées dans différents plans, les proximités entre variables s'interprètent en termes de corrélations, tandis que les proximités entre individus s'interprètent en termes de similitudes globales des valeurs observées. L'*analyse factorielle des correspondances* (ou *analyse des correspondances binaires*) a été conçue pour l'étude des tableaux de contingence obtenus par croisement de variables qualitatives. Cette analyse permet donc de traiter des variables qualitatives et est surtout adaptée à ce type de variables. Dans cette approche, les lignes et les colonnes ont un rôle symétrique et s'interprètent de la même façon. L'*analyse factorielle des correspondances multiples* est une extension de l'analyse factorielle des correspondances qui ne permet que le croisement de deux variables qualitatives. Elle est donc adaptée à la description de grands tableaux de variables qualitatives par exemple pour le traitement d'enquêtes. L'*analyse canonique* est très peu utilisée en pratique, son intérêt porte sur son aspect théorique. Elle cherche à analyser les relations entre deux groupes de variables de nature différente. De ce fait l'analyse factorielle des correspondances peut être vu comme analyse canonique particulière [CDG$^+$89], [LMP95].
- Une autre partie des méthodes cherche à classer les données de manière automatique. Ces méthodes sont complémentaires avec les précédentes pour synthétiser et analyser les données et répondre plus particulièrement à l'objectif fixé de caractériser les proximités entre individus et celles entre variables. Ces méthodes de classification sont soit à apprentissage supervisé (*i.e.* qui nécessitent une base de données d'apprentissage - ces méthodes sont appelées en statistique les analyses discriminantes) soit à apprentissage non-supervisée (*i.e.* qui ne nécessitent aucune donnée préalable).
  - Parmi les méthodes issues de l'analyse discriminante et directement rattachées à

l'analyse de données il y a l'analyse linéaire discriminante, la régression logistique, les k plus proches voisins ou encore les arbres de décision. D'autres méthodes issues de l'intelligence artificielle et du monde de la reconnaissance des formes peuvent être rattachées à l'analyse discriminante telles que le perceptron multicouche (et les autres réseaux de neurones) et les chaînes de Markov [Kun00] ou encore issues de la théorie de l'apprentissage statistique telle que les machines à vecteurs de supports [Vap99]. Si ces dernières ne sont pas toujours considérées comme faisant partie de l'analyse de données, elles sont parfaitement intégrées dans le *data mining*.

L'*analyse linéaire discriminante* est aussi appelée analyse factorielle discriminante car elle est en fait une analyse en composantes principales supervisée. Elle décrit les individus en classes (celles-ci sont données par une variable issue de l'apprentissage) et ensuite affecte de nouveaux individus dans ces classes. C'est donc une méthode à la fois descriptive et prédictive. Elle permet de traiter aussi bien des variables quantitatives que qualitatives.

La *régression logistique* consiste à exprimer les probabilités *a posteriori* d'appartenance à une classe $p(C/\mathbf{x})$ comme une fonction de l'observation [Sap90] [Cel03]. Bien souvent c'est la régression linéaire qui est employée, *i.e.* qu'il faut déterminer les coefficients β tels que :

$$\ln\left(\frac{p(C/\mathbf{x})}{1-p(C/\mathbf{x})}\right) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i. \tag{1.1}$$

L'approche des k *plus proches voisins* repose sur l'idée simple d'attribuer un nouvel individu à la classe majoritaire parmi ses k plus proches voisins (individus de la base d'apprentissage les plus proches au sens d'une certaine distance).

Les *arbres de décision* nécessitent souvent une construction délicate et difficilement généralisable si les données d'apprentissage sont peu représentatives de la réalité. La méthode CART (*Classification And Regression Tree*) possède une construction d'arbre aux propriétés intéressantes pour la segmentation [BFRS93].

– Les méthodes de classification automatique ne nécessitant pas d'apprentissage offrent un intérêt important lorsque les données sont complètement inconnues. Elles permettent ainsi de dégager des classes qui ne sont pas évidentes *a priori*. Les deux principales méthodes développées sont la méthode des centres mobiles (apparentée à la méthode des *k-means* ou des nuées dynamiques (comme un cas particulier)) et la classification hiérarchique ascendante ou descendante. Nous pouvons également citer les approches fondées sur les graphes et hypergraphes [Ber72].

La méthode des *centres mobiles* consiste à associer les individus à des centres de classes choisis aléatoirement, puis à recalculer ces centres jusqu'à obtenir une convergence. La difficulté consiste dans un choix astucieux des centres au départ pour une convergence plus rapide et dans le choix d'une distance appropriée.

La *classification hiérarchique ascendante* (resp. descendante) consiste à regrouper

les individus selon leur ressemblance (resp. dissemblance). Toute la difficulté est dans la définition d'une mesure de ressemblance et de la distance associée.

1.5 Les logiciels

Les méthodes d'analyse de données nées de la recherche universitaire sont depuis longtemps entrées dans le monde industriel. Il y a cependant peu de logiciels qui savent intégrer ces méthodes pour une recherche exploratoire aisée dans les données. Nous citons ici cinq logiciels :

- SAS :
  Ce logiciel est un logiciel de statistique très complet et très performant. Il a d'abord été développé pour l'environnement Unix, mais est maintenant accessible sous tout environnement. Il permet une puissance de calcul importante et ainsi est très bien adapté à tous traitements statistiques sur des données très volumineuses. Son manque de convivialité et surtout son prix fait qu'il est encore peu employé dans les entreprises qui ne se dédient pas complètement à la statistique. De nombreux cours universitaires de statistique sont proposés avec ce logiciel qui s'approche d'un langage (ex. Université de Rennes 1).
- Splus :
  Splus est à la fois un langage statistique et graphique interactif interprété et orienté objet. C'est donc à la fois un logiciel statistique et un langage de programmation. La particularité de ce langage est qu'il permet de mélanger des commandes peu évoluées à des commandes très évoluées. Il a été développé par Statistical Sciences autour du langage S, conçu par les *Bell Laboratories*. Depuis, Splus est devenu propriété de Mathsoft après le rachat de Statistical Sciences. Il est parfois employé pour l'enseignement (ex. Université Paul Sabatier de Toulouse III).
- R :
  Ce logiciel est la version gratuite de Splus. Il est téléchargeable sous www.r-project.org pour tous systèmes d'exploitation. Il souffre également de peu de convivialité et semble encore très peu employé en industrie. De part sa gratuité, il est de plus en plus employé pour la réalisation de cours de statistiques (ex. Université Paul Sabatier de Toulouse III, Université de Lyon 1).
- XlStat :
  Excel propose une macro payante permettant d'effectuer quelques méthodes d'analyse de données. Elle est cependant très limitée, utilisable qu'avec Excel sous Windows et de plus payante. Certaines écoles d'ingénieurs s'en contentent (ex. ENITAB, Bordeaux).
- UniWin Plus :
  Statgraphics est un logiciel de statistiques générales, qui propose un module d'analyse de données de treize méthodes. Développé uniquement pour les environnements Windows, l'accent est porté sur les interfaces graphiques. Statgraphics propose un grand nombre d'analyses statistiques et permet l'utilisation de beaucoup de for-

mat de données. Il est commercialisé par Sigma Plus. Statgraphics est enseigné par exemple à l'IUT de Vannes.

- Stalab :
  Ce logiciel développé par M. Jambu [Jam99b], [Jam99a] était initialement prévu pour Windows. Sa convivialité a permis un essor industriel qui semble s'être réduit. Il a été utilisé pour l'enseignement en écoles d'ingénieurs (ex. ENSSAT, Lannion).
- SPAD :
  Le logiciel SPAD supporté entre autre par A. Morineau est toujours maintenu à jour avec de nouvelles méthodes issues de la recherche universitaire. Sa version sous Windows est conviviale ce qui a poussé son achat par de plus en plus d'industriels. Le soucis de coller à une réalité industrielle fait qu'il est employé en enseignement (ex. IUT de Vannes, ENSIETA).

1.6 Plan

Ce document ne cherche pas à présenter l'ensemble des méthodes de l'analyse de données dont certaines ont été évoquées dans la section 1.4. Nous présentons ici les idées des principales méthodes, ces clés et les références[2] données permettront au lecteur d'approfondir les méthodes présentées et de comprendre les autres.

Nous commencerons ainsi par l'étude de quelques analyses factorielles. Le premier chapitre présente le principe général des analyses factorielles. Les chapitres 3, 4 et 5 présentent respectivement l'analyse en composantes principales, l'analyse factorielle des correspondances et l'analyse des correspondances multiples. Nous proposons ensuite au chapitre 6 l'étude d'une analyse discriminante : l'analyse factorielle discriminante qui peut également être vue comme une analyse factorielle. Dans le cadre des méthodes de classification non-supervisée nous présentons la classification hiérarchique au chapitre 7.

[2]Les références proposées ne sont pas exhaustives, il existe un grand nombre d'ouvrages de qualité dans le domaine de l'analyse de données.

Chapitre 2

Analyses Factorielles

2.1 Introduction

Les analyses factorielles constituent la plupart des analyses de données. Elles sont fondées sur un principe unique, c'est pour cela que nous pouvons parler de l'analyse factorielle [EP90]. Ce principe repose sur le fait que les deux nuages de points représentant respectivement les lignes et les colonnes du tableau étudié (tableau 1.1) sont construits et représentés sur des graphiques. Ces représentations des lignes et des colonnes fortement liées entre elles permettent une analyse plus aisée pour l'opérateur.

2.1.1 Les objectifs

Les analyses factorielles tentent de répondre à la question : tenant compte des ressemblances des individus et des liaisons entre variables, est-il possible de résumer toutes les données par un nombre restreint de valeurs sans perte d'information importante ? En effet en cherchant à réduire le nombre de variables décrivant les données, la quantité d'information ne peut être que réduite, au mieux maintenue. La motivation de cette réduction du nombre de valeurs vient du fait que des valeurs peu nombreuses sont plus faciles à représenter géométriquement et graphiquement (un des objectifs de l'analyse de données).

2.1.2 Domaines d'application

L'ensembles des méthodes d'analyses factorielles permettent de répondre à la plupart des problèmes posés par les applications auxquelles se consacre l'analyse de données. Le choix d'une analyse par rapport à une autre se fera en fonction du type de données (quantitatif, qualitatif, mais aussi textuelle) et de la quantité de données. Il est bien sûr possible lorsque le cas se présente d'appliquer une analyse sur les données quantitatives de la population puis une autre analyse sur les données qualitatives. Ainsi dans le cadre d'un enquête par exemple, une analyse en composantes principales peut faire ressortir les

liaisons entre les variables quantitatives, puis une analyse des correspondances multiples peut donner une représentation des variables qualitatives en fonction de leur *modalités*.

2.1.3 Les données

Dans ce chapitre, nous retenons la représentation des données sous forme de tableau (tableau 1.1, page 3), et les notations associées.

2.2 Principe général

Le principe général de l'analyse factorielle est fondé sur une double hypothèse. Supposons qu'il existe un vecteur colonne $\mathbf{u}_1$ à K composantes et un vecteur colonne $\mathbf{v}_1$ à I composantes tel que le tableau $X = \left\{x_i^k\right\}$ s'écrive $X = \mathbf{v}_1\mathbf{u}_1^t$, où $\mathbf{u}_1^t$ est le vecteur transposé de $\mathbf{u}_1$. Ainsi des $I+K$ valeurs des vecteurs $\mathbf{u}_1$ et $\mathbf{v}_1$, les $I.K$ valeurs de X sont retrouvées. Cette réduction devient vite intéressante dès lors que I et K sont assez grands. De plus elle n'entraîne aucune perte d'information. Cette hypothèse est malheureusement improbable en pratique.

Exemple 2.2.1 Considérons l'ensemble des notes des élèves de l'ENSIETA durant une année. Le nombre d'élèves est environ de 450, et nous pouvons considérer qu'ils obtiennent environ 30 notes chacun. Ainsi le tableau représentant l'ensemble des notes est constitué de 13 500 valeurs. La réduction présentée ci-dessus permet de réduire ce nombre à 480 valeurs sans perte d'information si l'hypothèse est valide. Pour que l'hypothèse soit vérifiée, il faudrait pouvoir déduire les notes de l'ensemble des élèves à partir de celles d'un seul élève et d'un vecteur de pondération. Ceci signifie que les notes sont dépendantes les unes des autres ou encore très fortement corrélées.

En pratique, il faut donc chercher une approximation de rang S pour X. C'est-à-dire ces analyses cherchent à écrire le tableau X tel que :

$$X = \mathbf{v}_1\mathbf{u}_1^t + \mathbf{v}_2\mathbf{u}_2^t + \ldots + \mathbf{v}_S\mathbf{u}_S^t + E, \tag{2.1}$$

où E est une matrice de I lignes et K colonnes de termes négligeables dite matrice *résiduelle*. Ainsi les $I.K$ valeurs initiales de X sont reconstituées de façon satisfaisante par les $S.(I+K)$ valeurs des S vecteurs $\mathbf{v}_q$ et $\mathbf{u}_q$. Les données sont donc soit considérées en tant qu'individus décrits par leurs K variables à l'aide des vecteurs $\mathbf{u}_q$ à K composantes, soit en tant que variables décrites par les I individus à l'aide des vecteurs $\mathbf{v}_q$ à I composantes.

La résolution de ce problème passe donc par la considération des deux nuages de points ou encore des deux représentations géométriques associées (figure 2.1). Nous obtenons ainsi I points dans l'espace $\mathbb{R}^K$ et K points dans l'espace $\mathbb{R}^I$. Plusieurs formes de nuages sont remarquables aussi bien pour les projections de l'espace des individus que pour celui des variables (figure 2.2). Par exemple, nous pouvons distinguer des formes sphériques ne traduisant aucune direction privilégiée, des formes allongées donnant une

direction privilégiée des dépendances, ou encore plusieurs sous-nuages faisant ainsi apparaître plusieurs sous-groupes de la population. D'autres formes sont remarquables telles que les formes triangulaires ou paraboliques [LMP95]. Le problème est de pouvoir rendre compte visuellement de la forme des nuages, pour ce faire l'idée est d'étudier les projections sur des droites ou mieux des plans (les projections dans un espace à 3 dimensions seraient intéressantes si l'œil humain n'était pas souvent trompé). Il faut donc chercher le sous-espace qui ajuste au mieux le nuage de points *i.e.* chercher à minimiser les déformations que la projection implique.

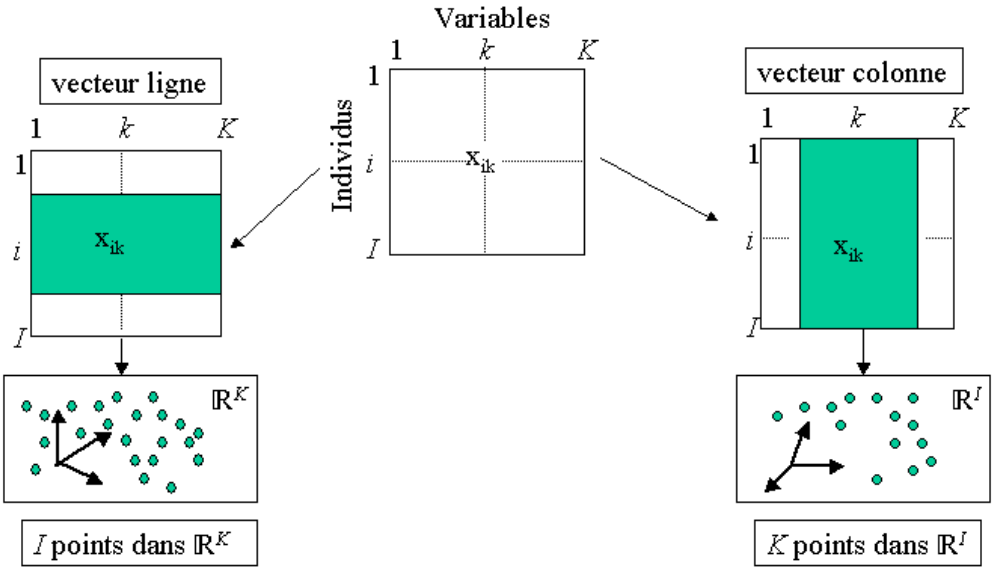


FIG. 2.1 – Les nuages de points.

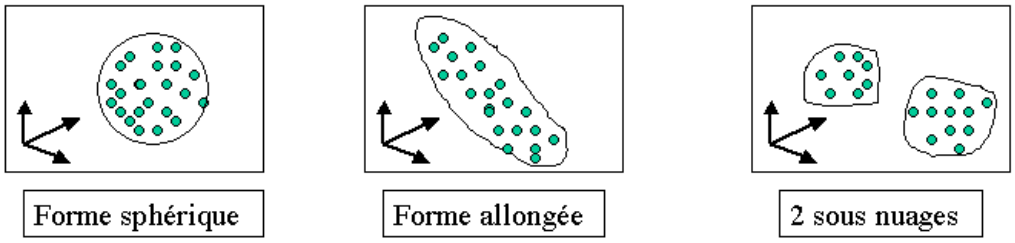


FIG. 2.2 – Les formes de nuages de points.

Nous allons donc chercher à ajuster au mieux le nuage des individus dans l'espace des variables (section 2.3) puis le nuage des variables dans l'espace des individus (section 2.4).

2.3 Ajustement du nuage des individus dans l'espace des variables

L'objectif est de fournir des images approchées du nuage des individus - que nous noterons N_I - dans $\mathbb{R}^K$. Nous considérons pour la visualisation des images planes de N_I. Nous faisons l'hypothèse que le nuage N_I est contenu dans un sous-espace vectoriel de dimension S inférieure à K, *i.e.* que nous supposons que la matrice E de l'équation (2.1) est nulle. Plus généralement, nous supposons que le nuage N_I est reconstitué de manière satisfaisante dans un sous-espace de dimension S. Nous pouvons ainsi reconstruire les I individus, et donc l'ensemble de la population et du tableau X associé à partir des coordonnées des individus sur S nouveaux axes. Les $I.K$ valeurs du tableau X sont donc remplacées par $I.S$ (coordonnées)+$K.S$ (composantes).

Exemple 2.3.1 Si nous considérons 1000 élèves qui obtiennent chacun 100 notes, et si $S = 10$, *i.e.* si les 1000 points-individus sont contenus dans un sous-espace de dimension 10, nous réduisons l'étude des $1000 \times 100 = 10^5$ valeurs de X à $1000 \times 10 + 100 \times 10 = 11000$ valeurs.

2.3.1 Droite d'ajustement

Dans un premier temps, cherchons un sous-espace vectoriel à une dimension, *i.e.* une droite d_1 passant par l'origine, qui ajuste au mieux le nuage N_I. Nous considérons donc le cas où $S = 1$. La projection sur la droite d_1 qui ajuste au mieux le nuage N_I donne la *dispersion* ou *inertie* maximale le long de la droite d_1.

Proposition 2.3.2 *Maximiser la dispersion le long de la droite d_1 revient à minimiser les distances des points du nuage N_I à la droite d_1, c'est-à-dire que la droite d_1 passe au plus près de tous les points du nuage N_I.*

Preuve En effet, en prenant les notations de la figure 2.3, maximiser la dispersion le long de d_1 revient à maximiser la somme $\sum_{i \in I} OH_i^2$, or par le théorème de Pythagore :

$$\sum_{i \in I} Oi^2 = \sum_{i \in I} OH_i^2 + \sum_{i \in I} iH_i^2, \tag{2.2}$$

le deuxième terme représentant les distances des points i de N_I à la droite d_1. □

Proposition 2.3.3 *Maximiser la dispersion le long de la droite d_1 revient à maximiser $\mathbf{u}_1^t X^t X \mathbf{u}_1$, avec $\mathbf{u}_1$ le vecteur unitaire de d_1. En fait, nous avons l'égalité :*

$$\sum_{i \in I} OH_i^2 = (X\mathbf{u}_1)^t (X\mathbf{u}_1) = \mathbf{u}_1^t X^t X \mathbf{u}_1, \tag{2.3}$$

qui représente l'inertie le long de l'axe d_1.

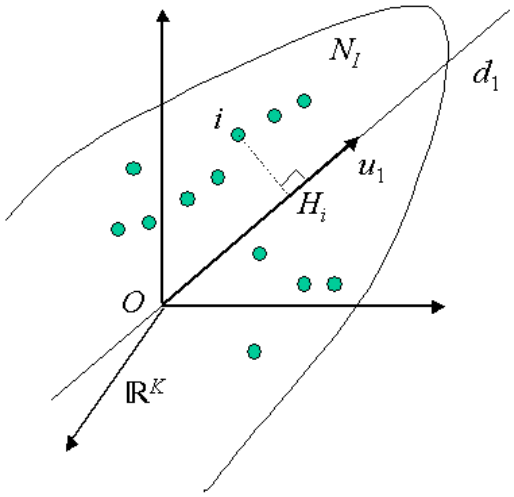


FIG. 2.3 – Le nuage N_I et sa droite d'ajustement.

Preuve La projection OH_i de Oi sur le sous-espace à une dimension d_1 porté par $\mathbf{u}_1$ est le produit scalaire de Oi par $\mathbf{u}_1$ ($OH_i =< Oi, \mathbf{u}_1 >$). Ainsi en munissant cet espace de la métrique euclidienne sans restreindre le problème :

$$OH_i = \mathbf{x}_i^t \mathbf{u}_1 = \sum_{k \in K} x_{ik} u_{1j}. \tag{2.4}$$

Les I composantes OH_i sont donc les I composantes de la matrice $X\mathbf{u}_1$, et donc :

$$\sum_{i \in I} OH_i^2 = (X\mathbf{u}_1)^t (X\mathbf{u}_1). \tag{2.5}$$

Nous avons ainsi démontré la proposition. □

Le problème revient donc à trouver u_1 qui maximise la forme quadratique $\mathbf{u}_1^t X^t X \mathbf{u}_1$ avec la contrainte $\mathbf{u}_1^t \mathbf{u}_1 = 1$. Le sous-espace à une dimension optimal au sens de l'inertie maximale est donc l'axe d_1 défini par le vecteur $\mathbf{u}_1$ solution de ce problème.

2.3.2 Plan d'ajustement

Cherchons maintenant à déterminer le sous-espace à deux dimensions s'ajustant au mieux au nuage N_I, nous considérons donc le cas où $S = 2$.

Proposition 2.3.4 *Le sous-espace à deux dimensions qui ajuste au mieux le nuage N_I contient $\mathbf{u}_1$.*

Preuve En effet, par un raisonnement par l'absurde, si ce sous-espace ne contient pas $\mathbf{u}_1$, alors il est définit par deux vecteurs $\mathbf{u}'$ et $\mathbf{u}''$ différents de $\mathbf{u}_1$. L'inertie le long des deux droites portées par $\mathbf{u}'$ et $\mathbf{u}''$ est donc inférieure à celle de l'inertie le long de la droite portée par $\mathbf{u}_1$. Il existe donc un sous-espace de dimension deux meilleur que celui défini par les deux vecteurs $\mathbf{u}'$ et $\mathbf{u}''$. Nous montrons ainsi la proposition. □

Le sous-espace à deux dimensions est donc caractérisé par l'axe d_1 et l'axe d_2 défini par le vecteur $\mathbf{u}_2$ orthogonal à $\mathbf{u}_1$ vérifiant donc :
- $\mathbf{u}_2^t X^t X \mathbf{u}_2$ est maximal,
- $\mathbf{u}_2^t \mathbf{u}_2 = 1$ (contrainte de normalité),
- $\mathbf{u}_2^t \mathbf{u}_1 = 0$ (contrainte d'orthogonalité).

2.3.3 Sous-espace d'ajustement

Dans le cas où $S \geq 2$, par récurrence, le sous-espace à S dimensions s'ajustant au mieux au nuage N_I contient les vecteurs $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{S-1}$. Ce sous-espace est engendré par le sous-espace $(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{S-1})$ de dimension $S-1$ et le vecteur $\mathbf{u}_S$ orthogonal à ce sous-espace (*i.e.* à tous les $\mathbf{u}_q$) et vérifiant :
- $\mathbf{u}_S^t X^t X \mathbf{u}_S$ est maximal,
- $\mathbf{u}_S^t \mathbf{u}_S = 1$.

Proposition 2.3.5 *Une base orthonormée du sous-espace vectoriel de dimension S, s'ajustant au mieux au sens des moindres carrés, au nuage N_I dans $\mathbb{R}^K$ est constituée par les S vecteurs propres $(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_S)$ correspondant aux S plus grandes valeurs propres $(\mu_1, \mu_2, \cdots, \mu_S)$ de la matrice $X^t X$.*

Remarque S est au plus égal au rang de la matrice $X^t X$, et dans le cas de l'égalité la matrice E de l'équation (2.1) est nulle.

Preuve Cette proposition peut se démontrer par la méthode de Lagrange, une autre approche est fondée sur certaines propriétés spectrales des matrices symétriques [LMP95]. Soit $L(\mathbf{u}_S)$ le Lagrangien :

$$L(\mathbf{u}_S) = \mathbf{u}_S^t X^t X \mathbf{u}_S - \lambda(\mathbf{u}_S^t \mathbf{u}_S - 1), \tag{2.6}$$

où λ est un multiplicateur de Lagrange *i.e.* une constante. Le maximum du Lagrangien est atteint lorsque la dérivée s'annule, *i.e.* :

$$\frac{\partial L}{\mathbf{u}_S} = 2X^t X \mathbf{u}_S - 2\lambda \mathbf{u}_S = 0. \tag{2.7}$$

Ainsi nous obtenons l'égalité $X^t X \mathbf{u}_S = \lambda \mathbf{u}_S$. Or d'après Lagrange, une condition nécessaire et suffisante pour que $f(\mathbf{u}_S) = \mathbf{u}_S^t X^t X \mathbf{u}_S$ soit extremum sachant que $g(\mathbf{u}_S) =$

$\mathbf{u}_S^t\mathbf{u}_S - 1 = 0$ (vérifiée par la contrainte de normalité), est qu'il existe un nombre λ tel que la dérivée du Lagrangien soit nulle. Le maximum est atteint si λ est la plus grande valeur propre de la matrice X^tX.

$\mathbf{u}_S$ est donc le vecteur propre correspondant à la plus grande valeur propre de la matrice X^tX et $\mathbf{u}_S^tX^tX\mathbf{u}_S = \lambda\mathbf{u}_S^t\mathbf{u}_S = \lambda$ est l'inertie projetée sur l'axe d_S. □

2.4 Ajustement du nuage des variables dans l'espace des individus

De la même façon que pour le nuage des individus N_I, nous cherchons une image du nuage des variables - que nous noterons N_K - dans $\mathbb{R}^I$. L'approche est identique à celle du nuage des individus, il suffit simplement de considérer X^t au lieu de X. Avec les notations de la figure 2.4, l'inertie le long de la droite D_S s'écrit $(X^t\mathbf{v}_S)(X^t\mathbf{v}_s) = \mathbf{v}_S^tXX^t\mathbf{v}_S$. Ainsi, l'axe factoriel (ou axe d'inertie) est déterminé par $\mathbf{v}_S$ vérifiant :

- $\mathbf{v}_S^tXX^t\mathbf{v}_S$ est maximal,
- $\mathbf{v}_S^t\mathbf{v}_S = 1$ (contrainte de normalité),
- $\mathbf{v}_S^t\mathbf{v}_q = 0$ pour tout $q = \{1, \cdots, S-1\}$ (contrainte d'orthogonalité).

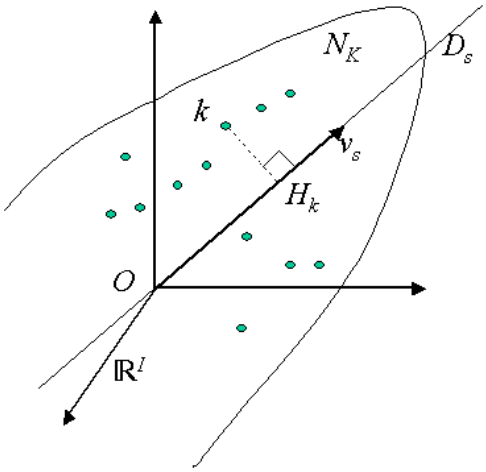


FIG. 2.4 – Le nuage N_I et sa droite d'ajustement.

Le sous-espace d'ajustement est obtenu de la même manière que dans le cas des individus, par la proposition suivante.

Proposition 2.4.1 *Une base orthonormée du sous-espace vectoriel de dimension S, s'ajustant au mieux au sens des moindres carrés, au nuage N_I dans $\mathbb{R}^I$ est constituée par les S vecteurs propres $(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_S)$ correspondant aux S plus grandes valeurs propres $(\nu_1, \nu_2, \cdots, \nu_S)$ de la matrice XX^t.*

Remarque S est au plus égal au rang de la matrice XX^t, qui est égal au rang de la matrice X^tX.

2.5 Relation entre les axes d'inertie et les facteurs des deux nuages

Nous montrons ici quelles sont les relations, dites relations de transition, entre les ajustements dans les deux espaces.

Notons :

- $\lambda_{d_S} = \mathbf{u}_S^t X^t X \mathbf{u}_S$, respectivement $\lambda_{D_S} = \mathbf{v}_S^t X X^t \mathbf{v}_S$ l'inertie le long de l'axe d_S, respectivement D_S.
- $F_S = X\mathbf{u}_S$, respectivement $G_S = X^t\mathbf{v}_S$ le *facteur* d'ordre S de N_I, respectivement de N_K. F_S est donc le vecteur issu de la projection du nuage N_I sur le $S^{\text{ème}}$ axe dans $\mathbb{R}^K$, de même G_S est le vecteur issu de la projection du nuage N_K sur le $S^{\text{ème}}$ axe dans $\mathbb{R}^I$.

Proposition 2.5.1 *L'inertie le long de l'axe d_S, λ_{d_S}, est égale à l'inertie le long de l'axe D_S, λ_{D_S}, nous la notons λ_S.*

Les formules de transition entre les deux espaces $\mathbb{R}^K$ et $\mathbb{R}^I$ sont données par les relations de fondamentales :

$$\begin{cases} \mathbf{v}_S = \dfrac{F_S}{\sqrt{\lambda_S}} \\ \\ \mathbf{u}_S = \dfrac{G_S}{\sqrt{\lambda_S}} \end{cases} \tag{2.8}$$

Preuve Par définition, nous avons dans l'espace $\mathbb{R}^K$:

$$X^tX\mathbf{u}_S = \mu_S\mathbf{u}_S, \tag{2.9}$$

et dans l'espace $\mathbb{R}^I$:

$$XX^t\mathbf{v}_S = \nu_S\mathbf{v}_S. \tag{2.10}$$

En multipliant par X dans l'équation (2.9), nous obtenons :

$$(XX^t)X\mathbf{u}_S = \mu_S(X\mathbf{u}_S), \tag{2.11}$$

et en multipliant par X^t dans l'équation (2.10), nous obtenons :

$$(X^tX)X^t\mathbf{v}_S = \nu_S(X^t\mathbf{v}_S). \tag{2.12}$$

Considérons dans un premier temps le cas où $S = 1$. ν_1 est par définition la plus grande valeur propre de XX^t. L'équation (2.11) pour $S = 1$ montre que $X\mathbf{u}_1$ est un vecteur propre

de XX^t, donc la valeur propre associée μ_1 est nécessairement telle que $\mu_1 \leq \nu_1$. De pus μ_1 est la plus grande valeur propre de X^tX. L'équation (2.12) montre que $X^t\mathbf{v}_1$ est un vecteur propre de X^tX, donc la valeur propre associée ν_1 est nécessairement telle que $\nu_1 \leq \mu_1$. Ainsi nous obtenons que $\mu_1 = \nu_1$.

De même, nous pouvons montrer que toutes les valeurs propres non nulles de X^tX et XX^t sont les mêmes, ainsi $\mu_S = \nu_S$. Le premier point de la proposition est donc démontré, puisque :

$$\mathbf{u}_S^t X^t X \mathbf{u}_S = \mathbf{v}_S^t X X^t \mathbf{v}_S = \lambda_S. \tag{2.13}$$

Pour démontrer le second point, nous constatons à partir de l'équation (2.11) que les facteurs F_S et les vecteurs unitaires $\mathbf{v}_S$ sont les vecteurs propres de la matrice XX^t, nous avons donc :

$$\mathbf{v}_S = \frac{F_S}{\|F_S\|}. \tag{2.14}$$

De plus $\|F_S\|^2 = \mathbf{u}_S^t X^t X \mathbf{u}_S = \lambda_S$. Nous montrons ainsi la première égalité de l'équation (2.8). La seconde égalité se montre de même en constatant que :

$$\mathbf{u}_S = \frac{G_S}{\|G_S\|}, \tag{2.15}$$

et $\|G_S\|^2 = \mathbf{v}_S^t X X^t \mathbf{v}_S = \lambda_S$. □

Les relations de transition entre les deux espaces peuvent se représenter par le schéma de dualité de la figure 2.5 représentant les relations entre les axes d'inertie d'un nuage et les facteurs de l'autre nuage.

Proposition 2.5.2 *Les relations de transitions entre les facteurs s'écrivent :*

$$\begin{cases} F_S(i) = \displaystyle\sum_{k\in K} x_{ik}\mathbf{u}_S(k) = \frac{\displaystyle\sum_{k\in K} x_{ik}G_S(k)}{\sqrt{\lambda_S}} \\ \\ G_S(k) = \displaystyle\sum_{i\in I} x_{ik}\mathbf{v}_S(i) = \frac{\displaystyle\sum_{i\in I} x_{ik}F_S(i)}{\sqrt{\lambda_S}} \end{cases} \tag{2.16}$$

Cette proposition montre que les facteurs des deux nuages doivent s'interpréter conjointement. L'analyse factorielle consiste donc à analyser simultanément le nuage N_I et le nuage N_K.

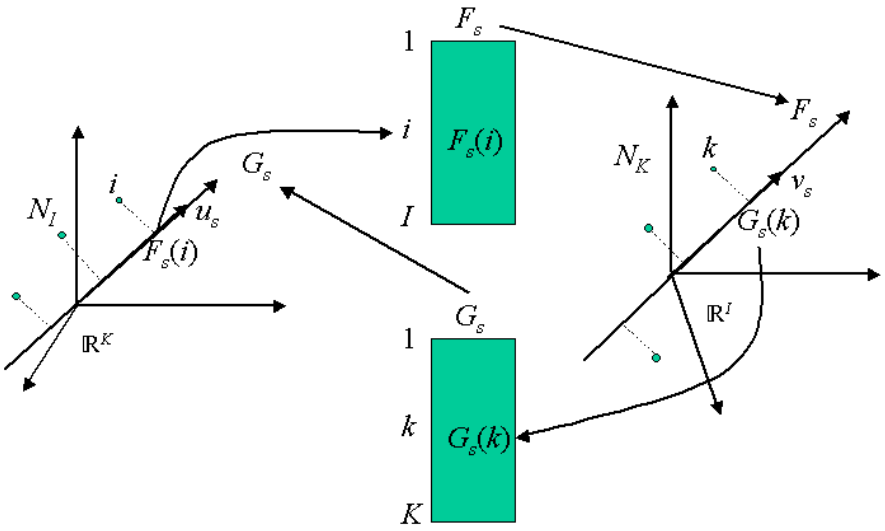


FIG. 2.5 – Schéma de dualité.

Preuve D'après les relations fondamentales de l'équation (2.8) nous obtenons les coordonnées de chaque point du nuage N_I sur les axes factoriels :

$$\mathbf{v}_S(i) = \frac{F_S(i)}{\sqrt{\lambda_S}} = \frac{\sum_{k\in K} x_{ik}\mathbf{u}_S(k)}{\sqrt{\lambda_S}}, \tag{2.17}$$

et les coordonnées de chaque point du nuage N_K sur les axes factoriels :

$$\mathbf{u}_S(k) = \frac{G_S(i)}{\sqrt{\lambda_S}} = \frac{\sum_{i\in I} x_{ik}\mathbf{v}_S(i)}{\sqrt{\lambda_S}}. \tag{2.18}$$

En développant les relations des équations (2.17) et (2.18), nous obtenons facilement les relations de transition de l'équation (2.16). □

2.6 Reconstruction des données

Il est possible de reconstruire de manière exacte le tableau de données X par une décomposition en valeurs singulières de la matrice X. En effet, puisque $\mathbf{u}_s$ est le $s^{\text{ème}}$ vecteur propre de norme 1 de la matrice X^tX, correspondant à la valeur propre λ_s et $\mathbf{v}_s$ est le $s^{\text{ème}}$ vecteur propre de norme 1 de la matrice XX^t, correspondant à la même valeur propre, nous avons :

$$X\mathbf{u}_s = \sqrt{\lambda_s}\mathbf{v}_s, \tag{2.19}$$

d'où

$$X \sum_{s \in K} \mathbf{u}_s \mathbf{u}_s^t = \sum_{s \in K} \sqrt{\lambda_s} \mathbf{v}_s \mathbf{u}_s^t. \tag{2.20}$$

Les vecteurs propres étant orthogonaux et de norme 1, nous obtenons :

$$X = \sum_{s \in K} \sqrt{\lambda_s} \mathbf{v}_s \mathbf{u}_s^t. \tag{2.21}$$

Cette formule de reconstruction du tableau X par décompositions en valeurs singulières à partir des valeurs propres λ_s (qui sont aussi les inerties), et des vecteurs propres associés $\mathbf{u}_s$ et $\mathbf{v}_s$ peut s'illustrer par la figure 2.6.

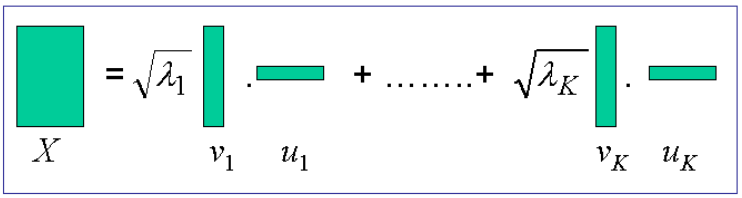


FIG. 2.6 – Décomposition en valeurs singulières du tableau X.

Cette reconstruction exacte suppose donc avoir $I.K$ valeurs contenues dans les K vecteurs propres $\mathbf{u}_s$ et $\mathbf{v}_s$. Nous avons vu dans la section 2.3.3 que nous cherchons le sous-espace qui s'ajuste au mieux aux nuages de points. S'ajuster au mieux signifie donc reconstituer au mieux les positions des points des nuages par un nouvel ensemble de coordonnées.

Premier plan d'ajustement Si λ_1 associée à $\mathbf{u}_1$ est grande par rapport aux autres valeurs propres, alors nous disons que la "reconstruction est bonne". D'un point de vue géométrique ceci signifie que le nuage de points s'allonge le long d'une droite. Lorsque le nuage est ainsi très étiré le long du premier axe, l'inertie du nuage de départ et la position des points sont bien reconstituée avec la seule information des coordonnées des projections des points du nuage.

S premiers axes d'ajustement Un repère formé par les S premiers axes factoriels permet de reconstituer les positions de départ avec une bonne précision, si $\lambda_1 + \cdots + \lambda_S$ représente une "bonne proportion" de la trace de la matrice $X^t X$. En effet, rappelons que $tr(X^t X) = \sum_{s \in K} \lambda_s = \sum_{i \in I, k \in K} x_{ik}^2$.

Nous obtenons ainsi une reconstruction approchée X^* du tableau X en se limitant aux S premiers axes factoriels, nous avons :

$$X \simeq X^* = \sum_{s=1}^{S} \sqrt{\lambda_s} \mathbf{v}_s \mathbf{u}_s^t. \tag{2.22}$$

Nous passons donc des $I.K$ valeurs du tableau X à $S(I+K)$ nombres pour reconstituer X. Ces nombres sont constitués par les S vecteurs $\sqrt{\lambda_s}\mathbf{v}_s$ ayant I composantes et les S vecteurs $\mathbf{u}_s$ ayant K composantes.

Toute la difficulté réside dans le choix de S, c'est-à-dire à partir de quelle valeur a-t-on une bonne reconstruction, ou encore une bonne proportion de la trace de X^tX ? Nous voyons donc l'importance de définir un indice de qualité de la reconstruction. La qualité globale de la reconstruction peut être mesurée par :

$$\tau_S = \frac{tr(X^{*t}X^*)}{tr(X^tX)} = \frac{\sum_{s=1}^{S} \lambda_s}{\sum_{s \in K} \lambda_s}. \tag{2.23}$$

Le coefficient τ_S est encore appelé *taux d'inertie* ou *pourcentage de la variance* relatif aux S premiers facteurs.

2.7 Conclusion

Nous avons présenté dans ce chapitre le principe général des analyses factorielles. Cette approche permet de représenter géométriquement de grands tableaux de données dans des sous-espaces sans perte d'information importante. La dimension des sous-espaces, *i.e.* l'approximation de la reconstruction de ces tableaux se fait en cherchant à minimiser la perte d'information. La quantité globale de reconstruction permet de quantifier cette perte d'information. Une fois la dimension du sous-espace choisie, les données sont représentées graphiquement par des projections sur les différents plans qui constituent le sous-espace. Bien sûr les premiers plans factoriels sont ceux contenant le plus d'information.

La décomposition en valeurs singulières présentée dans ce chapitre peut être appliquée à tous tableaux de données présentés comme sur le tableau 1.1. Cette décomposition fait appel à des distances euclidiennes, c'est-à-dire à des formes quadratiques définies positives. Les maximisations de l'inertie pour les ajustements des sous-espaces sont liées à ces distances. Il existe d'autres approches qui modifient ces distances ou la nature des sous-espaces [EP90], [LMP95]. En particulier ce qui est souvent recherché dans ces méthodes est la non-linéarité des projections, mieux adaptée aux données compliquées.

Avant d'appliquer cette approche générale à un tableau quelconque, il est important de tenir compte des données de départ. Pour se faire, nous allons les transformer en fonction de leur type. Ainsi dans les trois prochains chapitres nous allons voir comment transformer

des données quantitatives dans le cadre de l'analyse en composantes principales, et des données qualitatives dans les cas de l'analyse factorielle de correspondances et de celle des correspondances multiples.

Chapitre 3

Analyse en Composantes Principales

3.1 Introduction

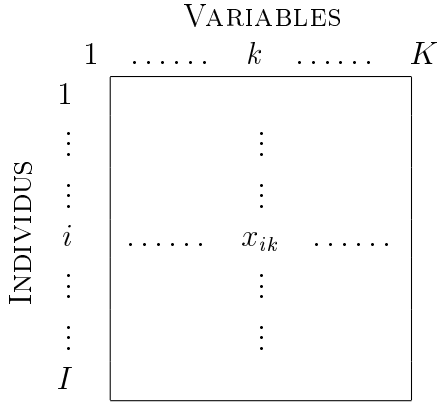
L'analyse en composantes principales - que nous notons par la suite ACP - est une des premières analyses factorielles, et certainement aujourd'hui l'une des plus employées. Dans [LMP95], nous trouvons l'historique de cette méthode qui fut conçue par Karl Pearson en 1901. Elle est sans doute à la base de la compréhension actuelle des analyses factorielles. Son utilisation a cependant été plus tardive avec l'essor des capacités de calculs.

Les principales variantes de l'ACP viennent des différences de transformations du tableau de données. Ainsi, le nuage de points peut être centré ou non, réduit ou non. Le cas le plus étudié, et que nous présentons ici, est lorsque le nuage de point est centré et réduit ; dans ce cas nous parlons d'ACP normée. D'autres variantes existent telle que l'analyse en composante curviligne [DH97] pour remédier au fait que les projections sont linéaires, ou encore l'analyse en composantes indépendantes pour la séparation de sources [Pha96].

Les données Les données pour l'ACP sont généralement présentées sous la forme du tableau précédemment vu dans le Chapitre 1 et que nous rappelons dans le tableau 3.1.

Ainsi les données sont constituées d'individus et de variables qui dans le cas de l'ACP doivent être quantitatives, continues, elles peuvent être homogènes ou non et sont *a priori* corrélées entre elles. Rappelons que nous notons x_{ik} la valeur de la variable k pour l'individu i, I désigne à la fois le nombre d'individus et l'ensemble des indices $I = \{1, \cdots, i, \cdots, I\}$, et K désigne à la fois le nombre d'individus et l'ensemble des indices $K = \{1, \cdots, k, \cdots, K\}$.

Les objectifs Les objectifs de l'ACP sont ceux d'une analyse factorielle, c'est-à-dire qu'elle cherche à représenter graphiquement les relations entre individus par l'évaluation de leurs ressemblances, ainsi que les relations entre variables par l'évaluation de leurs liaisons. Comme nous l'avons vu au chapitre précédent l'étude doit se faire simultanément. Le but final de ces représentations est l'interprétation par une analyse des résultats.



TAB. 3.1 – Représentation des données pour l'ACP.

Les domaines d'application De part la nature des données que l'ACP peut traiter, les applications sont très nombreuses. Il y a en fait deux façons d'utiliser l'ACP :

- soit pour l'étude d'une population donnée en cherchant à déterminer la typologie des individus et des variables. Par exemple, dans la biométrie, l'étude des mensurations sur certains organes peut faire apparaître des caractéristiques liées à des pathologies, ou encore en économie, l'étude des dépenses des exploitations par l'ACP peut permettre des économies de gestion.
- soit pour réduire les dimensions des données sans perte importante d'information, par exemple en traitement du signal et des images, où l'ACP intervient souvent en prétraitement pour réduire la quantité de données issues de traitements analogiques.

3.2 Principe de l'ACP

3.2.1 Les objectifs

Dans un premier temps reprenons les objectifs de l'ACP et détaillons-les. Nous avons vu que pour atteindre les objectifs de l'ACP il faut évaluer les *ressemblances* entre individus ainsi que les *liaisons* entre variables. Ces deux notions peuvent être interprétées de différentes façons, il est donc important de bien les définir.

Définition 3.2.1 *Deux individus se ressemblent, ou sont proches, s'ils possèdent des valeurs proches pour l'ensemble des variables.*

Cette définition sous entend une notion de proximité qui se traduit par une distance. Ainsi, nous définissons la distance entre deux individus i et j par :

$$d^2(i,j) = \sum_{k \in K} (x_{ik} - x_{jk})^2. \tag{3.1}$$

La métrique ici utilisée est donc euclidienne, mais de manière plus générale nous pouvons définir cette distance par :

$$d^2(i,j) = (\mathbf{x}_i - \mathbf{x}_j)^t M (\mathbf{x}_i - \mathbf{x}_j), \tag{3.2}$$

où M est une matrice symétrique définie positive de taille K.

Pour établir un bilan des ressemblances entre individus, nous cherchons à répondre à des questions du type :

- Quels sont les individus qui se ressemblent ?
- Quelles sont ceux qui sont différents ?
- Existe-t-il des groupes homogènes d'individus ?
- Est-il possible de mettre en évidence une typologie des individus ?

De la même façon que nous avons défini la *ressemblance* entre individus, il est essentiel de définir la *liaison* entre des variables.

Définition 3.2.2 *Deux variables sont liées si elles ont un fort coefficient de corrélation linéaire.*

Le coefficient de corrélation linéaire est donné par :

$$r(k,h) = \frac{\mathrm{cov}(k,h)}{\sqrt{\mathrm{var}(k)\,\mathrm{var}(h)}} = \frac{1}{I}\sum_{i\in I}\left(\frac{x_{ik}-\overline{x}_k}{s_k}\right)\left(\frac{x_{ih}-\overline{x}_h}{s_h}\right), \tag{3.3}$$

où $\overline{x}_k$ et s_k sont respectivement la moyenne et l'écart-type de la variable k.

Remarque Dans le cadre de l'ACP normée que nous présentons ici, le coefficient de corrélation est défini à partir de la covariance, cependant dans de rare cas l'ACP peut être fondée sur la matrice de covariance (ACP non-normée) ou encore sur la matrice des coefficients de corrélations des rangs. A partir du coefficient de corrélation de l'équation (3.3), il est possible de définir une distance entre deux variables k et h :

$$d(k,h) = \frac{1}{I}\sum_{i\in I}\left(\frac{x_{ik}-\overline{x}_k}{s_k} - \frac{x_{ih}-\overline{x}_h}{s_h}\right) = 2(1-r(k,h)). \tag{3.4}$$

De même que pour les individus, nous cherchons à établir un bilan des liaisons entre variables en répondant à des questions du type :

- Quelles sont les variables qui sont liées positivement entre elles ?
- Quelles sont celles qui s'opposent (*i.e.* liées négativement) ?
- Existe-t-il des groupes de variables corrélées entre elles ?
- Est-il possible de mettre en évidence une typologie des variables ?

Pondération Il est souvent souhaitable que les individus comme les variables jouent le même rôle. Cependant, dans certaines applications il peut être intéressant de pondérer différemment chaque individu. Soit p_i le poids affecté à chaque individu, par commodité ces poids sont pris tels que la masse totale soit égale à 1 ($\sum_{i\in I} p_i = 1$). Ainsi la moyenne de la variable k est définie par :

$$\overline{x}_k = \sum_{i\in I} p_i x_{ik}, \tag{3.5}$$

		Variables			
		1 ..... k			K
Individus	1				
	i	.....	$\frac{x_{ik} - \overline{x}_k}{s_k}$		.....
	I				

TAB. 3.2 – Représentation des données centrée-réduites pour l'ACP.

et le coefficient de corrélation devient :

$$r(k,h) = \sum_{i \in I} p_i \left(\frac{x_{ik} - \overline{x}_k}{s_k} \right) \left(\frac{x_{ih} - \overline{x}_h}{s_h} \right). \tag{3.6}$$

Nous retrouvons le cas particulier dans lequel les individus ont le même poids lorsque $p_i = \frac{1}{I}$.

De même, il est possible de ne pas accorder la même importance aux différentes variables. Soit m_k le poids associé à la variable k, la distance de l'équation (3.1) entre deux individus i et j devient :

$$d^2(i,j) = \sum_{k \in K} m_k (x_{ik} - x_{jk})^2. \tag{3.7}$$

En fait, ces poids ne modifient en rien les principes de l'ACP, nous considérons donc par la suite les cas où les individus et variables ont le même poids.

3.2.2 La transformation des données

Il existe plusieurs transformations utilisées. L'analyse centrée consiste à modifier les données du tableau X en remplaçant les valeurs des x_{ik} par $x_{ik} - \overline{x}_k$. Le fait de centrer les données présente dans le cas de l'ACP des propriétés intéressantes que nous exposons à la section 3.2.3. L'analyse centrée réduite ou encore normée, que nous présentons ici, est liée à la transformation des données du tableau X en remplaçant les valeurs des x_{ik} par $\frac{x_{ik} - \overline{x}_k}{s_k}$. Réduire les données permet d'uniformiser les unités de mesures. Par exemple, dans le cas d'une analyse sur la mensuration d'animaux, les dimensions dans le tableau X peuvent être exprimées en m ou en cm selon les variables. Ainsi le tableau de données X présenté sur le tableau 3.1 devient celui donné par le tableau 3.2.

3.2.3 L'analyse des nuages

Analyse du nuage des individus

Pour l'analyse du nuage N_I des individus, nous considérons donc le tableau 3.2 des données centrées réduites par ligne, comme nous l'avons déjà vu dans le cas général d'une analyse factorielle (*cf.* figure 2.1 page 11).

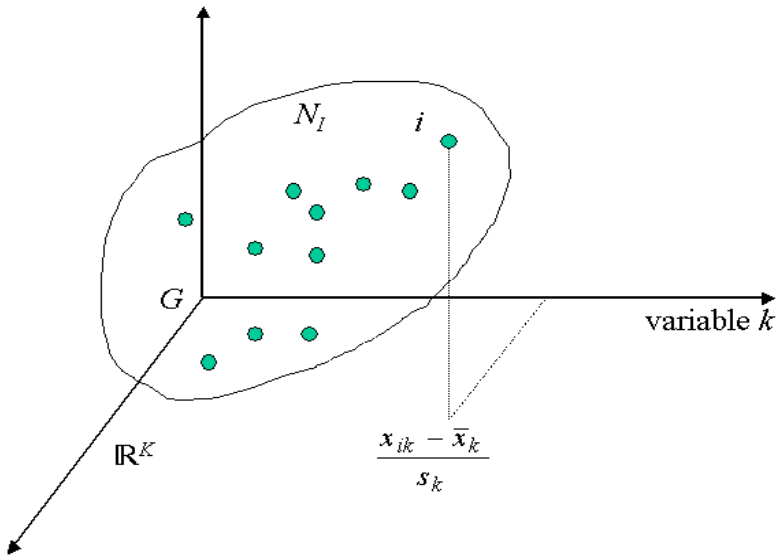


FIG. 3.1 – Nuage des individus N_I dans $\mathbb{R}^K$.

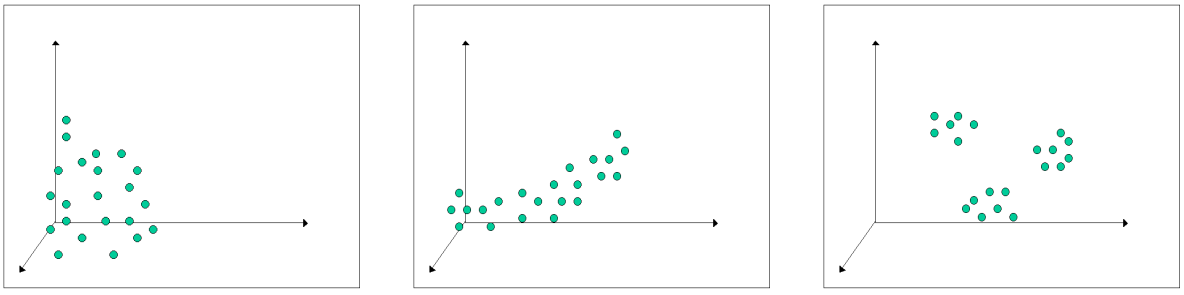


FIG. 3.2 – Différents types de nuages.

Ainsi le nuage N_I des individus est un espace vectoriel à K dimensions, dont chaque dimension représente une variable (*cf.* figure 3.1). Le fait d'avoir centré les données entraîne que l'origine des axes est confondu avec le centre de gravité G. Ce centre de gravité G peut s'interpréter comme l'individu moyen de la population. L'interprétation de ce nuage

N_I va se faire en décelant d'une part une partition de points et d'autre part des directions d'allongement. Ainsi sur la figure 3.2 nous représentons différents types de nuages possibles. Nous pouvons observer une absence de liens, ou bien par exemple une direction d'allongement suivant plutôt le premier axe, ou encore une partition des points en trois groupes. Si l'étude directe est envisageable dans un espace à trois dimensions, dès lors que $K > 3$ elle devient impossible. Nous avons donc recours à l'approche factorielle à partir de laquelle nous pouvons étudier différents plans de projection.

Analyse du nuage des variables

L'analyse du nuage N_K des variables se fait toujours à partir du tableau 3.2 des données centrées réduites, que nous considérons ici par colonne, comme nous l'avons déjà vu dans le cas général d'une analyse factorielle (figure 2.1 de la section 2.2).

La représentation du nuage N_K des variables se situe dans un espace vectoriel à I dimensions, chaque dimension représentant un individu de la population totale. La norme de chaque variable k est telle que :

$$\sum_{i \in I} \frac{1}{I} \left(\frac{x_{ik} - \overline{x}_k}{s_k} \right)^2 = 1. \tag{3.8}$$

Cette norme correspond également au coefficient de corrélation de la variable k avec elle même, donc $r(k,k) = 1$. Ainsi le nuage N_K est situé sur la sphère unité (de rayon 1) dans l'espace $\mathbb{R}^I$ (*cf.* figure 3.3). Il est intéressant de noter que le cosinus de l'angle entre les vecteurs représentant deux variables k et h est le produit scalaire $< k, h >= r(k,h)$. L'interprétation du coefficient de corrélation comme un cosinus est une propriété très importante puisqu'elle donne un support géométrique, donc visuel, au coefficient de corrélation. Cette propriété nécessite d'avoir au préalable centré les données, ce qui justifie une nouvelle fois cette transformation.

L'analyse du nuage N_K des variables se fera donc par l'étude des angles formés pour une variable k par Ok et les axes factoriels. Il est bon de noter que le centre de gravité du nuage N_K n'est pas l'origine de la sphère unité, à la différence du nuage N_I où le centre de gravité correspond au centre du repère lorsque les données sont centrées. Ainsi, ce sont les angles entre les vecteurs représentant les variables qui sont peu déformés par les projections et non pas les distances entre les points du nuage N_K.

Cette étude des angles est impossible à réaliser directement à cause de la dimension de $\mathbb{R}^I$. Elle se fera donc dans les plans de projection issus de l'approche factorielle.

3.2.4 L'ajustement

L'approche factorielle consiste donc à approcher ces nuages N_I et N_K dans des sous-espaces vectoriels permettant de fournir quelques images planes de ces nuages.

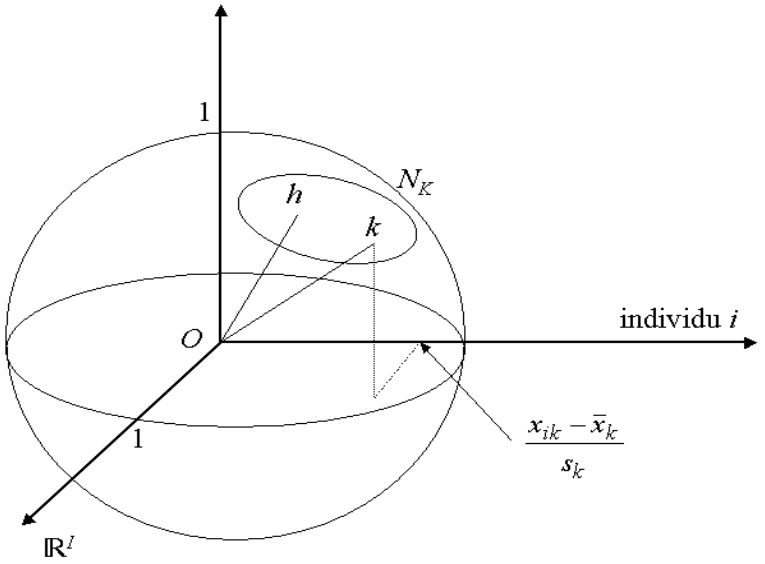


FIG. 3.3 – Nuage des variables N_K dans $\mathbb{R}^I$.

Ajustement du nuage des individus

Nous avons vu à la section 2.3 du chapitre précédent, qu'il faut chercher une suite $\{\mathbf{u}_s\,;s=1,\cdots,S\}$ de directions privilégiées - les axes factoriels - afin de fournir une représentation simplifiée du nuage N_I. Chaque direction $\mathbf{u}_s$ rend maximum l'inertie par rapport au centre de gravité G de la projection du nuage N_I sur l'axe factoriel $\mathbf{u}_s$. De plus les directions $\mathbf{u}_s$ sont orthogonales deux à deux.

Avec les notations de la figure 3.4 l'individu i se projette en H_i sur $\mathbf{u}_1$. Nous cherchons donc $\mathbf{u}_1$ qui rend maximum l'inertie $\sum_{i\in I} GH_i^2$. En effet, rendre maximum $\sum_{i\in I} GH_i^2$ revient à rendre minimum l'écart entre le nuage des individus et sa projection (*i.e.* $\sum_{i\in I} iH_i^2$), critère classique des moindres carrés. Ensuite, il faut trouver $\mathbf{u}_2$ orthogonal à $\mathbf{u}_1$ qui satisfait le même critère. Nous pouvons procéder ainsi jusqu'à l'obtention des S axes factoriels donnant une représentation suffisamment bonne.

Définition 3.2.3 *Les S axes factoriels $\{\mathbf{u}_s\,;s=1,\cdots,S\}$ sont appelées les facteurs principaux.*

Du fait d'avoir centré les données, ce critère permet d'interpréter les axes factoriels comme des directions d'allongement maximum du nuage N_K.

Ajustement du nuage des variables

Nous cherchons ici à obtenir des variables synthétiques $\{\mathbf{v}_s\,;s=1,\cdots,S\}$ et une représentation approchée des corrélations entre les variables. La démarche pour le nuage

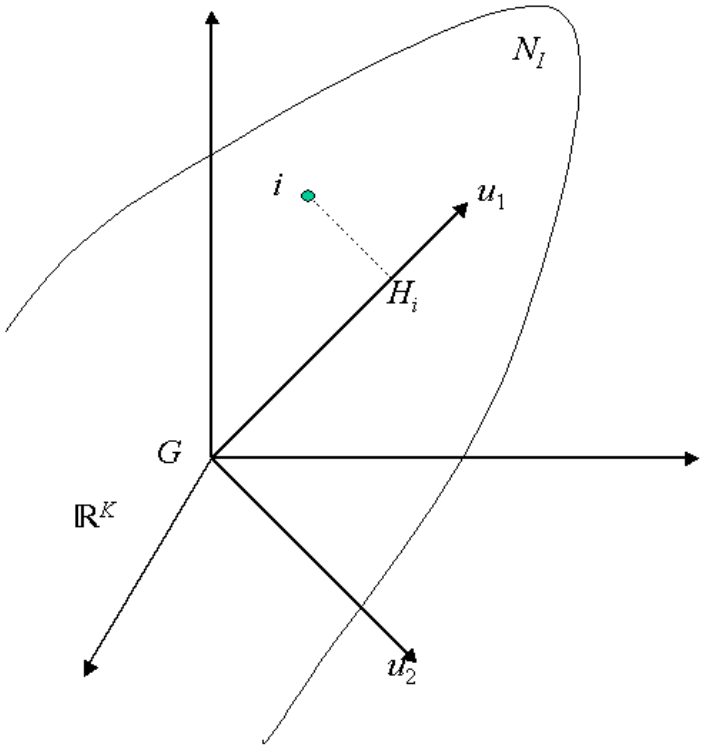


FIG. 3.4 – Ajustement du nuage N_I des individus pour l'ACP.

N_K reste la même que pour le nuage N_I des individus.

Ainsi, avec les notations de la figure 3.5, la variable k se projette en H_k sur $\mathbf{v}_1$. Nous cherchons le premier axe factoriel en déterminant le vecteur $\mathbf{v}_1$ qui rend maximum $\sum_{k \in K} OH_k^2$. Puis, nous cherchons le vecteur $\mathbf{v}_2$ orthogonal à $\mathbf{v}_1$ qui satisfait ce même critère. Nous poursuivons cette démarche jusqu'à l'obtention des S vecteurs recherchés.

Le vecteur $\mathbf{v}_1$ définit une nouvelle variable qui est la combinaison linéaire la plus liée à l'ensemble des variables initiales du tableau X. Ainsi les S vecteurs $\{\mathbf{v}_s\,;\, s = 1, \cdots, S\}$ étant orthogonaux deux à deux, les S nouvelles variables correspondantes sont non corrélées entre elles.

Définition 3.2.4 *Les S nouvelles variables (axes factoriels) $\{\mathbf{v}_s\,;\, s = 1, \cdots, S\}$ sont appelées les composantes principales.*

Ce sont ces vecteurs qui sont à l'origine du nom de cette analyse factorielle.

La coordonnée d'une variable initiale de X sur $\mathbf{v}_s$ est son coefficient de corrélation avec $\mathbf{v}_s$ du fait que les variables étudiées sont centrées réduites. Ainsi le vecteur $\mathbf{v}_1$ qui rend maximum $\sum_{k \in K} OH_k^2$ équivaut à la combinaison linéaire la plus liée à l'ensemble des variables initiales (la liaison étant entendu au sens du critère maximisant la somme des

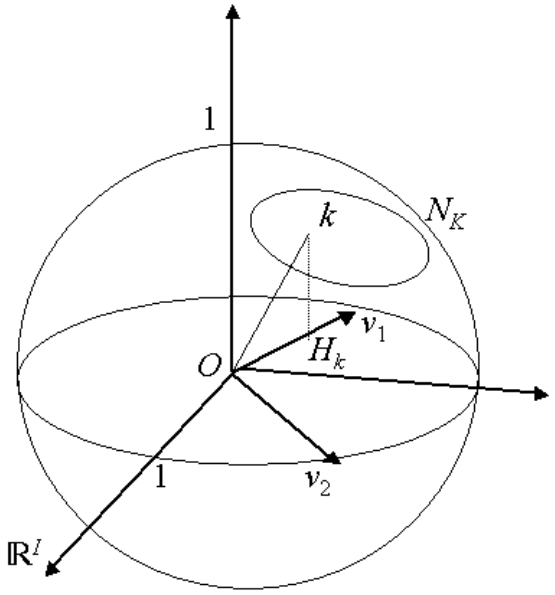


FIG. 3.5 – Ajustement du nuage N_K des variables pour l'ACP.

moindres carrés des corrélations). C'est donc la variable qui synthétise le mieux l'ensemble des variables initiales. Les axes factoriels résument donc l'ensemble des variables initiales du tableau X.

Effet de taille Un problème peut apparaître lorsque dans une population les variables sont toutes corrélées positivement deux à deux. Dans ce cas, elles forment des angles aigus et le centre de gravité G_K du nuage N_K est loin de l'origine de la sphère unité (*cf.* figure 3.6). Le premier axe factoriel est alors proche de la direction OG_K, ce qui fait qu'il représente mal le nuage N_K car toutes les projections des variables sont proches les unes des autres. En effet le premier axe factoriel rend toujours compte de la position du nuage N_K par rapport à l'origine.

3.3 Représentation simultanée

Nous avons vu à la section 2.5 qu'il existe des relations de transition entre les deux espaces $\mathbb{R}^K$ et $\mathbb{R}^I$. L'ACP permet pour une interprétation simultanée du nuage N_I et du nuage N_K de représenter ces deux nuages simultanément sur les plans issus des premiers axes factoriels. Nous devons cependant prendre garde au fait que les deux nuages ne sont en réalité pas dans les mêmes espaces qui ont des dimensions différentes. Cette représentation simultanée est essentiellement pragmatique.

En effet, le nuage des individus N_I et le nuage des variables N_K sont deux représenta-

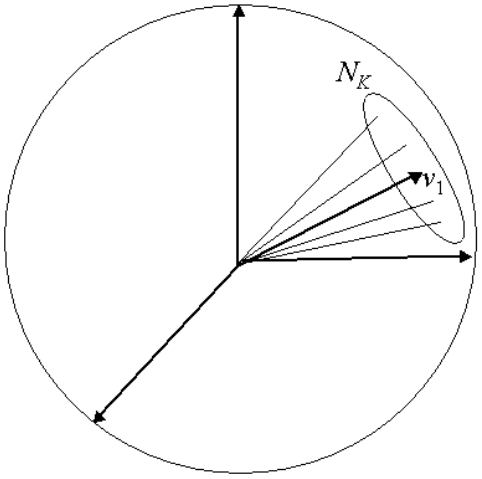


FIG. 3.6 – L'effet de taille dans $\mathbb{R}^I$.

tions du même tableau de données X. Ainsi des relations fortes (*relation de dualité*) lient ces deux nuages. Tout d'abord, l'inertie totale des deux nuages est la même :

$$\lambda = \frac{1}{I} \sum_{i \in I} \sum_{k \in K} \left(\frac{x_{ik} - \overline{x}_k}{s_k} \right)^2 . \tag{3.9}$$

De plus, les projections de tous les points du nuage des individus N_I sur le premier axe factoriel $\mathbf{u}_1$ constituent une nouvelle variable (appelée premier facteur, notée F_1) qui se confond à la norme près à la première composante principale (illustrées sur la figure 3.7). Ainsi le vecteur F_1 dans $\mathbb{R}^I$ est colinéaire à $\mathbf{v}_1$ (axe factoriel de N_K). Il en est de même pour les projections sur les autres facteurs qui correspondent aux composantes principales de même rang. De manière symétrique, les coordonnées des projections du nuage N_K sur

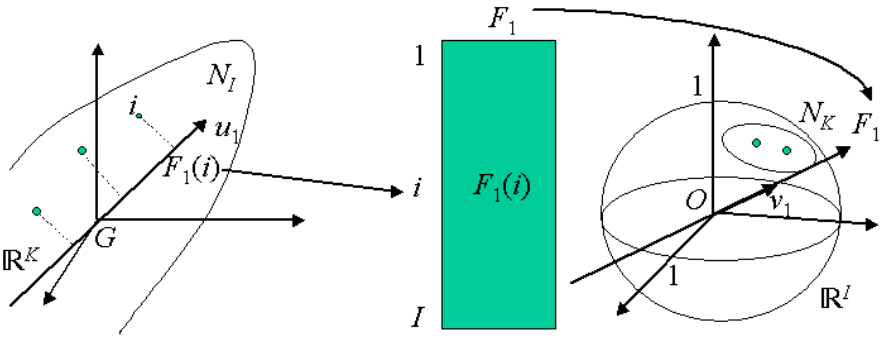


FIG. 3.7 – Forme de dualité exprimant le nuage N_I en fonction du nuage N_K.

l'axe factoriel $\mathbf{v}_1$ constituent un nouvel individu (premier facteur, noté G_1), ce que nous

représentons sur la figure 3.8. Ce vecteur G_1 de $\mathbb{R}^K$ est colinéaire à $\mathbf{u}_1$ (axe factoriel de N_I). Cette notion d'individu type est moins employée que celle de composante principale. Il est souvent plus facile de tenter de se ramener à des individus réels comme individu type.

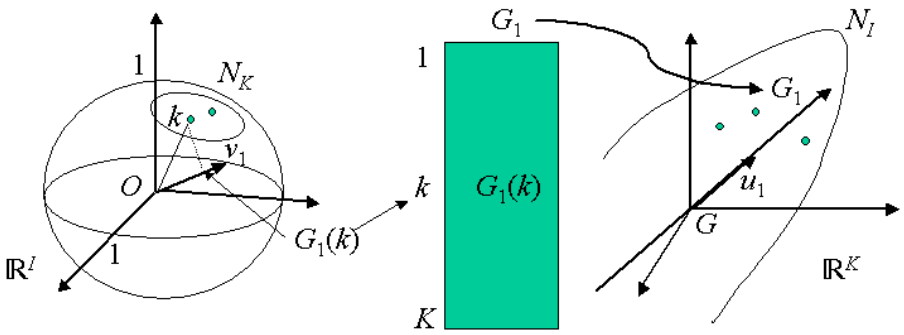


FIG. 3.8 – Forme de dualité exprimant le nuage N_K en fonction du nuage N_I.

Les relations algébriques des deux dualités précédentes au rang s sont données par :

$$\begin{cases} F_s(i) = \dfrac{1}{\sqrt{\lambda_s}} \displaystyle\sum_{k \in K} \frac{x_{ik} - \overline{x}_k}{s_k} G_s(k) \\ \\ G_s(i) = \dfrac{1}{I} \dfrac{1}{\sqrt{\lambda_s}} \displaystyle\sum_{i \in I} \frac{x_{ik} - \overline{x}_k}{s_k} F_s(k) \end{cases} \tag{3.10}$$

où λ_s est l'inertie projetée du nuage N_I (ou du nuage N_K) sur l'axe factoriel au rang s. Notons que les facteurs peuvent être négatifs.

Cette représentation est donc essentiellement une aide pour l'interprétation.

3.4 Interprétation

A partir des relations données précédemment, nous pouvons définir quelques règles pour l'interprétation :

- Un individu sera du côté des variables pour lesquelles il a de fortes valeurs, inversement il sera du côté opposé des variables pour lesquelles il a de faibles valeurs.
- Plus les valeurs d'un individu sont fortes pour une variable plus il sera éloigné de l'origine suivant l'axe factoriel décrivant le mieux cette variable.
- Deux individus à une même extrémité d'un axe (*i.e.* éloignés de l'origine) sont proches (*i.e.* se ressemblent).
- Deux variables très corrélées positivement sont du même côté sur un axe.

- Il n'est pas possible d'interpréter la position d'un individu par rapport à une seule variable, et réciproquement, il n'est pas possible d'interpréter la position d'une variable par rapport à un seul individu. Les interprétations doivent se faire de manière globale.

Les axes factoriels donnent des images approchées des nuages de points N_I et N_K. Il est donc nécessaire de définir des indicateurs pour mesurer la qualité de l'approximation. L'étude d'un plan de projection des sous-espaces vectoriels doit toujours se faire conjointement avec l'étude des indicateurs. En effet, deux points (individus ou variables) peuvent se trouver très proches dans un plan de projection, alors qu'en réalité ils sont éloignés. Nous présentons ici les principales aides à l'interprétation que nous retrouvons dans [EP90].

Qualité de représentation d'un élément (individu ou variable) par un axe La qualité de représentation d'un élément i par l'axe s est donnée par le rapport de l'inertie de la projection de l'élément i sur l'axe s et de l'inertie totale de l'élément i :

$$QLT_s(i) = \frac{(OH_i^s)^2}{(Oi)^2} = \cos^2\theta, \tag{3.11}$$

où θ est l'angle entre (Oi) et l'axe s. Ainsi si $QLT_s(i)$ est proche de 1, alors i est proche de l'axe s et donc du plan de projection contenant l'axe s.

Cette qualité se généralise au plan. Si un individu i est proche du plan, sa distance à G (l'individu moyen) dans le plan est proche de la valeur réelle. De même les distances dans le plan entre deux individus bien représentés sont proches de la réalité.

Qualité de représentation d'un nuage par un axe Cette qualité est donnée par le pourcentage d'inertie associé à un axe, c'est-à-dire le rapport de l'inertie de la projection du nuage sur l'axe et de l'inertie totale du nuage :

$$QLT_N = \frac{\sum_{i \in N} (OH_i^s)^2}{\sum_{i \in N} (Oi)^2}. \tag{3.12}$$

Cette qualité mesure "l'importance" d'un axe factoriel. Bien sûr les premiers axes auront plus d'importance que les suivants. Nous devons juger ces pourcentages en fonction de la taille du tableau. Par exemple, 10% est une valeur faible si le tableau comporte 10 variables ; c'est une valeur forte dans le cas de 100 variables.

Du fait de la dualité, il est équivalent de calculer ces pourcentages d'inertie à partir du nuage N_I des individus ou N_K des variables.

Contribution d'un élément à l'inertie d'un axe La contribution d'un élément i à l'inertie d'un axe s est donnée par le rapport de l'inertie de la projection de i sur l'axe s

et de l'inertie de la projection de l'ensemble du nuage sur l'axe s :

$$CT_s(i) = \frac{(OH_i^s)^2}{\sum_{i \in N} (Oi)^2}. \tag{3.13}$$

La contribution est importante si elle est proche de 1 pour les variables et doit être rapportée au tableau pour les individus. Ce rapport permet de mettre en évidence le sous-ensemble d'éléments ayant participé essentiellement à la construction de l'axe. L'interprétation devra en premier lieu s'appuyer sur ces éléments.

Pour aider à l'interprétation nous proposons de suivre le plan suivant :

- Choisir le nombre d'axes. Notons que le choix du nombre d'axes à retenir reste un problème car il n'y a pas de solutions rigoureuses. Les valeurs propres permettent de choisir ce nombre par exemple de telle sorte que le pourcentage d'information cumulée soit compris en 80% et 90% environ ou tel que toutes les valeurs propres soient supérieures à 1 ou encore lorsque un saut important sur l'histogramme des valeurs propres ou sur les recherches de paliers de celles-ci est observé. De plus le nombre d'axes ne doit pas être trop grand.
- Etudier les indicateurs de la qualité des approximations.
- Interpréter les facteurs simultanément :
  - à l'aide des contributions des individus,
  - à l'aide des coordonnées des variables (interpréter par axe et par plan),
  - à l'aide des coordonnées des individus.
- Mettre en évidence des typologies.

Il est possible de faire intervenir des éléments illustratifs (appelés également supplémentaires) afin d'aider l'opérateur à interpréter. Ces éléments, individus ou variables, n'interviennent pas dans la construction des axes factoriels, mais sont représentés pour l'étape d'interprétation. Dans le cas des variables, il s'agit de variables quantitatives qui peuvent être continues ou nominales. L'ajout d'éléments illustratifs doit rester exceptionnels, car ils n'appartiennent normalement pas au champ strict de l'étude. Il peut cependant parfois être intéressant de supprimer un individu provoquant un effet de taille dans le calcul des axes, et de le faire apparaître pour interpréter ses projections en fonction des autres individus.

3.5 Conclusion

Dans un premier temps résumons l'analyse en composantes principales à l'aide des neuf étapes de la figure 3.9 :

- 1 : La première étape concerne la mise en forme des données brutes.
- 2 : La deuxième étape consiste à centrer et réduire les données. Elles sont centrées afin d'obtenir des propriétés intéressantes, et réduites pour uniformiser les unités de mesure.
- 3 : Le tableau est considéré comme juxtaposition de lignes.

- 4 : Le tableau est considéré comme juxtaposition de colonnes.
- 5 : Les individus sont représentés dans un espace à K dimensions. Dans le nuage N_I nous nous intéressons aux distances inter-individuelles qui déterminent les ressemblances. Le centre de gravité G représente un individu moyen.
- 6 : Les variables sont représentées dans un espace à I dimensions. Nous nous intéressons ici aux angles des points. Le cosinus de l'angle est le coefficient de corrélation. Toutes les variables sont équidistantes de l'origine car les données ont été réduites, ainsi le nuage N_K se situe sur une hypersphère.
- AF : Analyse Factorielle. Cette phase permet de mettre en évidence une suite de directions. Dans l'étape 7 ces directions sont des directions d'allongement, et dans l'étape 8 les axes s'interprètent comme des variables synthétiques.
- 7 : Cette étape consiste à projeter les points du nuage N_I sur le premier plan factoriel. C'est un premier ajustement, il peut y en avoir d'autres à suivre. Les distances s'interprètent alors comme des ressemblances entre les individus.
- 8 : Cette étape consiste à projeter les points du nuage N_K sur le premier plan factoriel. Ici aussi, c'est un premier ajustement, et il peut y en avoir d'autres à suivre. Les coordonnées représentent les coefficients de corrélation avec les facteurs sur les individus.
- Les relations de transition expriment les résultats d'une analyse factorielle (AF) dans un espace en fonction des résultats de l'autre.
- 9 : Cette étape est la représentation simultanée de nuages de points qui se trouvent initialement dans des espaces de dimensions différentes. Cette représentation issue des relations de transition permet des interprétations des axes simultanées.

L'ACP est une méthode puissante pour synthétiser et résumer de vastes populations décrites par plusieurs variables quantitatives. Elle permet entre autre de dégager de grandes catégories d'individus et de réaliser un bilan des liaisons entre les variables. Par cette analyse nous pouvons mettre en évidence de grandes tendances dans les données telles que des regroupements d'individus ou des oppositions entre individus (ce qui traduit un comportement radicalement différent de ces individus) ou entre variables (ce qui traduit le fait que les variables sont inversement corrélées). Les représentations graphiques fournies par l'ACP sont simples et riches d'informations. L'ACP peut être une première analyse pour l'étude d'une population dont les résultats seront enrichis par une autre analyse factorielle ou encore une classification automatique des données.

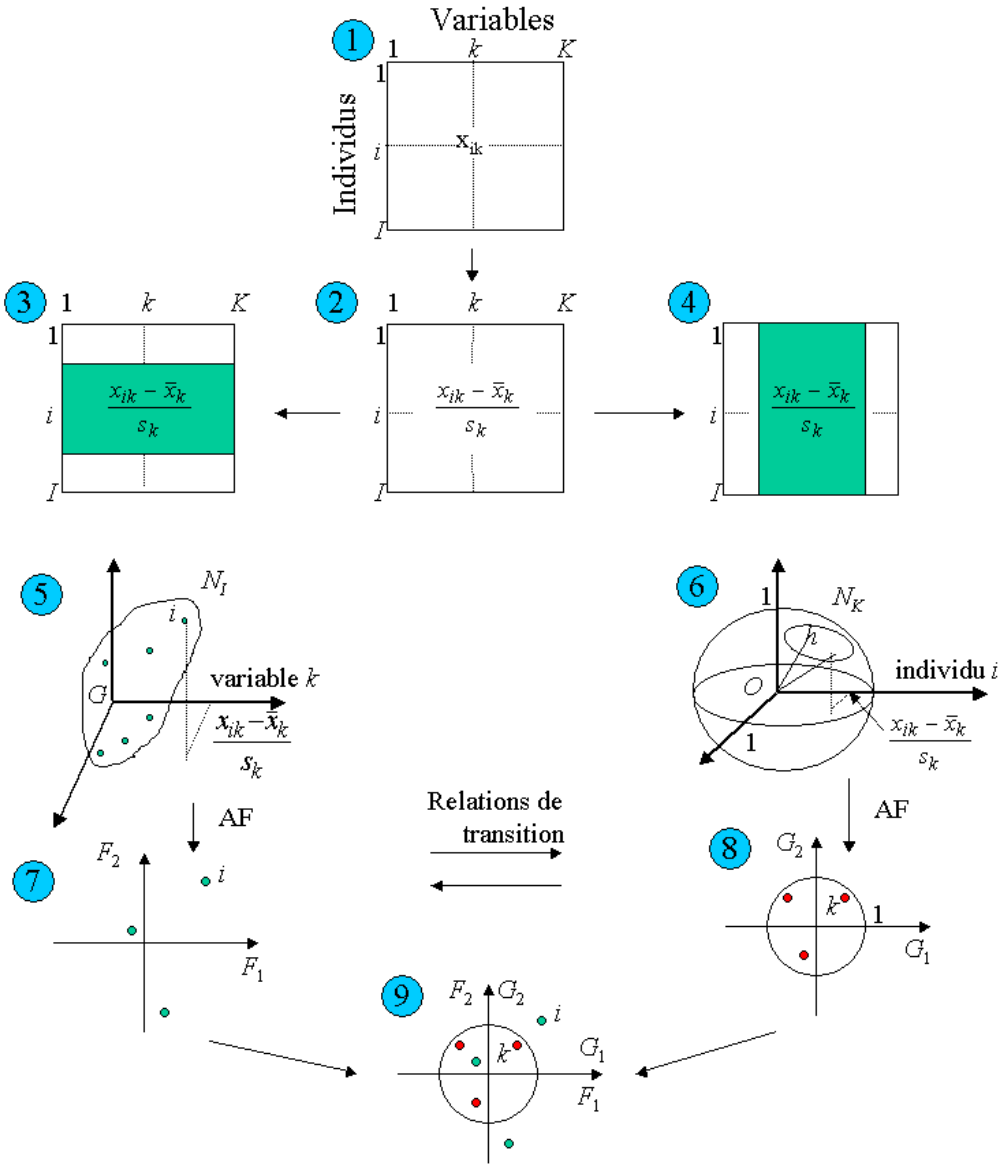


FIG. 3.9 – Résumé de l'ACP.

Chapitre 4

Analyse Factorielle des Correspondances

4.1 Introduction

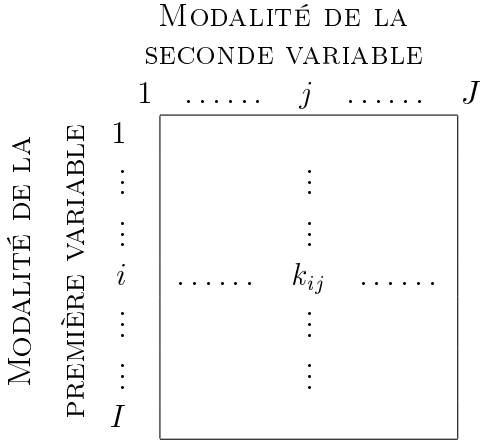
L'analyse factorielle des correspondances a été introduite par [Ben80b] sous le nom d'analyse des correspondances. Elle porte également le nom d'analyse des correspondances binaires en relation avec l'analyse des correspondances multiples que nous présentons ensuite. Nous la notons par la suite AFC. Cette analyse peut être présentée sous de nombreux points de vues, notamment comme un cas particulier de l'analyse canonique ou encore de l'analyse factorielle discriminante. Elle peut aussi être étudiée comme une ACP avec une métrique spéciale (celle du χ^2) [Sap90]. Nous la présentons ici suivant les points dégagés par une analyse factorielle vus au chapitre 2.

4.1.1 Les domaines d'application

Très tôt cette analyse a été utilisée en pratique car elle est conçue pour les *tableaux de contingence* et permet ainsi l'étude des liaisons (dites aussi correspondances) existant entre deux variables nominales. Les domaines d'application de l'AFC sont donc différents de ceux de l'ACP qui est adaptée aux tableaux de mesures hétérogènes ou non.

Pour cette analyse aussi nous pouvons donner une longue liste des disciplines ayant trouvé réponse à leur problème par l'AFC. Ainsi, l'écologie, la zoologie, la psychologie, l'économie, et d'autres encore dans lesquelles il peut être intéressant d'étudier les liaisons entre deux variables nominales, ont fourni un grand nombre de données.

L'AFC conçue pour les tableaux de contingence (*i.e.* fréquences), peut être appliquée aux tableaux de mesures homogènes (*i.e.* même système d'unités), aux tableaux de notes, de rangs, de préférences, aux tableaux à valeurs logiques (0 ou 1), et encore aux tableaux issus de questionnaires d'enquêtes.



Tab. 4.1 – Représentation des données pour l'AFC.

4.1.2 Les données

Les données, à la différence de l'ACP, doivent être organisées en tableaux de contingence (appelés aussi tableau de dépendance ou tableau croisé).

Définition 4.1.1 *Un tableau de contingence est un tableau d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus.*

L'AFC peut également être étendue aux variables quantitatives homogènes en définissant simplement quelques modalités pour ces variables. Par extension, elle s'applique aussi aux tableaux individus-variables pour des variables quantitatives homogènes, dans ce cas les individus sont considérés comme des variables.

Nous devons donc considérer les données brutes organisées de la façon décrite sur le tableau 4.1. Dans ce cas, I représente le nombre de lignes et l'ensemble des lignes $I = \{1, \cdots, I\}$, J représente le nombre de colonnes et l'ensemble des colonnes $J = \{1, \cdots, J\}$, et k_{ij} est le nombre d'individus possédant à la fois la modalité i de la première variable et la modalité j de la seconde variable. Nous avons donc :

$$\sum_{i \in I} \sum_{j \in J} k_{ij} = n, \tag{4.1}$$

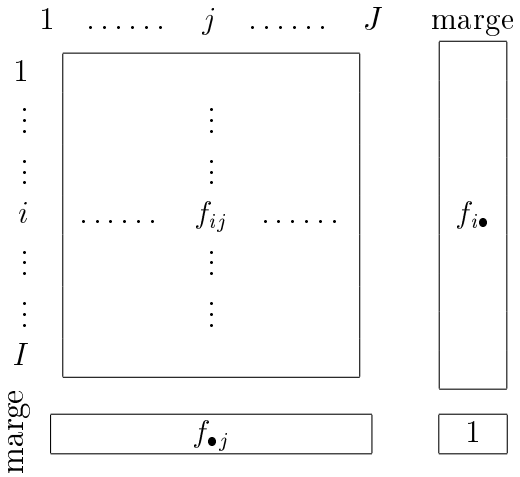
avec n le nombre total d'individus de la population initiale. Nous constatons que sur ce type de tableau les lignes et les colonnes jouent un rôle symétrique.

Davantage que le tableau 4.1, c'est le tableau des fréquences relatives 4.2 qui est considéré. Les fréquences f_{ij} sont données par :

$$f_{ij} = \frac{k_{ij}}{n}, \tag{4.2}$$

et les marges par :

$$f_{i\bullet} = \sum_{j \in J} f_{ij}, \tag{4.3}$$



TAB. 4.2 – Tableau des fréquences relatives pour l'AFC.

et

$$f_{\bullet j} = \sum_{i \in I} f_{ij}. \tag{4.4}$$

Nous avons ainsi :

$$\sum_{i \in I} f_{i\bullet} = \sum_{j \in J} f_{\bullet j} = \sum_{i \in I} \sum_{j \in J} f_{ij} = 1. \tag{4.5}$$

Liaisons entre les variables Nous avons vu que l'AFC considère un tableau de contingence ou de fréquence pour étudier les liaisons entre les deux variables à l'initiative du tableau. Nous ne pouvons plus définir les liaisons par les coefficients de corrélation comme pour l'ACP (*cf.* Chapitre 3).

Définition 4.1.2 *Il y a indépendance entre les deux variables considérées si :*

$$f_{ij} = f_{i\bullet} f_{\bullet j}, \ \forall i \in I, \ \forall j \in J. \tag{4.6}$$

Définition 4.1.3 *Nous disons qu'il y a liaison entre ces deux variables, ou que ces deux variables sont liées si elles ne sont pas indépendantes.*

Ainsi nous pouvons dire que :

- Si f_{ij} est supérieur au produit des marges, les modalités i et j s'associent plus que sous l'hypothèse d'indépendance. Nous dirons que les deux modalités i et j "s'attirent".
- Si f_{ij} est inférieur au produit des marges, les modalités i et j s'associent moins que sous l'hypothèse d'indépendance. Nous dirons qu'il y a "répulsion" entre les deux modalités i et j.

Sous l'hypothèse d'indépendance nous avons :
- en considérant le tableau comme un ensemble de lignes :

$$\frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}, \forall i \in I, \forall j \in J, \tag{4.7}$$

- en considérant le tableau comme un ensemble de colonnes :

$$\frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}, \forall i \in I, \forall j \in J. \tag{4.8}$$

Dans l'équation (4.7), le terme de droite $f_{\bullet j}$ s'interprète comme le pourcentage de la population totale possédant la modalité j, et le terme $\frac{f_{ij}}{f_{i\bullet}}$ représente ce même pourcentage dans la sous-population possédant la modalité i.

Ainsi il y a indépendance lorsque les lignes du tableau de fréquences sont proportionnelles. Par symétrie il en est de même pour les colonnes.

4.1.3 Les objectifs

Les objectifs sont les mêmes que ceux de l'ACP dans le sens où l'AFC cherche donc à obtenir une typologie des lignes et une typologie des colonnes, puis de relier ces deux typologies. Il faut donc faire ressortir un bilan des ressemblances entre lignes (respectivement colonnes) en répondant aux questions du type :
- Quels sont les lignes (respectivement colonnes) qui se ressemblent ?
- Quelles sont celles qui sont différentes ?
- Existe-t-il des groupes homogènes de lignes (respectivement colonnes) ?
- Est-il possible de mettre en évidence une typologie des lignes (respectivement des colonnes) ?

La notion de ressemblance entre deux lignes ou deux colonnes diffère cependant de l'ACP. En effet, deux lignes (respectivement deux colonnes) sont *proches* si elles s'associent de la même façon à l'ensemble des colonnes (respectivement des lignes), *i.e.* elles s'associent trop ou trop peu par rapport à l'indépendance.

Il faut donc chercher les lignes (respectivement colonnes) dont la répartition s'écarte le plus de l'ensemble de la population, celles qui se ressemblent entre elles et celles qui s'opposent. Afin de relier la typologie des lignes avec l'ensemble des colonnes, chaque groupe de lignes est caractérisé par les colonnes auxquelles ce groupe s'associe peu ou fortement. Par symétrie, chaque groupe de colonnes est caractérisé par les lignes auxquelles ce groupe s'associe peu ou fortement. Ainsi nous pouvons décomposer la liaison entre deux variables en une somme de tendances simples et interprétables et mesurer leur importance respective.

4.2 Principe de l'AFC

Nous allons présenter le principe de l'AFC et la démarche à suivre en illustrant les étapes par un tableau de données de faible dimension. De ce fait l'AFC ne se justifie

		Couleurs des cheveux				Total
		brun	châtain	roux	blond	
Couleurs des yeux	marron	68	119	26	7	220
	noisette	15	54	14	10	93
	vert	5	29	14	16	64
	bleu	20	84	17	94	215
Total		108	286	71	127	592

TAB. 4.3 – Tableau de contingence.

		Couleurs des cheveux				Profil moyen
		brun	châtain	roux	blond	
Couleurs des yeux	marron	11,4	20,1	4,3	1,1	37,1
	noisette	2,5	9,1	2,3	1,6	15,7
	vert	0,8	4,8	2,3	2,7	10,8
	bleu	3,3	14,1	2,8	15,8	36,3
Profil moyen		18,2	48,3	11,9	21,4	$\simeq 100$

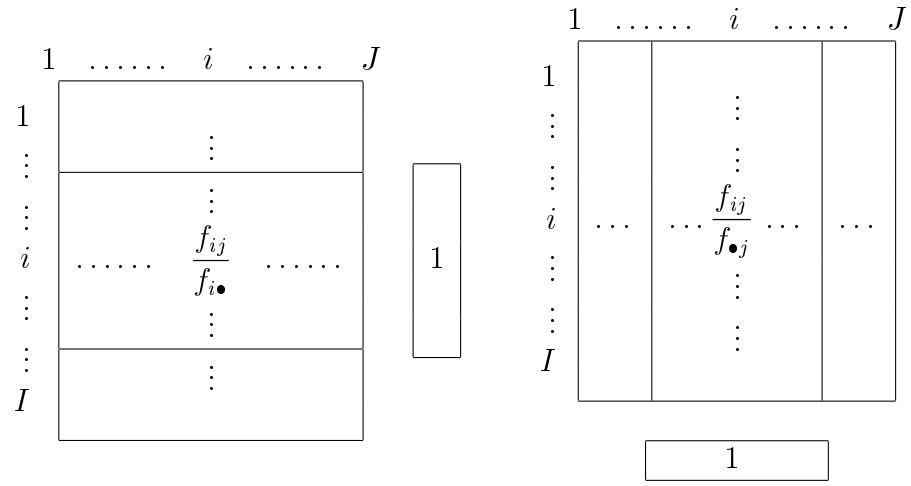
TAB. 4.4 – Tableau des fréquences observées.

pas vraiment, puisque les conclusions sont immédiates dès le tableau de contingence, cependant nous verrons clairement le principe et les propriétés de cette analyse.

Prenons l'exemple simple de la répartition de 592 femmes selon les couleurs des yeux et des cheveux (exemple proposé par Cohen en 1980 et repris dans [LMP95]). Le tableau 4.3 de contingence donne le nombre de femmes possédant à la fois une des quatre modalités de la couleur des cheveux et une des quatre modalités de la couleurs de yeux. Ainsi $I = J = 4$ et $n = 592$. Le tableau des fréquences 4.4 correspondant permet de ne plus tenir compte du nombre de femmes total. Ainsi nous pouvons nous demander s'il y a indépendance entre la couleur des yeux et celle des cheveux, ou encore quelles sont les associations entre ces couleurs. Sur cet exemple les réponses sont assez faciles, cependant lorsque la taille des données est plus importante, nous ne pouvons travailler directement sur le tableau des données brutes. Comme dans l'ACP, il y a une étape de transformation des données.

4.2.1 La transformation des données

Nous allons considérer le tableau d'une part comme une suite de lignes, puis comme une suite de colonnes (*cf.* tableau 4.5). Lorsque le tableau est considéré en ligne les données sont normalisées en divisant par $f_{i\bullet}$, la nouvelle ligne ainsi créée est appelée *profil-ligne*. Cette normalisation a pour but de considérer les liaisons entre les deux variables au travers de l'écart entre les pourcentages en lignes. Dans cette exemple $\dfrac{f_{ij}}{f_{i\bullet}}$ représente la probabilité d'avoir les cheveux de couleur j sachant que la couleur des yeux est i. Le profil-ligne i est la probabilité conditionnelle définie par i sur l'ensemble des colonnes. Un raisonnement similaire peut être fait pour les colonnes du fait du rôle symétrique joué par les lignes



TAB. 4.5 – Les profil-ligne et profil-colonne.

		Couleurs des cheveux				Profil
		brun	châtain	roux	blond	moyen
Couleurs des yeux	marron	30,9	54,0	11,8	3,1	$\simeq 100$
	noisette	16,1	58,0	15,0	10,7	$\simeq 100$
	vert	7,8	45,3	21,8	25,0	$\simeq 100$
	bleu	9,3	39,0	7,9	43,7	$\simeq 100$
Profil moyen		18,2	48,3	11,9	21,4	$\simeq 100$

TAB. 4.6 – Profils-lignes (exprimés en pourcentages-lignes arrondis).

et les colonnes. Ainsi $\frac{f_{ij}}{f_{\bullet j}}$ représente la fréquence pour une femme d'avoir les yeux d'une couleur i sachant qu'elle a les cheveux de couleur j. Si nous reprenons notre exemple sur les couleurs de cheveux et des yeux, nous obtenons les profils-lignes et les profils-colonnes donnés respectivement par les tableaux 4.6 et 4.7. Le tableau 4.6 représente donc les probabilités conditionnelles d'avoir les cheveux de la couleur j sachant que les yeux ont la couleur i. Le tableau 4.7 fournit la répartition de la couleur des yeux selon les modalités de la couleur des cheveux. Nous avons donc par exemple 31 chances sur 100 que les femmes qui ont les yeux marrons aient les cheveux de couleur brun, et 63 chances sur 100 que les femmes qui ont les cheveux de couleur brun aient les yeux marrons. Nous savons aussi à partir du tableau 4.4 que 11 femmes sur 100 ont à la fois les yeux marrons et les cheveux de couleur brun.

4.2.2 La ressemblance entre profils

La ressemblance entre deux lignes ou entre deux colonnes est définie par une distance entre profils. La distance employée est celle du χ^2 et elle est définie de façon symétrique

		Couleurs des cheveux				Profil
		brun	châtain	roux	blond	moyen
Couleurs des yeux	marron	62,9	41,6	36,6	5,5	37,1
	noisette	13,8	18,8	19,7	7,8	15,7
	vert	4,6	10,1	19,7	12,5	10,8
	bleu	18,5	29,3	23,9	74,0	36,3
Profil moyen		$\simeq 100$	$\simeq 100$	$\simeq 100$	$\simeq 100$	$\simeq 100$

TAB. 4.7 – Profils-colonnes (exprimés en pourcentages-colonnes arrondis).

pour les lignes et les colonnes. Ainsi entre deux lignes i et i' elle est donnée par :

$$d_{\chi^2}(i,i') = \sum_{j\in J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 , \tag{4.9}$$

et entre deux colonnes j et j' par :

$$d_{\chi^2}(j,j') = \sum_{i\in I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2 . \tag{4.10}$$

La matrice diagonale $\dfrac{1}{f_{\bullet j}}$ définie la métrique dans $\mathbb{R}^J$, tandis que $\dfrac{1}{f_{i\bullet}}$ définie celle dans $\mathbb{R}^I$. Cette pondération $\dfrac{1}{f_{\bullet j}}$ équilibre l'influence des colonnes sur la distance entre les lignes en augmentant les termes concernant les modalités rares.

Remarque D'autres distances pourraient être employées, cependant la distance euclidienne usuelle entre les points-lignes ou entre les points-colonnes exprimés à partir du tableau de fréquence ne traduit que les différences d'effectifs entre deux modalités. La distance euclidienne entre les profils-lignes ou entre les profils-colonnes permet de bien modéliser les ressemblances entre deux modalités. Par exemple pour la distance entre deux profils-lignes est donnée par :

$$d(i,i') = \sum_{j\in J} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 . \tag{4.11}$$

Cependant cette distance favorise les colonnes ayant une masse $f_{\bullet j}$ important. Ainsi dans notre exemple elle favorise les couleurs de cheveux bien représentées tel que le châtain. C'est pour cette raison que la distance retenue dans l'équation (4.9) (respectivement (4.10)) l'écart entre les profils est pondéré par l'inverse de la masse de la colonne (respectivement de la ligne). Cette distance est nommée distance du χ^2 car elle proportionnelle à la statistique du χ^2 de Karl Pearson. De plus cette distance du χ^2 possède une propriété fondamentale nommée l'*équivalence distributionnelle*. Cette propriété permet d'associer

deux modalités d'une même variable qui possède des profils identiques en une modalité unique affectée de la somme de leurs masses, sans modifier ni les distances entre les modalités de cette variable, ni les distances entre les modalités de l'autre variable. Ainsi, si deux colonnes proportionnelles d'un tableau sont regroupées, les distances entre profils-lignes sont inchangées, et réciproquement. Ceci permet de regrouper des modalités voisines pour ainsi réduire le nombre de modalités et donc la complexité de l'interprétation en garantissant une certaine invariance des résultats.

4.2.3 Les nuages des deux profils

Le nuage des profils-lignes

Lorsque nous nous intéressons aux modalités de la première variable, il faut considérer les données comme une juxtaposition de profils-lignes. Ainsi chaque profil-ligne i peut être représenté comme un point de l'espace $\mathbb{R}^J$ dont chacune des J dimensions représente une modalité de la seconde variable (*cf.* figure 4.1). L'utilisation de la distance entre

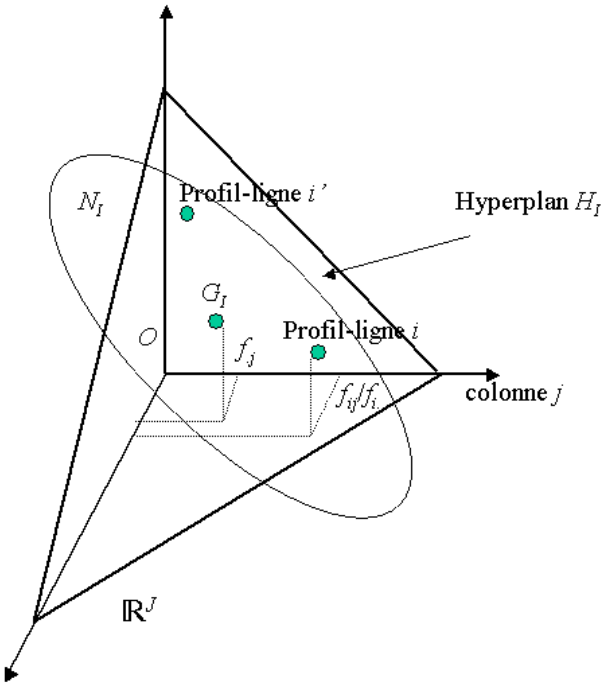


FIG. 4.1 – Le nuage N_I des profils-lignes dans l'espace $\mathbb{R}^J$.

deux profils est celle χ^2, elle revient à affecter le poids $\dfrac{1}{f_{\bullet j}}$ à la $j^{\text{ème}}$ dimension de $\mathbb{R}^J$. Du fait que la somme de chaque profil-ligne est égale à 1, le nuage N_I appartient à un hyperplan, noté H_I. Pour l'AFC les poids affectés à chaque point du nuage sont imposés et ne sont pas identiques. Le point i a pour poids la fréquence marginale $f_{i\bullet}$. Ce poids est

naturel puisqu'il est proportionnel à l'effectif de la classe d'individus qu'il représente. La coordonnée du point i sur l'axe j est donnée par $\frac{f_{ij}}{f_{i\bullet}}$.

Le barycentre des points de N_I munis de ces poids, noté G_I, est la moyenne pondérée de tous les points sur tous les axes j. La coordonnée de G_I sur l'axe j est donc donnée par :

$$\sum_{i \in I} f_{i\bullet} \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}. \tag{4.12}$$

Le barycentre s'interprète comme un profil-moyen. Dans l'étude des lignes, il sert de référence pour étudier dans quelle mesure et de quelle façon une classe d'individus diffère de l'ensemble de la population. Ceci se fait par l'étude de l'écart entre le profil de cette classe et le profil moyen. Ainsi l'étude de la dispersion du nuage autour de son barycentre équivaut à l'étude de l'écart entre profils et marge ou encore à l'étude de la liaison entre les deux variables.

Le nuage des profils-colonnes

La construction du nuage des profils-colonnes est identique à celle du nuage des profils-lignes du fait de la symétrie entre les lignes et les colonnes en AFC. Ainsi, lorsque nous nous intéressons aux modalités de la seconde variable, il faut considérer les données comme une juxtaposition de profils-colonnes. Chaque profil-colonne j peut être représenté comme un point de l'espace $\mathbb{R}^I$ dont chacune des I dimensions représente une modalité de la première variable (*cf.* figure 4.2). Le point i a pour coordonnée sur l'axe i la proportion $\frac{f_{ij}}{f_{\bullet j}}$, et le poids qui lui est associé est $f_{i\bullet}$. Le nuage N_J appartient à un hyperplan noté H_J. De plus le barycentre des points de N_J munis de leur poids a pour coordonnée sur l'axe i :

$$\sum_{j \in J} f_{\bullet j} \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}. \tag{4.13}$$

Ce barycentre s'interprète également comme un profil moyen et joue le même rôle pour l'étude de la liaison entre les deux variables.

4.2.4 L'ajustement des deux nuages

Les deux hyperplans H_I et H_J sont de grande dimension si la taille des données est importante. Nous ne pouvons donc pas les étudier directement. Ainsi, nous cherchons à fournir des images planes des nuages N_I et N_J. La démarche reste la même que celle présentée au Chapitre 2.

Ainsi, pour l'ajustement du nuage des profils-lignes, nous cherchons une suite d'axes orthogonaux deux à deux $\{\mathbf{u}_s\,; s = 1, \cdots, S\}$ sur lesquels le nuage N_I est projeté. Chaque axe $\mathbf{u}_s$ doit rendre maximum l'inertie projetée du nuage N_I. En pratique, nous devons centrer le nuage N_I, ainsi le centre de gravité G_i devient l'origine des axes. Une fois le

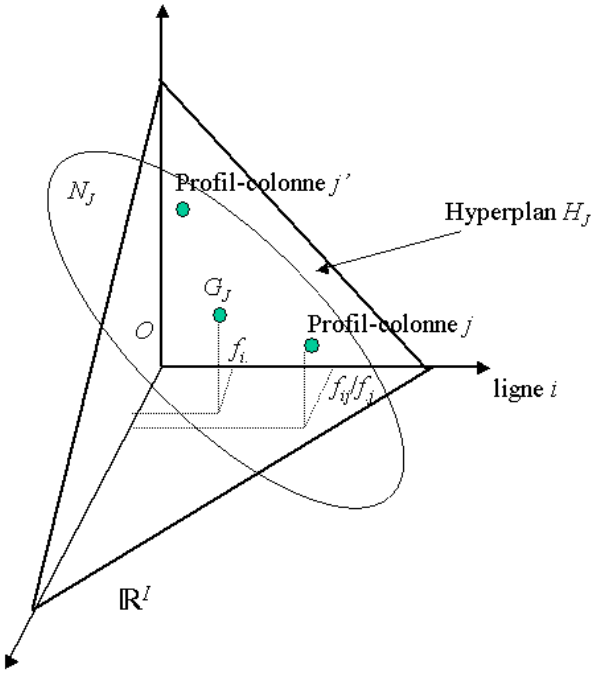


FIG. 4.2 – Le nuage N_J des profils-colonnes dans l'espace $\mathbb{R}^I$.

nuage centré, la modalité i a pour coordonnée $\dfrac{f_{ij}}{f_{i\bullet} - f_{\bullet j}}$ sur le $j^{\text{ème}}$ axe. Cette coordonnée exprime la différence entre la répartition de la classe i et celle de la population totale sur l'ensemble des modalités de la seconde variable. La recherche des axes qui rendent maximum l'inertie du nuage centré revient donc à chercher les classes qui s'écartent le plus du profil de l'ensemble de la population. Chaque profil-ligne étant muni d'un poids $f_{i\bullet}$, l'inertie est donnée par :

$$\sum_{i \in I} f_{i\bullet} \left(\frac{f_{ij}}{f_{i\bullet} - f_{\bullet j}} \right)^2 . \tag{4.14}$$

L'ajustement du nuage des profils-lignes N_I dans $\mathbb{R}^J$ revient donc à chercher le premier vecteur unitaire $\mathbf{u}_1$ qui rende cette inertie maximale, puis par chercher le vecteur unitaire $\mathbf{u}_2$ orthogonal à $\mathbf{u}_1$ qui vérifie le même critère, *etc.*

Cette démarche est semblable à celle de l'ACP, à l'exception du fait que les lignes interviennent au travers de leur profil, que la distance entre les profils est celle du χ^2 et que chaque élément i est affecté d'un poids $f_{i\bullet}$.

Puisqu'en AFC les lignes et les colonnes jouent un rôle symétrique, l'ajustement du nuage N_J est semblable à celui de N_I. Ainsi les images planes du nuage N_J doivent être telles que les distances entre les points de l'image ressemblent le plus possible aux distances entre les points du nuage N_J. L'analyse du nuage N_J se fait également par rapport au barycentre G_J.

4.2.5 Représentation simultanée

En AFC, la représentation simultanée des deux nuages N_I et N_J repose sur une dualité plus riche qu'en ACP car les lignes et les colonnes représentent des éléments de même nature. Les deux nuages N_I et N_J sont deux représentations du même tableau en le considérant en tant que profils-lignes et profils-colonnes. L'analyse du tableau passe donc par les analyses des nuages qui ne sont pas indépendantes.

Remarque sur l'inertie L'inertie du nuage N_I est donnée par :

$$I_{N_I} = \sum_{i \in I} f_{i\bullet} \sum_{j \in J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 = \sum_{i \in I} \sum_{j \in J} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}. \tag{4.15}$$

Nous constatons que l'inertie du nuage N_J est identique à celle du nuage N_I. Cette inertie représente la liaison entre les deux variables. En effet la statistique du χ^2 habituellement employée pour mesurer la liaison entre deux variables est la somme du rapport avec pour numérateur le carré de la différence de l'effectif observé et de l'effectif théorique et pour dénominateur l'effectif théorique :

$$\chi^2 = \sum_{i \in I} \sum_{j \in J} \frac{(n f_{ij} - n f_{i\bullet} f_{\bullet j})^2}{n f_{i\bullet} f_{\bullet j}} = n I_{N_I} = n I_{N_J}. \tag{4.16}$$

Ainsi la statistique du χ^2 est égale, au coefficient n près, à l'inertie totale du nuage N_I et du nuage N_J. Ceci justifie une nouvelle fois l'emploi de la distance du χ^2.

Nous avons vu au chapitre 2 que les inerties associées à chaque axe de même rang dans chacun des nuages sont égales, ainsi que les facteurs de même rang sur les lignes et les colonnes sont liés par des relations de transition. Ces relations donnent un sens à une représentation simultanée. Le schéma de dualité de la figure 4.3 représente les relations de transition (appelées également barycentriques, ou encore quasi-barycentriques) données par :

$$\begin{cases} F_S(i) = \dfrac{1}{\sqrt{\lambda_S}} \displaystyle\sum_{j \in J} \frac{f_{ij}}{f_{i\bullet}} G_S(j) \\ \\ G_S(j) = \dfrac{1}{\sqrt{\lambda_S}} \displaystyle\sum_{i \in I} \frac{f_{ij}}{f_{\bullet j}} F_S(i) \end{cases} \tag{4.17}$$

$F_S(i)$ représente la projection de la ligne i sur l'axe de rang S du nuage N_I, $G_S(j)$ représente la projection de la colonne j sur l'axe de rang S du nuage N_J, et λ_S est la valeur commune de l'inertie associée à chacun de ces axes. Elle est donnée par :

$$\lambda_S = \sum_{i \in I} f_{i\bullet} \left[F_S(i) \right]^2 = \sum_{j \in J} f_{\bullet j} \left[G_S(j) \right]^2. \tag{4.18}$$

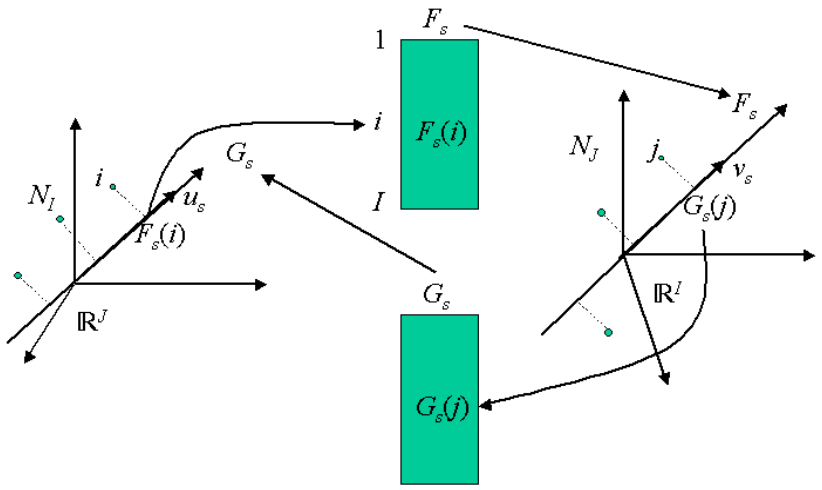


FIG. 4.3 – Le schéma de dualité pour l'AFC.

La projection de la ligne i sur l'axe S est le barycentre des projections des colonnes j sur l'axe S, chaque colonne j étant affectée du poids f_{ij}. Cette propriété est appelée propriété barycentrique.

La représentation simultanée s'obtient en superposant les projections de chacun des deux nuages N_I et N_J sur des plans engendrés par des axes de même rang pour les deux nuages. Bien sûr les deux nuages ne sont pas forcément dans le même espace. Si la représentation simultanée n'est pas adoptée par tous pour l'ACP, elle se justifie beaucoup mieux pour l'AFC. En fait pour pouvoir réellement superposer les deux nuages, il faudrait avoir les mêmes barycentres car chaque nuage devrait alors être contenu dans l'autre. Cette représentation est possible en forçant les centres de gravité pour approcher la solution idéale. Les relations seront alors quasi-barycentriques.

4.3 Interprétation

La représentation simultanée des lignes et des colonnes permet une interprétation aisée des projections. Ainsi la position relative de deux points d'un même ensemble (ligne ou colonne), s'interprète en tant que distance. La position d'un point d'un ensemble et tous les points d'un autre ensemble s'interprète en tant que barycentre. Attention cependant, toute association entre une ligne et une colonne suggérée par une proximité sur le graphique doit être contrôlée sur le tableau.

Reprenons l'exemple précédent sur la couleur des yeux et des cheveux. La représentation simultanée sur le premier plan factoriel (*cf.* figure 4.4) montre par exemple que les femmes aux yeux bleus et aux yeux marrons sont éloignées. En confirmation avec le tableau, nous remarquons que les femmes aux yeux bleus auront tendance à avoir les cheveux blonds, ainsi que pour celles aux yeux marrons qui seront davantage brunes. Les femmes aux cheveux roux auront plutôt les yeux verts ou noisettes. La modalité des

cheveux châtains est proche de l'origine, elle représente donc un profil moyen et n'est rattachée à aucune couleur de cheveux.

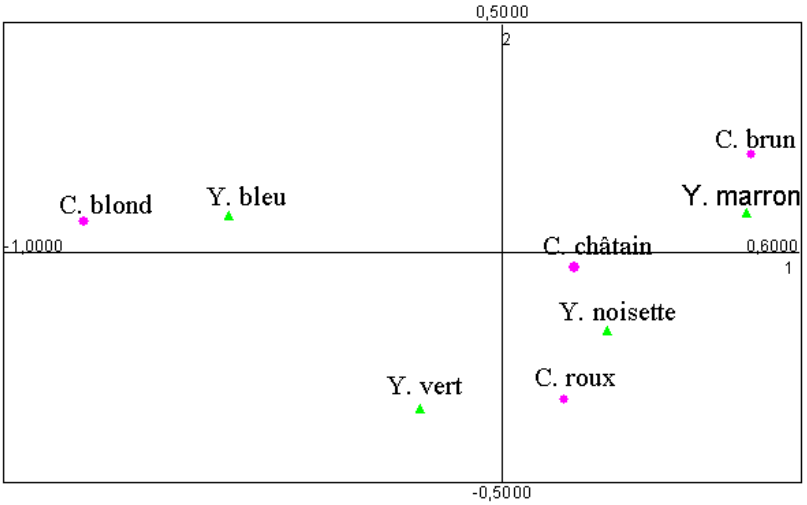


FIG. 4.4 – Représentation simultanée dans le premier plan sur l'exemple de Cohen.

Pour l'interprétation, il peut être utile à partir des nuages de points d'en déduire les relations d'indépendance et l'inertie totale et des axes. Nous reprenons les principaux cas étudiés dans [LMP95] sur la figure 4.5. Ainsi une inertie faible du nuage entraîne un nuage concentré autour du centre de gravité tandis qu'une inertie forte donne un nuage dilaté. L'indépendance des variables donne une forme sphérique au nuage, ce qui entraîne aucune direction privilégiée pour les axes, l'inertie des axes est donc dans ce cas faible. Au contraire l'existence d'une dépendance provoque un étirement du nuage dans une direction donnée.

Lorsque les nuages de points sont scindés en plusieurs sous-nuages, il est possible de réorganiser les données du tableau en ordonnant les coordonnées des lignes et des colonnes de façon à regrouper les fréquences nulles (*cf.* figure 4.6). Ceci permet alors d'étudier les sous-nuages indépendamment en considérant les parties du tableau correspondant.

Généralement, l'interprétation se limite aux premiers facteurs, nous considérons ainsi une approximation du tableau initial. Les calculs de reconstruction de l'analyse factorielle s'appliquent ici. Il est possible de montrer que :

$$f_{ij} - f_{i\bullet} f_{\bullet j} = f_{i\bullet} f_{\bullet j} \sum_{s \in S} \frac{F_s(i) G_s(j)}{\sqrt{\lambda_s}}. \tag{4.19}$$

Cette formule présente la décomposition de l'écart du tableau relativement à l'hypothèse d'indépendance en une somme de tableaux dont chacun ne dépend que d'un couple de facteurs (F_s, G_s) d'un même rang. Elle décompose ainsi la liaison des deux variables en éléments simples. En effet, chaque terme $f_{i\bullet} f_{\bullet j} F_s(i) G_s(j)$ représente la liaison simple entre

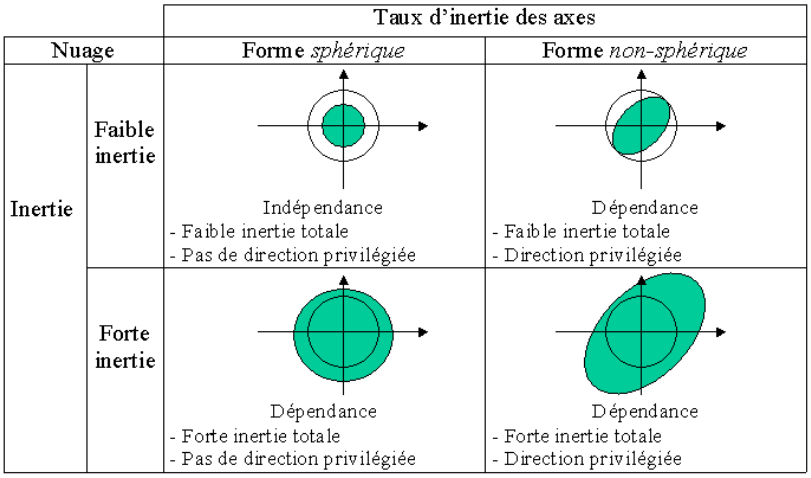


FIG. 4.5 – Inertie et dépendance.

les modalités i et j. Ainsi, si $F_s(i)$ et $G_s(j)$ sont du même signe, la case (i, j) du tableau exprime une attirance, sinon elle exprime une répulsion. L'attirance et la répulsion seront d'autant plus grande que la valeur absolue du produit $F_s(i)G_s(j)$ est grande.

Puisque le tableau est approché, lorsqu'une partie seulement est considérée pour l'analyse, il est important d'employer des indicateurs pour l'interprétation. Ceux utilisés pour l'AFC sont les mêmes que ceux de l'ACP que nous avons vu à la section 3.4 du chapitre précédent. Nous pouvons donc étudier la qualité de représentation d'un élément par un axe ou un plan. La qualité de représentation d'une ligne par un axe s est donnée par le rapport de l'inertie projetée du point sur l'axe s par l'inertie totale du point :

$$\frac{f_{i\bullet}F_s(i)^2}{f_{i\bullet}d^2(G_I, i)}, \tag{4.20}$$

et la qualité de représentation d'une ligne par un plan défini par les axes s et t est donnée par :

$$\frac{f_{i\bullet}F_s(i)^2}{f_{i\bullet}d^2(G_I, i)} + \frac{f_{i\bullet}F_t(i)^2}{f_{i\bullet}d^2(G_I, i)}. \tag{4.21}$$

La qualité de représentation d'un nuage par un plan est mesurée par le rapport de l'inertie projetée du nuage sur l'axe s et de l'inertie totale du nuage :

$$\frac{\lambda_s}{\sum_{s \in S} \lambda_s}, \tag{4.22}$$

multipliée par 100, elle représente le pourcentage d'inertie.

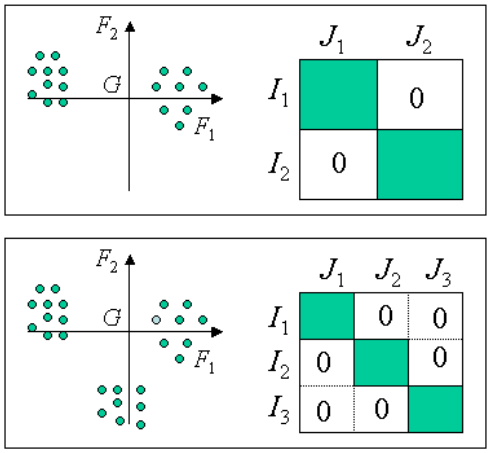


FIG. 4.6 – Relation entre la forme du nuage de points et le tableau.

Enfin la contribution d'un élément à l'inertie d'un axe est mesurée par le rapport de l'inertie du point et de l'inertie du nuage. Lorsque l'élément est une ligne, la contribution à l'inertie d'un axe s est donnée par :

$$\frac{f_{i\bullet} F_s(i)^2}{\lambda_s}, \tag{4.23}$$

et dans le cas d'un plan formé des axes s et t :

$$\frac{f_{i\bullet} \left[F_s(i)^2 + F_t(i)^2\right]}{\lambda_s + \lambda_t}. \tag{4.24}$$

Il est aussi possible, comme pour l'ACP, d'ajouter des éléments supplémentaires, illustratifs qui sont projetés sur les plans étudiés. Leur utilisation pour l'AFC est plus fréquente que pour l'ACP, car il peut y avoir beaucoup de variables pour une étude donnée qui ne sont pas considérées dans cette analyse. Les projections sur les axes factoriels des profils-lignes ou des profils-colonnes de ces éléments n'interviennent pas dans les calculs de ces axes.

Pour une bonne interprétation des plans de projection en AFC, nous proposons de suivre le plan suivant :

- Choisir le nombre d'axes de projection à étudier. Ce choix peut se faire par la même approche que celle décrite pour l'ACP.
- Etudier les valeurs propres. Les valeurs propres proches de 1 traduisent une forte liaison entre les lignes et les colonnes.
- Etudier la contribution des lignes et des colonnes de la même façon que pour l'ACP.
- Etudier les coordonnées des éléments actifs :
  - ceux qui présentent une forte contribution,
  - les extrêmes avec une forte qualité de représentation (pour qualifier le facteur).

4.4 Conclusion

Nous résumons l'AFC en neuf étapes illustrées par la figure 4.7 :

- 1 : Cette première étape donne le tableau de contingence des modalités communes aux deux variables. Les lignes et les colonnes jouent des rôles symétriques.
- 2 : Cette deuxième étape modifie le tableau en fréquences. Ces fréquences font apparaître des lois de probabilités.
- 3 : Nous considérons ici le tableau comme une juxtaposition de lignes après transformation en divisant par $f_{i\bullet}$. Ces lignes sont appelées profil-lignes et peuvent être interprétées comme des probabilités conditionnelles.
- 4 : Nous considérons ici le tableau comme une juxtaposition de colonnes après transformation en divisant par $f_{\bullet j}$. Ces colonnes sont appelées profil-colonnes et peuvent être interprétées comme des probabilités conditionnelles.
- 5 : Les profils-lignes qui constituent le nuage N_I sont projetés dans l'espace $\mathbb{R}^J$. Le nuage N_I se situe dans un hyperplan H_I. Le nuage N_I est analysé par rapport au centre de de gravité G_I qui constitue un profil moyen.
- 6 : Les profils-colonnes qui constituent le nuage N_J sont projetés dans l'espace $\mathbb{R}^I$. Le nuage N_J se situe dans un hyperplan H_J. Le nuage N_J est analysé par rapport au centre de gravité G_J qui constitue un profil moyen.
- AF : Analyse Factorielle. Elle permet de mettre en évidence une suite de directions orthogonales, d'étudier les projections en 7 et 8 en fonction de leurs proximités entre elles et par rapport à l'origine qui correspond à un profil moyen.
- 7 : Cette étape consiste en la projection du nuage N_I sur le premier plan factoriel. Les distances correspondent à des ressemblances entre les modalités.
- 8 : Cette étape consiste en la projection du nuage N_J sur le premier plan factoriel. Les distances correspondent à des ressemblances entre les modalités.
- Relations de transition : ces relations expriment les résultats d'une AF en fonction des résultats de l'autre.
- 9 : Les relations de transition permettent des interprétations simultanées des axes. Cette représentation simultanée facilite l'interprétation. Attention toute association entre un point-ligne et un point-colonne suggérée par une proximité doit être contrôlée sur le tableau.

L'ACP et l'AFC sont différentes en plusieurs points, elles fournissent des éclairages complémentaires. L'AFC est une méthode puissante pour synthétiser et résumer de vastes tableaux de contingence. En pratique elle est appliquée à beaucoup d'autres tableaux, notamment les tableaux individus-variables. Les individus sont alors considérés comme une variable.

Dans le cas de tableaux de contingence, le principal objectif de cette analyse est de dégager les liaisons entre deux variables. L'analyse des correspondances multiples que nous exposons dans le chapitre suivant permet l'étude des liaisons entre plus de deux variables.

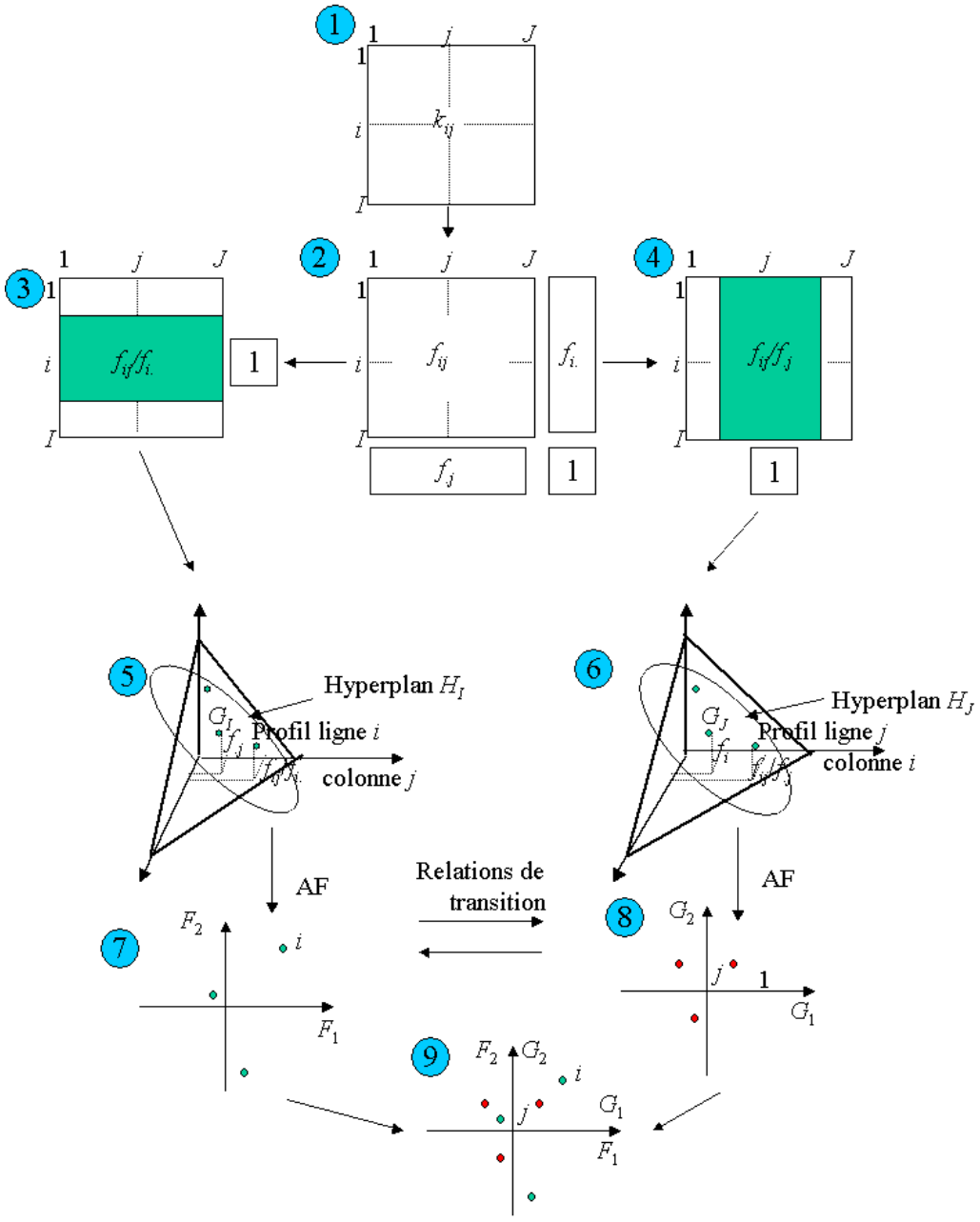


FIG. 4.7 – Résumé de l'AFC.

Chapitre 5

Analyse des Correspondances Multiples

5.1 Introduction

L'AFC peut se généraliser de plusieurs manières dans le cas où non plus deux variables sont mises en correspondance, mais deux ensembles de variables. La généralisation la plus simple et la plus employée est l'analyse des correspondances multiples. Nous la notons dans ce qui suit ACM. Cette analyse a particulièrement été étudiée par B. Escofier [EP90] et J.P. Bezécri [Ben80b].

5.1.1 Les domaines d'application

Cette analyse très simple est non plus adaptée aux tableaux de contingence de l'AFC, mais aux tableaux *disjonctifs complets* que nous décrivons ci-dessous. Ces tableaux sont des tableaux logiques pour des variables codées. Les propriétés de tels tableaux font de l'ACM une méthode spécifique aux règles d'interprétation des représentations simples. Elle permet donc l'étude des liaisons entre plus de deux variables qualitatives, ce qui étend le spectre d'étude de l'AFC.

L'ACM est donc très bien adaptée au traitement d'enquêtes lorsque les variables sont qualitatives (ou rendues qualitatives). Il est également possible de n'appliquer cette méthode plusieurs fois en ne prenant en compte que quelques variables.

5.1.2 Les données

L'ACM permet l'étude de tableaux décrivant une population de I individus et J variables qualitatives. Une variable qualitative (ou nominale) peut être décrite par une application de l'ensembles des I individus dans un ensemble fini non structuré, par exemple non ordonné. Ces variables qualitatives peuvent être codées par un *codage condensé* qui attribue une valeur à chaque modalité. Par exemple les modalités pour la couleur d'un vin peuvent être 1 pour le rouge, 2 pour le blanc et 3 pour le rosé. Les données peuvent donc être représentées sous la forme d'une matrice $\mathbf{X}$ décrite par le tableau 5.1, où I représente à la fois le nombre d'individus et l'ensemble des individus $I = \{1, \cdots, I\}$, J représente

		Variables			
		1	.....	j	..... J
Individus	1				
	⋮			⋮	
	i	.....		x_{ij}	.....
	⋮			⋮	
	I				

Tab. 5.1 – Représentation des données sous forme de codage condensé pour l'ACM.

à la fois le nombre de variables et l'ensemble des variables $J = \{1, \cdots, J\}$ et x_{ij} est le codage condensé de l'individu i pour la variable j.

Les x_{ij} représentant une codification, en prendre la moyenne n'a aucun sens. Ces données ne peuvent donc pas être traitées par l'ACP ou l'AFC précédemment étudiées. Ce tableau présente donc des spécificités dont l'analyse factorielle doit tenir compte par une méthode spécifique.

5.1.3 Les objectifs

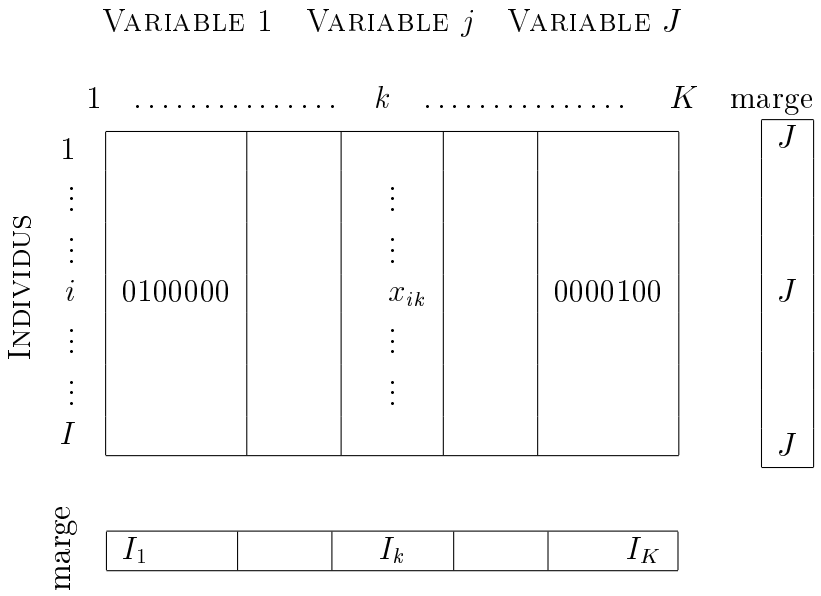
Les objectifs que cette méthode spécifique, l'ACM, doit remplir sont les mêmes que ceux de l'ACP ou de l'AFC. Il s'agit d'obtenir une typologie des lignes et des colonnes et relier ces deux typologies. Nous aurons ici trois familles d'éléments à étudier, les individus, les variables et les modalités des variables. Cette étude se fait par la définition de ressemblances et liaisons pour ces trois familles que nous détaillons dans la section suivante. Afin d'établir un bilan des ressemblances entre individus, comme en ACP nous cherchons à répondre à des questions du type :

- Quels sont les individus qui se ressemblent ?
- Quelles sont ceux qui sont différents ?
- Existe-t-il des groupes homogènes d'individus ?
- Est-il possible de mettre en évidence une typologie des individus ?

Les mêmes types de questions se posent pour les variables et les modalités.

5.2 Principe de l'ACM

Le principe de base de l'ACM repose dans un premier temps sur une transformation des données du tableau 5.1 pour modifier la codification en nombres binaires. L'analyse applique ensuite le même principe que l'AFC, en transformant ce tableau disjonctif complet ainsi obtenu en profils-lignes et en profils-colonnes. La distance du χ^2 est également employée pour définir les liaisons.



Tab. 5.2 – Représentation des données sous forme de codage condensé pour l'ACM.

5.2.1 La transformation des données

Une autre représentation du tableau 5.1 est le tableau disjonctif complet. Il représente les individus en ligne, alors que les colonnes représentent les modalités des variables (et non plus les variables) (*cf.* tableau 5.2). Ainsi, à l'intersection de la ligne i avec la colonne k, la valeur x_{ik} vaut 1 si l'individu i possède la modalité k et 0 sinon. Ce tableau porte le nom de *disjonctif complet*, car l'ensemble des valeurs x_{ik} d'un même individu pour les modalités d'une même variable, comporte la valeur 1 une fois (complet) et une fois seulement (disjonctif). Chaque modalité k est relié à une variable j. Nous avons ainsi trois familles d'éléments les individus, les variables et les modalités.

Notons K_j le nombre des modalités de la variable j et également l'ensemble des modalités de cette variable $K_j = \{1, \cdots, K_j\}$. Ainsi $K = \sum_{j \in J} K_j$ est à la fois le nombre des modalités toutes variables confondues et l'ensemble $K = \{1, \cdots, K\}$. Nous avons donc les égalités suivantes :

$$\sum_{k \in K_j} x_{ik} = 1, \forall (i, j), \tag{5.1}$$

$$\sum_{k \in K} x_{ik} = J, \forall i, \tag{5.2}$$

$$\sum_{i \in I} x_{ik} = I_k, \forall k, \tag{5.3}$$

	couleur	origine	appréciation
Individu 1	2	1	4
Individu 2	2	1	3
Individu 3	2	2	1
Individu 4	3	2	3
Individu 5	1	1	2
Individu 6	3	1	2
Individu 7	3	2	1
Individu 8	1	2	3

TAB. 5.3 – Exemple du vin : tableau initial.

et

$$\sum_{k \in K_j} I_k = I, \forall j. \tag{5.4}$$

Les propriétés intéressantes de l'ACM sont essentiellement dues aux propriétés des tableaux disjonctifs complets. Notons surtout que c'est un tableau binaire dont les lignes sont de sommes constantes à J et dont les colonnes sont regroupées par paquet correspondant à une variable avec pour somme par ligne égale à 1.

Exemple 5.2.1 Pour une meilleure compréhension de cette transformation, nous pouvons l'illustrer par un exemple. Nous supposons avoir des données issues d'une enquête sur l'appréciation du vin. Nous pouvons considérer trois variables : la couleur, l'origine et l'appréciation de l'individu. Nous reprenons les trois modalités rouge, blanc et rosé pour la couleur codées respectivement par 1, 2 et 3. Nous considérons uniquement deux origines : Bordeaux et Côte du Rhône, codées par 1 et 2, et quatre modalités pour l'appréciation : mauvais, moyen, bon et très bon codées respectivement par 1, 2, 3 et 4. Nous avons ainsi trois variables ($J = 3$) et neuf modalités ($K = 9$). Les résultats de l'enquête fictive sont donnés dans le tableau 5.3. Ainsi, par exemple l'individu 5 a moyennement apprécié un Bordeaux rouge. Le tableau disjonctif complet 5.4 déduit ce tableau initial présente les mêmes informations. Ainsi, l'individu 5 présente les modalités : rouge, Bordeaux, moyen.

Lorsque le nombre de variables est réduit à deux ($J = 2$), les données peuvent être représentées sous la forme d'un tableau de contingence, comme dans l'AFC, mettant ainsi en correspondance les modalités des deux variables. Il est possible d'étendre ce tableau à une hypertable de contingence lorsque $J = 3$ (*cf.* figure 5.1), où K_1 (respectivement K_2 et K_3) représente le nombre de modalités de la première (respectivement deuxième et troisième) variable et I_{klm} est le nombre d'individus possédant les modalités k (de la première variable), l (de la deuxième variable) et m (de la troisième variable). Cependant dès que J augmentent le nombre de cases devient très important et l'hypertable est alors difficile à manier et à représenter.

	Couleur			Origine		Appréciation			
	Rouge	Blanc	Rosé	Bordeaux	Côtes du Rhône	Mauvais	Moyen	Bon	Très Bon
Ind. 1	0	1	0	1	0	0	0	0	1
Ind. 2	0	1	0	1	0	0	0	1	0
Ind. 3	0	1	0	0	1	1	0	0	0
Ind. 4	0	0	1	0	1	0	0	1	0
Ind. 5	1	0	0	1	0	0	1	0	0
Ind. 6	0	0	1	1	0	0	1	0	0
Ind. 7	0	0	1	0	1	1	0	0	0
Ind. 8	1	0	0	0	1	0	0	1	0

TAB. 5.4 – Exemple du vin : tableau disjonctif complet.

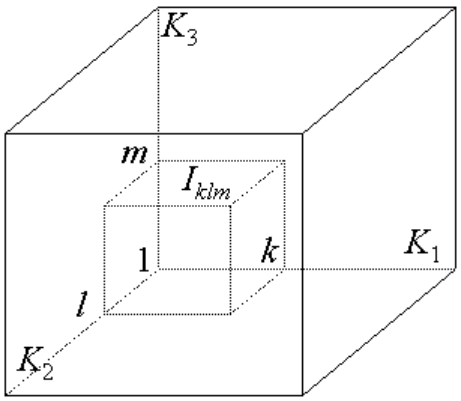
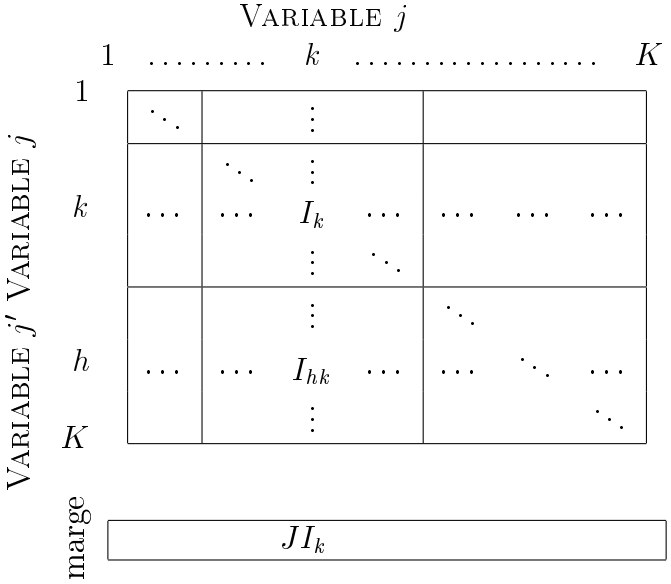


FIG. 5.1 – Hypertable de contingence pour $J = 3$.

L'hypertable étant problématique pour un grand nombre de variables, pour généraliser le tableau de contingence à deux variables, il est possible de considérer les tableaux de contingence entre variables prises deux à deux. Nous obtenons ainsi une juxtaposition de tableaux de contingence. Un tel tableau est appelé *tableau de Burt* du nom de son auteur (*c.f.* tableau 5.5).

Comme en ACP nous cherchons une typologie des individus. La notion de *ressemblance* est déterminée par le nombre de modalités en commun. Par exemple dans une enquête d'opinion, il est important de mettre en évidence une classe d'individus déterminées par des variables. Pour l'étude des variables deux points de vue s'offrent à nous. Nous pouvons caractériser les liaisons entre deux variables qualitatives en considérant les modalités, ou encore chercher à résumer l'ensemble des liaisons à l'aide de quelques variables numériques qui synthétisent l'ensemble des variables. Les catégories socio-professionnelles peuvent ainsi résumer une variable "statut social". La richesse de l'ACM provient de l'étude d'une troisième classe d'éléments, les modalités. De la même façon que les individus, nous



TAB. 5.5 – Représentation des données sous forme du tableau de Burt.

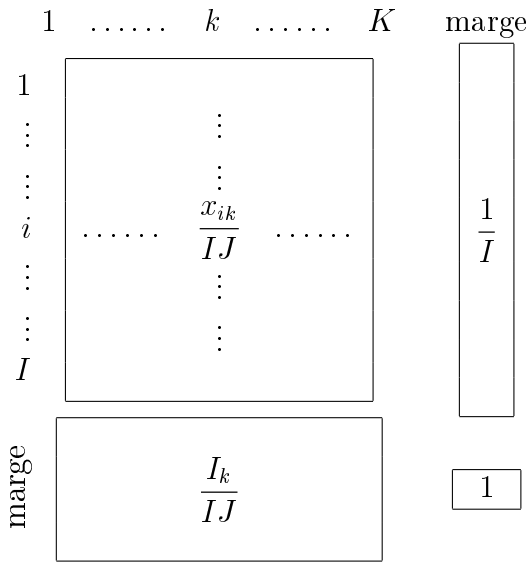
pouvons chercher à établir un bilan des ressemblances entre modalités. Les ressemblances entre modalités peuvent être définies à partir du tableau disjonctif complet, ou bien à partir du tableau de Burt. Dans le premier cas une colonne est une variable indicatrice, ainsi deux modalités se ressemblent si elles sont présentes ou absentes chez beaucoup d'individus. Dans le cas du tableau de Burt, une ligne ou une colonne correspond à une classe d'individus, ainsi deux modalités se ressemblent si elles s'associent beaucoup ou peu aux mêmes modalités. Ces deux points de vue aboutissent aux mêmes résultats. L'ACM peut donc être vue comme une AFC du tableau disjonctif complet ou comme une AFC du tableau de Burt.

La richesse apportée par ces trois éléments, ne doit pas occulter l'unicité du tableau, et donc des conclusions parfois redondantes. Il sera donc préféré l'étude des modalités en priorité.

5.2.2 L'analyse factorielle des correspondances du tableau disjonctif complet

Comme pour l'AFC, nous allons considérer le tableau disjonctif complet en profils-lignes et en profils-colonnes. Pour se faire nous modifions ce tableau pour considérer les fréquences (*cf.* tableau 5.6). Les fréquences f_{ik} sont données par $\frac{x_{ik}}{IJ}$. De plus les marges sont données par :

$$f_{i\bullet} = \sum_{k \in K} \frac{x_{ik}}{IJ} = \frac{1}{I}, \tag{5.5}$$



TAB. 5.6 – Mise en fréquences du tableau disjonctif complet.

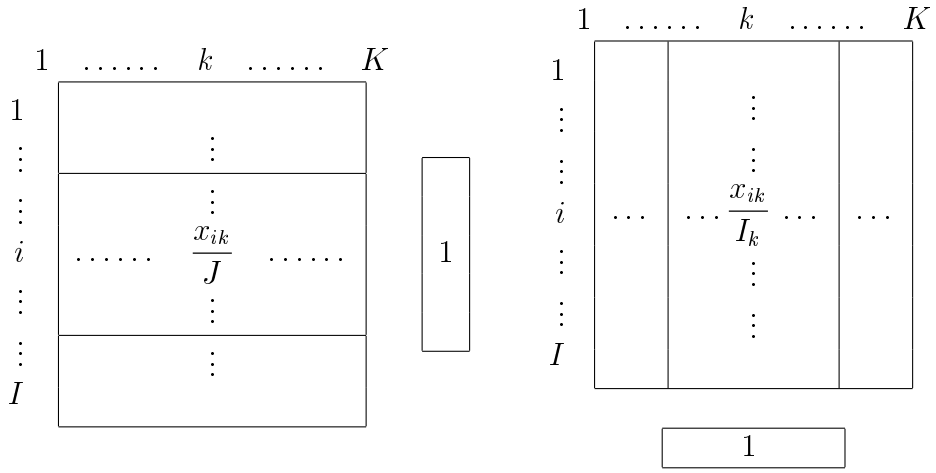
et

$$f_{\bullet k} = \sum_{i \in I} \frac{x_{ik}}{IJ} = \frac{i_k}{IJ}. \tag{5.6}$$

Une fois les fréquences calculées, il faut considérer le tableau en profils-lignes et profils-colonnes (*cf.* tableau 5.7). Ainsi le tableau est de nouveau modifié de façon à ce que pour les profils-lignes la marge des lignes soit 1 et pour les profils-colonnes la marge des colonnes soit 1. Ainsi chaque case est composée respectivement de $\frac{x_{ik}}{J}$ et $\frac{x_{ik}}{I_k}$.

L'analyse des nuages

Chaque individu du nuage des individus N_I est représenté par les modalités qu'il possède. La marge étant constante, la transformation en profils-lignes ne modifie en rien les données. Ainsi le nuage N_I appartient à un hypercube noté H_I d'arrête $\frac{1}{J}$, puisque le profil d'une ligne est soit 0 soit $\frac{1}{J}$ (*cf.* figure 5.2). Un individu i est un point de $\mathbb{R}^K$ qui a pour coordonnée sur l'axe k la valeur $\frac{x_{ik}}{J}$ avec un poids identique pour chaque individu (car la marge est constante) de $\frac{1}{I}$. Le barycentre G_I du nuage N_I a pour coordonnée $\frac{I_k}{IJ}$ sur l'axe k. La ressemblance entre deux individus est définie par les modalités de chacun des individus. Si les deux individus présentent globalement les mêmes modalités, alors ils se ressemblent. La distance qui caractérise cette ressemblance entre deux individus i et l



TAB. 5.7 – Les profil-lignes et profil-colonnes pour l'ACM.

est définie par :

$$d^2(i,l) = \sum_{k \in K} \frac{IJ}{I_k} \left(\frac{x_{ik}}{J} - \frac{x_{lk}}{J} \right)^2 = \frac{1}{J} \sum_{k \in K} \frac{I}{I_k} \left(x_{ik} - x_{lk} \right)^2 . \tag{5.7}$$

Cette expression est remarquable car $(x_{ik} - x_{lk})^2 = 1$ si un seul individu possède la modalité k et 0 sinon. Cette distance croît logiquement avec le nombre de modalités qui diffèrent pour les individus i et l, ce qui est recherché. Le poids de la modalité k dans la distance est l'inverse de sa fréquence : $\frac{I}{I_k}$. Ainsi si un individu possède une modalité rare, il sera éloigné de tous les autres individus et du centre de gravité.

Chaque modalité peut être représentée par le profil-colonne, c'est-à-dire par les valeurs prises par tous les individus pour la modalité considérée. Ainsi une modalité k est un point de l'espace $\mathbb{R}^I$ et a pour coordonnée $\frac{x_{ik}}{I_k}$ sur l'axe i avec un poids constant de $\frac{I_k}{IJ}$ (*cf.* figure 5.3). Le barycentre G_K du nuage N_K a pour coordonnée $\frac{1}{I}$ sur l'axe i. Ainsi le nuage N_K appartient à l'hypercube d'arrête $\frac{1}{I_k}$, noté H_K, puisque le profil d'une colonne est soit 0 soit $\frac{1}{I_k}$.

La ressemblance entre deux modalités k et h est donnée par la distance :

$$d^2(k,h) = \sum_{i \in I} I \left(\frac{x_{ik}}{I_k} - \frac{x_{ih}}{I_h} \right)^2 . \tag{5.8}$$

En notant que $(x_{ik})^2 = x_{ik}$ qui ne prennent que les valeurs 1 ou 0, cette distance peut s'écrire :

$$d^2(k,h) = \frac{I}{I_k I_h} \left(I_k + i_h - 2 \sum_{i \in I} x_{ik} x_{ih} \right), \tag{5.9}$$

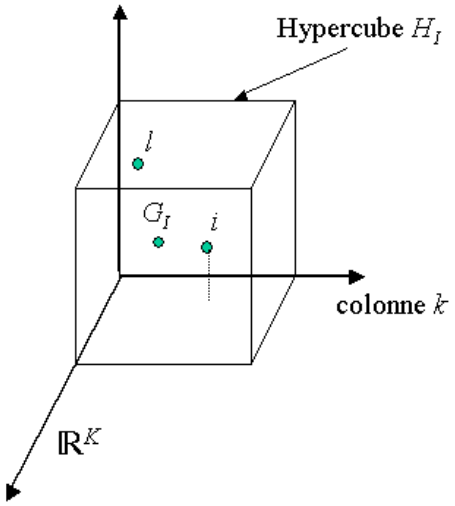


FIG. 5.2 – Représentation du nuage des individus N_I dans l'espace $\mathbb{R}^K$.

ce qui est le nombre d'individus possédant une et une seule des deux modalités h ou k multiplié par $\frac{I}{I_k I_h}$. Cette distance croît donc avec le nombre d'individus possédant une et une seule des deux modalités k et h et décroît avec l'effectif de chacune de ces modalités. Ainsi, par construction, deux modalités d'une même variable sont éloignées l'une de l'autre (puisqu'elles ne peuvent pas être possédées par le même individu). Deux modalités possédées par exactement les mêmes individus sont confondues, tandis que les modalités rares sont éloignées de toutes les autres et du centre de gravité G_K.

La représentation simultanée

Il est possible, comme pour l'AFC, de représenter simultanément les deux nuages N_I et N_K grâce à la dualité existant entre ces deux nuages. Avec les notations données par le schéma de dualité sur la figure 5.4, les relations de transitions s'écrivent :

$$\begin{cases} F_S(i) = \dfrac{1}{\sqrt{\lambda_S}} \displaystyle\sum_{k \in K} \frac{x_{ik}}{J} G_S(k) \\ \\ G_S(k) = \dfrac{1}{\sqrt{\lambda_S}} \displaystyle\sum_{i \in I} \frac{x_{ik}}{I_k} F_S(i) \end{cases} \tag{5.10}$$

où $F_S(i)$ représente la projection de la ligne i sur l'axe de rang S de N_I, tandis que $G_S(k)$ représente la projection de la ligne k sur l'axe de rang S de N_K. λ_S représente toujours la valeur commune de l'inertie associée à chacun de ces axes de rang S des nuages N_I et N_K. Ces relations s'interprètent facilement car les x_{ik} ne prennent que les valeurs 0 ou 1. Ainsi $F_S(i)$ est placé au coefficient $\frac{1}{\sqrt{\lambda_S}}$ près, au barycentre des modalités que

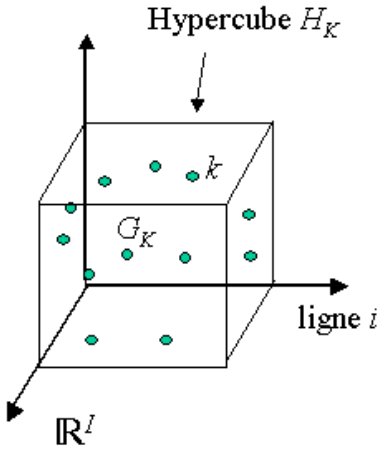


FIG. 5.3 – Représentation du nuage des modalités N_K dans l'espace $\mathbb{R}^I$.

l'individu i possède. Inversement, $G_S(k)$ est placé au coefficient $\frac{1}{\sqrt{\lambda_S}}$ près, au barycentre des individus qui possèdent la modalité k. Les modalités peuvent ainsi être vues comme barycentre d'une classe d'individus, ou comme une modalité d'une variable.

Il faudra cependant tenir compte lors de l'interprétation, que malgré cette équivalence entre les facteurs $F_S(i)$ et $G_S(k)$, les modalités et les individus n'évoluent pas dans le même espace ($\mathbb{R}^K$ pour les premiers et $\mathbb{R}^I$ pour les seconds).

La représentation simultanée n'est pas toujours facile à interpréter, car en pratique le nombre d'individus et de modalités pouvant être grand, le graphique devient vite encombré. Elle permet cependant de bien caractériser les répartitions et les classes d'individus.

5.2.3 L'analyse factorielle des correspondances du tableau de Burt

Nous avons vu que l'ACM peut être vue comme une analyse factorielle des correspondances du tableau disjonctif complet ou encore du tableau de Burt. L'analyse à partir du tableau disjonctif complet fournit des représentations des barycentres de classes d'individus, cependant au lieu de calculer les axes d'inertie du nuage d'individus, puis de projeter les barycentres sur ces axes, nous pouvons analyser directement le nuage des barycentres obtenu par le tableau de Burt.

En fait, ces deux approches fournissent exactement les mêmes résultats. Sans détailler davantage cette approche, les transformations des données ainsi que les relations de transitions issues de l'analyse factorielle des correspondances du tableau de Burt sont données par exemple dans [Pag03] ou [LMP95].

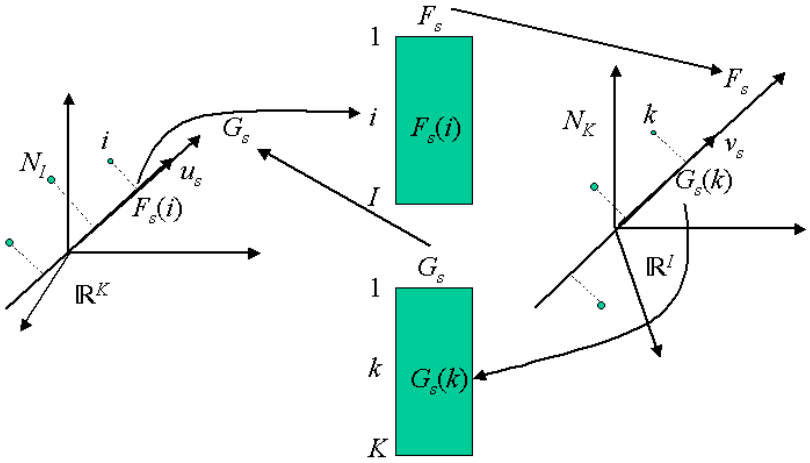


FIG. 5.4 – Schéma de dualité pour l'ACM.

5.2.4 Les variables quantitatives

Initialement prévue pour les variables qualitatives, l'ACM peut traiter également les variables quantitatives, sous condition qu'elles soient rendues qualitatives. Ceci a un double intérêt. Tout d'abord, rendre des variables quantitatives en variables qualitatives permet d'obtenir un tableau homogène et ainsi d'analyser l'ensemble de ces variables par une même analyse. Un autre intérêt est qu'une ACM sur des variables quantitatives codées en classe peut mettre en évidence des liaisons entre variables non linéaires, que l'ACP ne peut dévoiler. Or les liaisons non linéaires sont très fréquentes en pratique.

Pour se faire, il suffit de découper l'intervalle de variation en sous-intervalles qui définissent autant de modalités. Ainsi diminuer le nombre de classes, c'est regrouper des individus de plus en plus différents et augmenter le nombre de classes, c'est obtenir des classes plus nombreuses et à faible effectif. Il est préférable de garder un nombre inférieur à huit classes pour pouvoir espérer analyser ces classes correctement. Bien souvent quatre ou cinq classes suffisent. Trois classes peuvent par exemple être interprétées comme trois modalités mauvais, moyen et bon. Il faut également choisir correctement les classes, par exemple en regardant s'il n'existe pas de seuils pour la variable mesurée, déterminés par exemple par l'histogramme. Dans le cas où la variable possède une répartition homogène de ses valeurs, il est possible de faire un découpage systématique, par exemple avec des intervalles réguliers, ou encore avec un nombre d'individus identique dans chaque classe.

5.3 Interprétation

Nous avons vu que deux individus se ressemblent s'ils présentent globalement les mêmes modalités. Cette ressemblance se traduit par une proximité des individus dans l'espace $\mathbb{R}^K$ ou en pratique dans l'espace de projection choisi pour la représentation simultanée. De même si deux modalités d'une même variable sont proches dans l'espace

de projection, ceci se traduit par une ressemblance entre les groupes d'individus qui les ont choisies. La proximité de deux modalités de variables différentes s'interprète en terme d'*association*. Ainsi deux modalités de variables différentes s'associent si elles concernent globalement les mêmes individus. En fait ces modalités correspondent alors aux points moyens des individus.

En ce qui concerne la proximité entre modalités et individus, l'interprétation peut se faire en considérant les modalités comme barycentre de classe d'individus. Il est souvent nécessaire de se reporter au tableau de données pour vérifier les conclusions.

Nous rappelons que sur la représentation simultanée, les nuages N_I et N_K ne sont pas dans les mêmes espaces. Il est donc important d'avoir recours à des indicateurs sur la qualité de représentation. Les indicateurs pour l'interprétation de l'ACM sont les mêmes que ceux de l'ACP et de l'AFC déjà donnés, ici pour les individus et les modalités. Ainsi nous pouvons étudier la qualité de représentation de chaque individu et de chaque modalité par un axe ou par un plan. La contribution d'un individu ou d'une modalité à l'inertie d'un axe ou d'un plan doit aussi être considérée. La notion de variable doit également être prise en compte. Ainsi la contribution d'une variable à l'inertie d'un axe peut être obtenue en sommant les contributions des modalités de cette variable à l'inertie du même axe. Nous obtenons ainsi un indicateur de liaison entre la variable et le facteur.

Les éléments supplémentaires Les éléments supplémentaires ou illustratifs peuvent être des variables (et leurs modalités) ou bien des individus. Les variables supplémentaires permettent d'enrichir l'interprétation des axes sans avoir participées à leur construction. Une variable supplémentaire couramment employée est la variable qualitative obtenue par la classification hiérarchique (*cf.* chapitre 7). Les individus supplémentaires exclus de l'analyse peuvent être situés par rapport aux individus actifs, ou à des groupes d'individus actifs dans une optique de discrimination.

Il est aussi courant de regrouper les modalités de faible effectif (qui n'ont pas de signification statistique) pour ensuite les représenter en tant qu'éléments supplémentaires.

Afin de ne rien oublier pour l'interprétation des résultats, nous proposons de suivre le plan suivant :

- Définir le nombre de modalités des variables quantitatives, s'il y a des variables quantitatives intéressantes pour l'étude.
- Choisir le nombre d'axes de projection. Ce choix se fait toujours de la même façon que pour l'ACP ou l'AFC.
- Etudier les valeurs propres qui représentent l'inertie de chaque axe.
- Etudier la contribution des lignes et des modalités de la même façon que l'ACP.
- Etudier la contribution des variables en sommant les contributions des modalités d'une variable pour un facteur donné.
- Etudier les coordonnées des modalités et des individus actifs.
- Etudier les coordonnées des variables, des modalités et des individus supplémentaires s'il y en a.

5.4 Conclusion

Pour conclure ce chapitre, commençons par résumer l'ACM en dix étapes illustrées sur la figure 5.5 :

- 1 : Cette première étape donne le tableau des données une fois que les variables qualitatives sont codées de manière condensée. Les lignes représentent les individus et les colonnes les variables.
- 2 : Cette deuxième étape transforme le tableau de l'étape précédente en tableau disjonctif complet. Les lignes représentent toujours les individus, mais à présent les colonnes représentent les modalités. Cette deuxième étape peut également être la transformation du tableau de Burt. Dans ce cas, il y a symétrie entre les lignes et les colonnes qui représentent une classe d'individus.
- 3 : A partir de cette étape nous appliquons l'AFC. Nous transformons le tableau disjonctif complet en fréquences.
- 4 : Nous considérons ici le tableau comme une juxtaposition de lignes après transformation en multipliant par I. Ces lignes sont appelées les profils-lignes.
- 5 : Nous considérons ici le tableau comme une juxtaposition de colonnes après transformation en multipliant par $\frac{IJ}{I_k}$. Ces colonnes sont appelées profil-colonnes.
- 6 : Les profils-lignes qui constituent le nuage N_I sont projetés dans l'espace $\mathbb{R}^K$. Le nuage N_I se situe dans un hypercube H_I.
- 7 : Les profils-colonnes qui constituent le nuage N_K sont projetés dans l'espace $\mathbb{R}^I$. Le nuage N_K se situe dans un hypercube H_K.
- AF : Analyse Factorielle. Elle permet de mettre en évidence une suite de directions orthogonales, d'étudier les projections en 8 et 9 en fonction de leurs proximités entre elles et par rapport à l'origine qui correspond à un profil moyen.
- 8 : Cette étape consiste en la projection du nuage N_I sur le premier plan factoriel. Les distances correspondent à des ressemblances entre les individus.
- 9 : Cette étape consiste en la projection du nuage N_K sur le premier plan factoriel. Les distances correspondent à des ressemblances entre les modalités.
- Relations de transition : ces relations expriment les résultats d'une AF en fonction des résultats de l'autre. Ce sont des relations barycentriques.
- 10 : Les relations de transition permettent des interprétations simultanées des axes. Cette représentation simultanée facilite l'interprétation. Attention toute association entre un point-ligne et un point-colonne suggérée par une proximité doit être contrôlée sur le tableau.

L'ACM est donc une analyse factorielle qui permet l'étude de plusieurs variables qualitatives, de ce fait elle est une généralisation de l'AFC. Elle est donc applicable aux tableaux de variables qualitatives, mais aussi quantitatives après construction de classes à partir de celles-ci. Le fait de pouvoir interpréter l'ACM de plusieurs façons rend cette méthode très riche et d'emploi facile. Elle peut être très complémentaire de l'ACP et bien sûr des méthodes de classification.

Les méthodes de classification permettent de regrouper les individus en classes selon leurs ressemblances. Deux types d'approches sont possibles soit nous considérons des données sur lesquelles nous connaissons les différentes classes et nous tentons d'affecter un nouvel individu dans une des classes connues, soit nous n'avons aucun *a priori* sur les classes. Nous allons étudier ces deux types de classification dans les chapitres suivants.

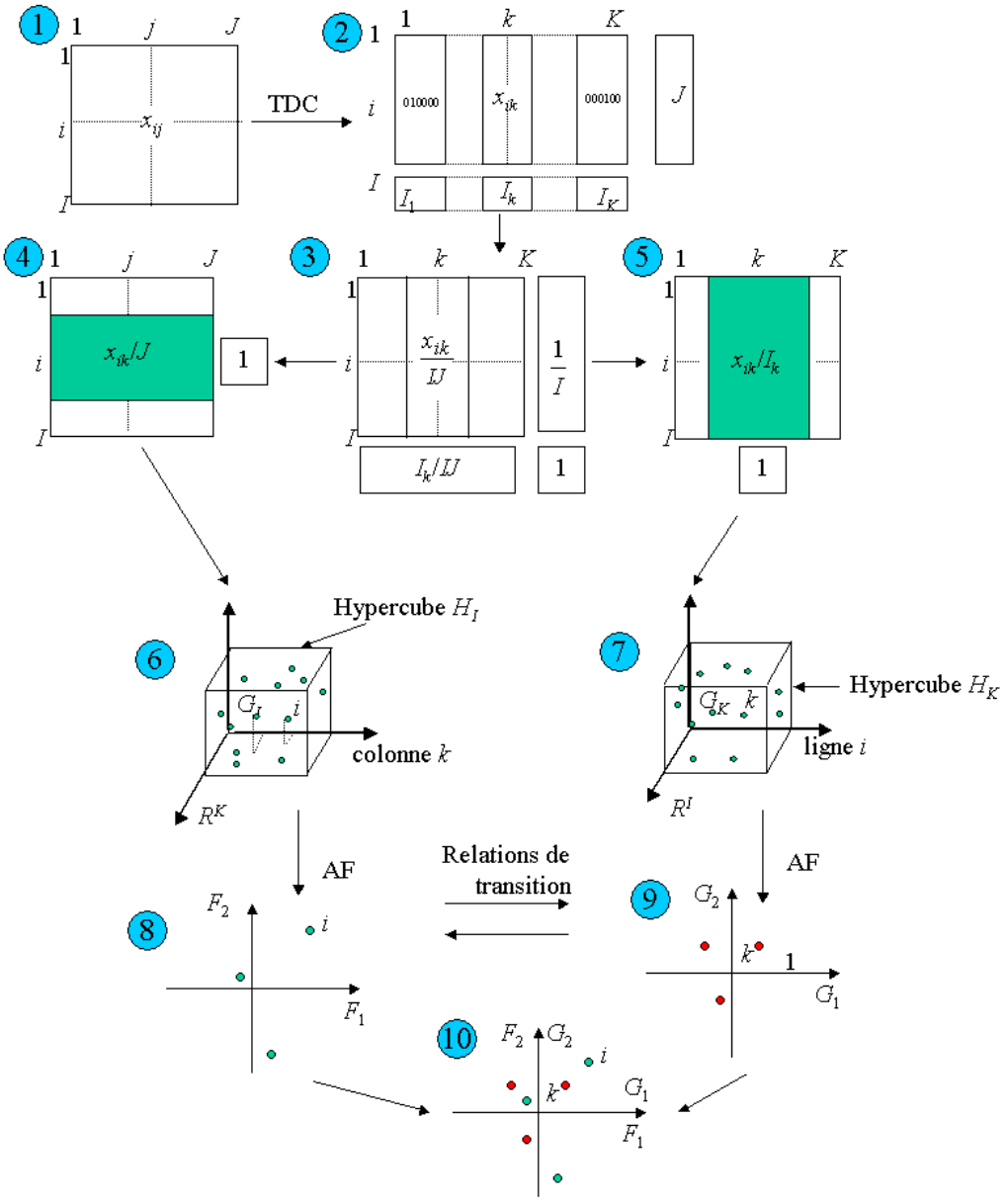


FIG. 5.5 – Résumé de l'ACM.

Chapitre 6

Analyse Factorielle Discriminante

6.1 Introduction

L'analyse factorielle discriminante est une des nombreuses méthodes de l'analyse discriminante. Sous ce nom sont regroupées des méthodes de classification qui nécessitent une connaissance des classes préexistantes. Dans le domaine de la reconnaissance des formes c'est ce qui est appelé classification supervisée ou encore à apprentissage supervisé. Parmi ces méthodes peuvent être comptés la régression logistique (méthode semi-paramétrique), les k-plus proches voisins, les arbres de décisions (méthode non paramétrique qui favorise le dialogue homme-machine) ou encore des méthodes issues de l'intelligence artificielle souvent considérées comme des "boîtes noires" telles que le perceptron multicouche et les autres réseaux de neurones, les chaînes de Markov [Kun00] ou les machines à vecteurs de support [Vap99]. Un aspect important de la classification supervisée est l'échantillonnage des données pour réaliser l'apprentissage. Différentes approches d'échantillonnage existent telles que la technique de Jackknife, du bootstrap ou de la validation croisée [LMP95], [Sap90], nous ne les détaillons pas ici.

La plupart des méthodes qui ne sont pas issues de l'intelligence artificielle peuvent être décrites par deux étapes :

- l'étape de *discrimination* qui cherche à déterminer sur les données d'apprentissage une fonction qui discrimine au mieux les données,
- l'étape de *classement* qui cherche à affecter une nouvelle donnée à une classe, à l'aide de la fonction établie dans l'étape précédente.

La *régression logistique* consiste à exprimer les probabilités *a posteriori* d'appartenance à une classe $p(C/\mathbf{x})$ comme une fonction de l'observation [Sap90] [Cel03]. Bien souvent c'est la régression linéaire qui est employée, *i.e.* qu'il faut déterminer les coefficients β tels que :

$$\ln\left(\frac{p(C/\mathbf{x})}{1-p(C/\mathbf{x})}\right) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i. \tag{6.1}$$

Il est donc nécessaire d'estimer les paramètres des lois de probabilité, en supposant connue cette loi. Selon la loi retenue, il est possible de traiter des variables quantitatives, ou

binaires. La fonction de discrimination est ainsi définie, pour le classement d'un nouvel individu, la règle bayésienne peut être appliquée.

La *classification bayésienne* est une autre approche probabiliste qui suppose connues les probabilités *a priori* et les distributions des probabilités d'appartenance à chaque classe. Dans ce cas c'est une méthode optimale. En pratique, ces probabilités sont estimées à partir de données d'apprentissage. Nous présentons brièvement cette méthode très utilisée en classification, comme méthode de classement de l'analyse factorielle discriminante à la section 6.2.2.

Les *arbres de décision* sont des méthodes de discrimination, souvent employées pour la segmentation. La représentation sous forme d'arbres permet une interprétation rapide et aisée des résultats. La construction de l'arbre (*i.e.* l'étape de discrimination) est effectuée sur les données d'apprentissage, puis l'étape de classement peut être réalisé pour de nouveaux individus. L'idée de la construction est simple, et se décompose comme suit :

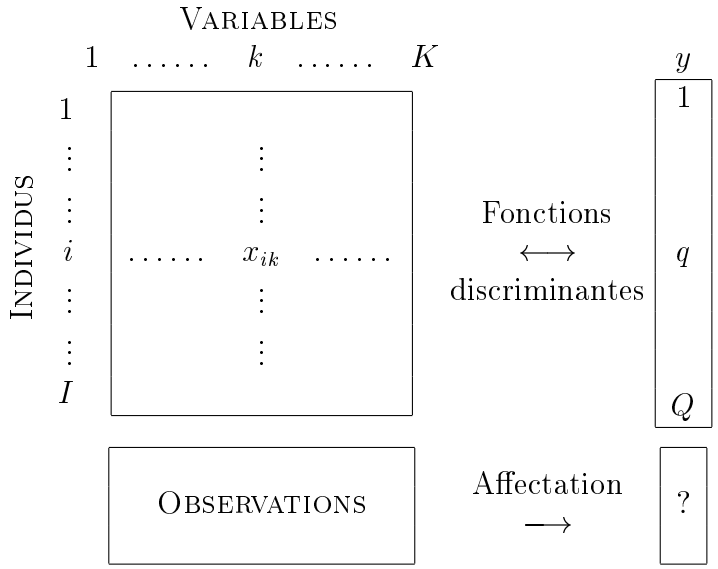
- chercher la variable qui produit la meilleure division (par exemple celle qui donne la variance intraclasse la plus faible),
- diviser en deux nœuds intermédiaires, les individus selon cette variable,
- chercher les variables qui produisent les meilleurs divisions des nœuds intermédiaires,
- poursuivre ainsi jusqu'à n'obtenir que des nœuds terminaux.

Cependant l'arbre optimal est difficile à déterminer. En effet, il faut définir un critère de division, un critère d'arrêt *i.e.* une règle pour déclarer si le nœud est terminal ou intermédiaire. De plus la complexité est importante pour des arbres à plus de deux branches (*i.e.* des arbres qui ne sont pas binaires). En outre, elle est difficilement généralisable si les données d'apprentissage sont peu représentatif de la réalité. La méthode CART (*Classification And Regression Tree*) qui est un cas particulier des arbres binaires possède une construction d'arbres aux propriétés intéressantes pour la segmentation qui résout en partie ces problèmes [BFRS93].

L'approche des k *plus proches voisins* repose sur l'idée simple d'attribuer un nouvel individu à la classe majoritaire parmi ses k plus proches voisins (individus de la base d'apprentissage les plus proches au sens d'une certaine distance). C'est donc une approche géométrique. Nous présentons plus en détails cette approche, comme méthode de classement de l'analyse factorielle discriminante à la section 6.2.2.

L'analyse factorielle discriminante est une méthode descriptive et prédictive fondée sur un modèle paramétrique. Elle est également appelée analyse linéaire discriminante (*Linear Analysis Discriminant* (LDA) en anglais). Nous conservons ici le nom d'analyse factorielle discriminante, et nous utilisons l'abréviation AFD. En effet, cette méthode peut être vu comme une analyse factorielle, car son aspect descriptif fait appel à des calculs d'axes principaux. C'est une méthode avant tout prédictive qui discrimine les individus selon des classes connues. Son aspect prédictif de classement de nouveaux individus peut en fait faire appel à d'autres méthodes de classification géométriques ou probabilistes.

L'analyse factorielle discriminante peut également être vu comme une analyse canonique particulière ou encore comme une extension de la régression multiple présentée par exemple dans [LMP95].



Tab. 6.1 – Représentation des données pour l'AFD.

6.1.1 Les domaines d'application

L'AFD est une approche très utilisée, et fait à présent partie de tout bon logiciel de statistique ou d'apprentissage. Les domaines d'application sont très nombreux pour résoudre des problèmes tels que l'aide au diagnostic (par exemple en médecine pour la prédiction de maladies), pour la prédiction de risques (par exemple en météorologie pour prédire un risque d'avalanche ou en finance pour prédire un comportement boursier), pour le contrôle de qualité (par exemple prévision de qualité d'un produit agro-alimentaire par des mesures) ou encore pour la reconnaissance des formes (par exemple en traitement d'images). C'est une méthode importante dans le métier d'ingénieurs puisque l'aspect essentiel de l'AFD (et des méthodes de l'analyse discriminante en général) est l'aide à la décision. Son intérêt vient également du fait qu'elle fournit des résultats *stables*, *i.e.* peu dépendants des données d'apprentissage et *robuste*, *i.e.* peu dépendants des hypothèses. Elle est ainsi considérée comme une approche de référence à laquelle sont souvent comparées les autres méthodes.

6.1.2 Les données

Nous disposons de I individus ou observations décrits par K variables et répartis en Q classes données par la variable nominale y (*cf.* tableau 6.1). Les Q classes sont *a priori* connues. La variable nominale y possède donc Q modalités. I représente à la fois le nombre d'individus et l'ensemble des individus $I = \{1, \cdots, I\}$, K représente à la fois le nombre de variables et l'ensemble des variables $K = \{1, \cdots, K\}$, et Q représente à la fois le nombre de modalités de la variable y et l'ensemble $Q = \{1, \cdots, Q\}$. x_{ik} est la valeur de la variable k pour l'individu i.

6.1.3 Les objectifs

A partir du tableau 6.1, nous constatons que deux objectifs se dessinent :
- Le premier objectif consiste à déterminer les fonctions linéaires discriminantes sur l'échantillon d'apprentissage, *i.e.* la combinaison linéaire des K variables explicatives dont les valeurs séparent au mieux les Q classes. Il s'agit donc d'une étape de *discrimination* des classes.
- Le second objectif consiste à déterminer la classe de nouveaux individus pour lesquels nous observons les valeurs des K variables explicatives. Cette étape est une étape d'*affectation* d'un nouvel individu dans une classe. Il s'agit d'un problème de *classement* par opposition au problème de *classification* qui est la construction de classes les plus homogènes possibles dans un échantillon.

Exemple 6.1.1 Supposons un service dans un hôpital qui comprend 500 patients. Dans ce service sont rencontrées essentiellement cinq pathologies. Il est aisé de réaliser une vingtaine d'examens et des analyses peu coûteuses. Cependant pour déterminer une des cinq pathologies il est nécessaire d'entreprendre des interventions très coûteuses. Les données sont ainsi constituées de 500 individus et 20 variables, de plus la variable nominale y est composée de cinq modalités. L'étape de discrimination tente de répondre à des questions du type : est-il possible de prévoir avec les vingt examens et analyses, les pathologies des 500 patients sans avoir recours à des interventions plus coûteuses ? Alors que l'affectation tente de répondre à des questions du type : Est-il possible de prédire la pathologie d'un nouveau patient en n'effectuant que les examens et analyses peu coûteux ?

En fait derrière ces deux questions il en existe une autre d'ordre plus général à laquelle tente de répondre l'analyse factorielle discriminante : Est-ce qu'un grand nombre de données d'accès facile peut contenir une information décrite par une appartenance à une classe, plus délicate à déterminer ?

6.2 Principe de l'AFD

6.2.1 La discrimination

L'idée du principe de la discrimination repose sur le fait que la discrimination visuelle est plus aisée si :
- les centres de gravité de chaque sous-nuage appartenant à une seule classe sont éloignés,
- chaque sous-nuage appartenant à une seule classe sont les plus homogènes possibles autour de ces centres de gravité.

Pour ce faire il faut maximiser les variances interclasses (entre les classes) et minimiser les variances intraclasses (à l'intérieur des classes). Nous parlons également de variances externes et internes.

La figure 6.1 représente un nuage N_I des individus partitionnés en trois classes dans l'espace $\mathbb{R}^K$. Notons I_q le nombre d'individus dans la classe q et l'ensemble des individus

de la classe q, $I_q = \{A, \cdots, I_q\}$. $\mathbf{G}$ représente le centre de gravité du nuage des individus dans $\mathbb{R}^K$, et $\mathbf{g}_q$ le centre de gravité de la partition des individus appartenant à la classe q. Le centre de gravité de la classe q est donné par le vecteur :

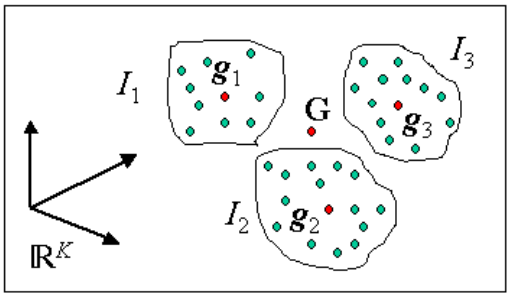


FIG. 6.1 – Représentation du nuage N_I des individus partitionnés dans l'espace $\mathbb{R}^K$.

$$\mathbf{g}_q = \frac{1}{I_q} \sum_{i \in I_q} \mathbf{x}_i. \tag{6.2}$$

La matrice de covariance interclasse est définie par :

$$\mathbf{B} = \frac{1}{I} \sum_{q \in Q} I_q (\mathbf{g}_q - \mathbf{G})(\mathbf{g}_q - \mathbf{G})^t, \tag{6.3}$$

et la matrice de covariance intraclasse qui est la somme pondérée des covariances interclasses est donnée par :

$$\mathbf{W} = \frac{1}{I} \sum_{q \in Q} \sum_{i \in I_q} (\mathbf{x}_i - \mathbf{g}_q)(\mathbf{x}_i - \mathbf{g}_q)^t. \tag{6.4}$$

Proposition 6.2.1 Formule de décomposition de Huygens

L'inertie totale du nuage N_I est égale à la somme de l'inertie interclasse et de l'inertie intraclasse.

Cette proposition s'énonce également par le fait que la covariance totale du nuage est la somme de la covariance interclasse et de la covariance intraclasse :

$$\mathbf{V} = \mathbf{B} + \mathbf{W}. \tag{6.5}$$

La figure 6.2 illustre cette proposition. Le même nuage est représenté deux fois en reliant les points pour le calcul de la covariance totale à gauche et de la somme des covariances interclasse et intraclasse à droite.

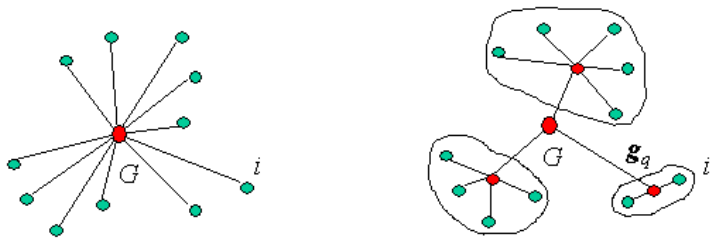


FIG. 6.2 – Illustration de la formule de Huygens.

Preuve La matrice de covariance totale est donnée par :

$$v_{kk'} = \frac{1}{I}\sum_{i\in I}(x_{ik} - G_k)(x_{ik'} - G_{k'}) = \frac{1}{I}\sum_{q\in Q}\sum_{i\in I_q}(x_{ik} - G_k)(x_{ik'} - G_{k'}), \tag{6.6}$$

où

$$G_k = \frac{1}{I}x_{ik}. \tag{6.7}$$

Or

$$(x_{ik} - G_k) = (x_{ik} - g_{qk}) + (g_{qk} - G_k), \tag{6.8}$$

nous remarquons ainsi que

$$\sum_{i\in I_q}(x_{ik} - g_{qk})(g_{qk'} - G_{k'}) = \sum_{i\in I_q}(g_{qk} - G_k)(x_{ik'} - g_{qk'}) = 0. \tag{6.9}$$

Donc uniquement deux des quatre termes de la partie droite de l'équation (6.6) sont non nuls et nous pouvons écrire :

$$v_{kk'} = b_{kk'} + w_{kk'}, \tag{6.10}$$

avec

$$b_{kk'} = \frac{1}{I}\sum_{q\in Q} I_q(g_{qk} - G_k)(g_{qk'} - G_{k'}), \tag{6.11}$$

et

$$w_{kk'} = \frac{1}{I}\sum_{q\in Q}\sum_{i\in I_q}(x_{ik} - g_{qk})(x_{ik'} - g_{qk'}), \tag{6.12}$$

ce qui démontre la proposition. □

Fonctions linéaires discriminantes

L'AFD consiste à trouver les combinaisons linéaires définissant de nouveaux axes tels que les projections des Q centres de gravité sur ces axes doivent être les plus éloignées, tandis que les projections de chaque sous-nuage sur ces axes doivent être les plus regroupées autour des projections des centres de gravité.

La marche à suivre est identique à celle d'une analyse factorielle. La première combinaison linéaire est donc celle qui maximise la variance interclasse et minimise la variance intraclasse. Puis, la deuxième combinaison linéaire est celle qui est non corrélée à la première et qui discrimine au mieux les classes au sens du même critère (maximisation de la variance interclasse et minimisation de la variance intraclasse). Les autres combinaisons linéaires sont déterminées de la même façon. Ces combinaisons linéaires sont appelées *fonctions linéaires discriminantes*.

Une combinaison linéaire $\mathbf{a}$ est un vecteur dans l'espace $\mathbb{R}^K$. La valeur de $\mathbf{a}$ pour un individu i est donnée par :

$$a(i) = \sum_{k \in K} a_k (x_{ik} - g_{qk}). \tag{6.13}$$

La variance de la variable $\mathbf{a}$ est définie par :

$$\text{var}(\mathbf{a}) = \frac{1}{I} \sum_{i \in I} a^2(i) = \frac{1}{I} \sum_{i \in I} \left[\sum_{k \in K} a_k (x_{ik} - g_{qk}) \right]^2, \tag{6.14}$$

ou encore

$$\text{var}(\mathbf{a}) = \frac{1}{I} \sum_{i \in I} \sum_{k \in K} \sum_{k' \in K} a_k a_{k'} (x_{ik} - g_{qk})(x_{ik'} - g_{qk'}) = \sum_{k \in K} \sum_{k' \in K} a_k a_{k'} v_{kk'}. \tag{6.15}$$

La variance de $\mathbf{a}$ est donc $\mathbf{a}^t \mathbf{V} \mathbf{a}$.

D'après l'équation (6.5), nous avons :

$$\mathbf{a}^t \mathbf{V} \mathbf{a} = \mathbf{a}^t \mathbf{B} \mathbf{a} + \mathbf{a}^t \mathbf{W} \mathbf{a}. \tag{6.16}$$

Le problème de l'AFD revient donc à trouver $\mathbf{a}$ tel que l'inertie des sous-nuages des individus I_q projetés sur $\mathbf{a}$ soit maximale (inertie interclasse $\mathbf{a}^t \mathbf{B} \mathbf{a}$) et chaque sous-nuage soit groupé donc l'inertie intraclasse $\mathbf{a}^t \mathbf{W} \mathbf{a}$ soit minimale. Chercher $\mathbf{a}$ tel que $\mathbf{a}^t \mathbf{B} \mathbf{a}$ soit maximale et $\mathbf{a}^t \mathbf{W} \mathbf{a}$ soit minimale est équivalant à chercher le maximum de la fonction :

$$f(\mathbf{a}) = \frac{\mathbf{a}^t \mathbf{B} \mathbf{a}}{\mathbf{a}^t \mathbf{V} \mathbf{a}}. \tag{6.17}$$

Il est encore équivalent de chercher le maximum de la forme quadratique $\mathbf{a}^t \mathbf{B} \mathbf{a}$ sous la contrainte quadratique $\mathbf{a}^t \mathbf{V} \mathbf{a} = 1$. Par la méthode du Lagrangien, nous pouvons montrer alors que :

$$\mathbf{B} \mathbf{a} = \lambda \mathbf{V} \mathbf{a}, \tag{6.18}$$

et lorsque la matrice $\mathbf{V}$ est inversible, nous obtenons :

$$\mathbf{V}^{-1} \mathbf{B} \mathbf{a} = \lambda \mathbf{a}. \tag{6.19}$$

Ainsi $\mathbf{a}$ est le vecteur propre de $\mathbf{V}^{-1} \mathbf{B}$ associé à la plus grande valeur propre λ.

Remarque Il faut donc diagonaliser $\mathbf{V}^1\mathbf{B}$ qui n'est pas *a priori* symétrique. Posons :

$$\mathbf{B} = \mathbf{C}^t\mathbf{C}, \tag{6.20}$$

avec :

$$c_{kq} = \sqrt{\frac{I_q}{I}}(g_{qk} - G_k). \tag{6.21}$$

Et posons :

$$\mathbf{a} = \mathbf{V}^{-1}\mathbf{C}\mathbf{v}. \tag{6.22}$$

L'équation (6.18) s'écrit alors :

$$\mathbf{C}\mathbf{C}^t\mathbf{V}^{-1}\mathbf{C}\mathbf{v} = \lambda\mathbf{C}\mathbf{v}. \tag{6.23}$$

Il suffit alors de diagonaliser la matrice symétrique $\mathbf{C}^t\mathbf{V}^{-1}\mathbf{C}$ d'ordre Q puis de déduire $\mathbf{a}$ à l'aide de $\mathbf{v}$.

En règle générale, il y a $Q-1$ valeurs propres donc $Q-1$ axes discriminants. C'est le cas si $I > K > Q$ et si les variables ne sont pas liées linéairement.

Cas de deux classes

Lorsqu'il n'y a que deux classes (*i.e.* $Q = 2$), nous sommes dans le cas d'un problème non sans importance de détection (et non plus de classification). Dans ce cas, il n'y a donc qu'un seul axe factoriel discriminant $\mathbf{a}$, déterminé par la droite passant par les centres de gravité des deux classes g_1 et g_2. Ainsi nous pouvons écrire :

$$\mathbf{B} = \mathbf{c}\mathbf{c}^t, \tag{6.24}$$

où le vecteur $\mathbf{c}$ de l'espace $\mathbb{R}^K$ est défini par :

$$c_k = \sqrt{\frac{I_1 I_2}{I}}(\mathbf{g}_1 - \mathbf{g}_2). \tag{6.25}$$

Nous avons donc :

$$\mathbf{V}^{-1}\mathbf{c}\mathbf{c}^t\mathbf{a} = \lambda\mathbf{a}, \tag{6.26}$$

ou encore :

$$\mathbf{c}^t\mathbf{V}^{-1}\mathbf{c}\mathbf{c}^t\mathbf{a} = \lambda\mathbf{c}^t\mathbf{a}. \tag{6.27}$$

Donc l'unique valeur propre est donnée par :

$$\lambda = \mathbf{c}^t\mathbf{V}^{-1}\mathbf{c}, \tag{6.28}$$

et l'unique fonction discriminante par :

$$\mathbf{a} = \mathbf{V}^{-1}\mathbf{c}. \tag{6.29}$$

λ est appelée *distance généralisée* entre les deux classes ou encore *distance de Mahalanobis*. Dans ce cas de deux classes, l'AFD est équivalente à la régression multiple [LMP95].

L'AFD peut aussi être vue comme une ACP des centres de gravité g_q de chaque classe avec une pondération pour ces individus donnée par la métrique $\mathbf{V}^{-1}$.

La représentation

Comme les autres méthodes factorielles, il est possible de représenter les individus dans les plans factorielles discriminants. Il est aussi possible comme pour l'ACP de représenter les variables en traçant le cercle de corrélation des K variables.

Afin de mesurer la qualité de la représentation, les mêmes indicateurs que l'ACP peuvent être employés. Par exemple la qualité de représentation d'un nuage par un axe $\mathbf{a}_s$ est donnée par le rapport :

$$\frac{\lambda_s}{\sum_{s \in S} \lambda_s}. \tag{6.30}$$

La contribution absolue du centre de gravité $\mathbf{g}_q$ à l'axe $\mathbf{a}_s$ est définie par :

$$\frac{I_q}{I} \left(\mathbf{a}_s^t \mathbf{V}^{-1} \mathbf{g}_q \right)^2, \tag{6.31}$$

et la contribution relative du centre de gravité $\mathbf{g}_q$ à l'axe $\mathbf{a}_s$ est définie par :

$$\frac{I_q}{I} \frac{1}{\lambda_s} \left(\mathbf{a}_s^t \mathbf{V}^{-1} \mathbf{g}_q \right)^2. \tag{6.32}$$

Dans une optique de classification, la qualité de la discrimination peut être définie par le rapport du nombre d'individus bien classés par le nombre total d'individus. Ce critère reste classique.

6.2.2 L'affectation

Lorsque les fonctions discriminantes ont été déterminées, nous souhaitons trouver la classe d'affectation d'un nouvel individu. Il existe plusieurs règles d'affectation (ou de classement) d'un nouvel individu i' dans une classe q. Nous en présentons ici quelques unes géométriques et probabilistes.

Distances aux centres de gravité

Une idée simple consiste à affecter un individu à la classe dont le centre de gravité est le plus près. Nous devons donc définir la distance entre le point individu i' décrit par le vecteur $\mathbf{x}_{i'}$ et le centre de gravité g_q du sous-nuage I_q. Rappelons ici quelques distances qui peuvent être envisagées.

– Distance euclidienne
  La distance euclidienne usuelle dans $\mathbb{R}^K$:

$$d_e^2(\mathbf{x}_{i'}, \mathbf{g}_q) = \sum_{k \in K} (x_{i'k} - g_{qk})^2. \tag{6.33}$$

Exprimons cette distance dans le nouvel espace. Notons :

$$z_r = \mathbf{u}_r^t(\mathbf{x}_{i'} - \mathbf{G}), \tag{6.34}$$

où $\mathbf{G}$ est le centre de gravité du nuage N_I défini par le vecteur $(G_k)_{k=1,..,K}$, r désigne l'axe principal issu de l'analyse, et $\mathbf{u}_r$ est le $r^{\text{ième}}$ vecteur propre normalisé de la matrice des covariances totales $\mathbf{V}$, définie précédemment, correspondant à la valeur propre λ_r. La distance euclidienne s'écrit alors :

$$d^2_{eV}(\mathbf{x}_{i'}, \mathbf{g}_q) = \sum_{r=1}^{r_{max}} (z_r - \bar{z}_{qr})^2, \tag{6.35}$$

où $\bar{z}_{qr} = \mathbf{u}^t(\mathbf{g}_q - \mathbf{G})$, r_{max} est le nombre de valeurs propres retenues, qui peut être ici le rang de la matrice $\mathbf{X}$ des données initiales.
La distance du nouvel individu i' décrit par le vecteur $\mathbf{x}_{i'}$ au centre de gravité $\mathbf{g}_q$ du sous-ensemble des individus I_q dans la métrique $\mathbf{V}^{-1}$ (*i.e.* sous la condition : $\mathbf{u}^t\mathbf{V}\mathbf{u} = 1$) est :

$$d^2_{eV^{-1}}(\mathbf{x}_{i'}, \mathbf{g}_q) = \sum_{r=1}^{r_{max}} \frac{(z_r - \bar{z}_{qr})^2}{\lambda_r}. \tag{6.36}$$

– Distance de Mahalanobis globale
Si nous remplaçons les données $\mathbf{X}$ par $\hat{\mathbf{X}}$ de terme général $\hat{x}_{ik} = x_{ik} - g_{qk}$, nous diagonalisons alors la matrice $\mathbf{W}$ au lieu de $\mathbf{V}$. Notons $\hat{\lambda}_r$ les valeurs propres de $\mathbf{W}$ et $\hat{z}_q$ les coordonnées de l'individu i' sur les nouveaux axes principaux $\hat{\mathbf{u}}_r$. La distance de $\mathbf{x}_{i'}$ au centre de gravité $\mathbf{g}_q$ dans la métrique $\mathbf{W}^{-1}$ s'écrit :

$$d^2_{Mg}(\mathbf{x}_{i'}, \mathbf{g}_q) = \sum_{r=1}^{r_{max}} \frac{(\hat{z}_r - \bar{\hat{z}}_{qr})^2}{\hat{\lambda}_r}. \tag{6.37}$$

– Distance de Mahalanobis locale
La distance de Mahalanobis locale est la distance de l'individu i' au centre de gravité $\mathbf{g}_q$ dans la métrique $\mathbf{W}_q^{-1}$, où $\mathbf{W}_q$ est la matrice des covariances internes de la classe I_q. Notons $w_{sq} = \mathbf{v}_{sq}^t(\mathbf{x}_{i'} - \mathbf{g}_q)$, où $\mathbf{g}_q$ est le centre de gravité du sous-nuage d'individus I_q décrit par le vecteur $(g_{qk})_{k=1,..,K}$, et w_{sq} est le $s^{\text{ième}}$ vecteur propre normalisé de $\mathbf{U}^q\mathbf{W}_q\mathbf{U}$ qui correspond à la valeur propre β_{sq}. La distance s'écrit alors :

$$d^2_{Ml}(\mathbf{x}_{i'}, \mathbf{g}_q) = \sum_{s=1}^{s_{max}(q)} \frac{(w_{sq} - \bar{w}_{qs})^2}{\beta_{sq}}, \tag{6.38}$$

où $\bar{w}_{qs} = \mathbf{v}_{qs}^t(\mathbf{g}_q - \mathbf{G})$, et $s_{max}(q)$ est le nombre de valeurs propres retenues dans le sous-nuage d'individus I_q.

– Distance du χ^2
La distance du χ^2 est déterminée par :

$$d^2_{\chi^2}(\mathbf{x}_{i'}, \mathbf{g}_q) = \sum_{k \in K} \frac{1}{s_{x_k}} \left(\frac{x_{i'k}}{s_{x_{i'}}} - \frac{g_{qk}}{s_{\mathbf{g}_q}} \right)^2, \tag{6.39}$$

où $s_{x_k} = \sum_{i \in I} x_{ik}$, $s_{x_{i'}} = \sum_{k \in K} x_{i'k}$ et $s_{\mathbf{g}_q} = \sum_{k \in K} \mathbf{g}_{qk}$. Dans le nouvel espace, nous avons donc :

$$d^2_{\chi^2}(\mathbf{x}_{i'}, \mathbf{g}_q) = \sum_{r=1}^{r_{max}} \frac{1}{s_{z_r}} \left(\frac{z_r}{s_z} - \frac{\bar{z}_{qr}}{s_{\bar{z}_q}} \right)^2. \tag{6.40}$$

Cependant cette distance s'applique habituellement aux tableaux de contingence comme nous l'avons vu pour l'AFC et l'ACM, elle convient donc peu à l'AFC en général.

– Distance de Minkowsky
Elle dépend d'un paramètre λ positif :

$$d_M(\mathbf{x}_{i'}, \mathbf{g}_q) = \left(\sum_{k \in K} |x_{i'k} - g_{qk}|^\lambda \right)^{\frac{1}{\lambda}}. \tag{6.41}$$

Dans le nouvel espace, nous avons :

$$d_M(\mathbf{x}_{i'}, \mathbf{g}_q) = \left(\sum_{r=1}^{r_{max}} |z_r - \bar{z}_{qr}|^\lambda \right)^{\frac{1}{\lambda}}. \tag{6.42}$$

Si $\lambda = 1$, nous avons la distance des valeurs absolues aussi nommée *distance de Manhattan*, du nom du quartier new-yorkais, $\lambda = 2$, nous retrouvons la distance euclidienne. Lorsque $\lambda \to +\infty$, nous obtenons la distance de Tchebychev :

$$d_T(\mathbf{x}_{i'}, \mathbf{g}_q) = \max_r |z_r - \bar{z}_{qr}|. \tag{6.43}$$

D'autres distances sont envisageables. Cependant, pour l'AFC il est généralement retenu la distance de Mahalanobis globale (métrique $\mathbf{W}^{-1}$) ou locale (métrique $\mathbf{W}_q^{-1}$, où $\mathbf{W}_q$ est la matrice des covariances internes au sous-nuage I_q). Cette dernière permettant de réduire les erreurs d'affectation lorsque les dispersions des classes sont très différentes.

Une autre approche géométrique est possible, non plus en considérant les centres de gravité, mais les individus proches du nouvel individu.

Règle des k plus proches voisins

Cette méthode d'affectation peut être employée directement pour la classification dans l'espace initial. Elle est très utilisée en reconnaissance des formes.

Le principe est simple, nous affectons le nouvel individu i' au sous-nuage d'individus I_q le plus représenté dans son voisinage. Le voisinage est étendu jusqu'à ce qu'il contienne k individus. Ainsi notons :

$$K_q(i') = card\left\{i \in I \text{ tel que } i \in I_q, i \in V_k(i')\right\}, \tag{6.44}$$

où $V_k(i')$ désigne le voisinage de l'individu i' formé par k individus. Cet ensemble peut se formaliser pour $k = 1$ par :

$$V_1(i') = \left\{i \in I \text{ tel que } d(i', i) \leq d(i', i'')\, \forall i'' \in I, i'' \neq i\right\}, \tag{6.45}$$

par récurrence, nous obtenons pour un k quelconque :

$$V_k(i') = V_k(i') \cup \left\{i \in I \setminus V_{k-1} \text{ tel que } d(i', i) \leq d(i', i'')\, \forall i'' \in I \setminus V_{k-1}, i'' \neq i\right\}. \tag{6.46}$$

Nous voyons que la aussi la définition d'une distance adéquate est importante. Il est possible d'employer une des distance précédemment présentées.

La décision est alors prise en cherchant le maximum de $K_q(i')$ (*i.e.* que l'individu i' est affecté à la classe $\underset{q \in Q}{\operatorname{argmax}} K_q(i')$). D'autres règles de décisions sont envisageables issues des méthodes de votes [Mar04].

Il existe une variante intéressante de cette approche la classification par k plus proches voisins flous.

Cette approche très coûteuse donne de bons résultats. C'est pourquoi elle sert souvent de méthode de comparaison en reconnaissance des formes avec d'autres approches moins coûteuses.

Ce type d'affectation ne prend cependant pas en compte les probabilités *a priori* de chaque classe.

Approche bayésienne

Cette approche probabiliste simple consiste à affecter l'individu i' au sous-nuage d'individus I_q pour lequel la probabilité $P(I_q/i')$ est maximale. Or d'après la règle de Bayes, nous avons :

$$P(I_q/i') = \frac{P(i'/I_q)P(I_q)}{\sum_{q' \in Q} P(i'/I_{q'})P(I_{q'})}. \tag{6.47}$$

Il suffit alors de maximiser $P(i'/I_q)P(I_q)$. Cependant pour estimer cette probabilité il faut connaître les probabilités *a priori* $P(I_q)$, ce qui n'est pas toujours le cas. Elles peuvent être estimées, mais il faut alors être sûr de la capacité de généralisation des données d'apprentissage. Il faut de plus estimer la probabilité $P(i'/I_q)$ qui nécessite :

- soit une estimation à partir des fréquences et dans ce cas il faut encore être sûr de la capacité de généralisation des données d'apprentissage,

- soit faire l'hypothèse de la distribution. La distribution gaussienne qui peut être justifiée par la loi forte des grands nombres est souvent employée. De plus elle ne nécessite que l'estimation de deux paramètres (la moyenne et la variance).

Dans ce dernier cas, lorsque les distributions gaussiennes d'appartenance à chaque sous-nuage sont de même matrice de covariance intraclasse et s'il y a équiprobabilité des classes (les probabilités *a priori* $P(I_q)$ sont identiques), alors l'approche bayésienne est équivalente à affecter la classe du plus proche voisin en utilisant la distance de Mahalanobis locale (*cf.* [LMP95] pour plus de détails).

Il existe d'autres méthodes d'affectation, car en fait toute méthode de classification peut être employée pour cette étape de classement. Bien souvent, les approches les plus simples donnent de meilleurs résultats, au dépend d'un coût plus important.

6.3 Conclusion

L'AFD est une méthode très utilisée de nos jours. Sa simplicité de mise en œuvre fait que nous la retrouvons dans de nombreux logiciels. Elle est adéquate pour la représentation des données dans des espaces qui discriminent au mieux les individus selon des classes connues. Cette représentation permet de dégager des informations à partir d'un grand nombre de données souvent difficile à interpréter. Elle permet également l'affectation de nouveaux individus dans les classes existantes. Il est alors possible de rendre la méthode adaptative pour tenir compte de ces nouvelles observations.

Il peut s'avérer très enrichissant de l'employer en complément d'une autre analyse factorielle telles que l'ACP ou l'ACM.

Chapitre 7

Classification

7.1 Introduction

La classification sans *a priori* est depuis longtemps une problématique importante issue surtout de l'étude des phénomènes naturelles et de la biologie en particulier. Toutes les méthodes ainsi développées appartiennent à une science la *taxonomie* littéralement la science des lois de l'ordre [Ben80a]. Les méthodes de classification font parties intégrante de l'analyse de données. Dans le domaine de la reconnaissance des formes elle porte le nom de classification non-supervisée. Le terme anglais pour classification est *clustering*, les classes étant des *clusters*. Le terme anglais *classification* désigne davantage classement *i.e.* le fait d'affecter des objets à des classes prédéfinies, voire analyse de données en général.

7.1.1 Les objectifs

La classification a pour principal objectif de rassembler les éléments (individus ou variables) qui se ressemblent et/ou de séparer ceux qui diffèrent. C'est-à-dire qu'il s'agit de créer des classes homogènes les plus éloignées les unes des autres. Si cet objectif est facilement compréhensible, il n'en est pas moins compliqué à atteindre. Nous sous-entendons lorsque nous cherchons à classer des éléments, qu'il existe des regroupements, soit en nombre inconnu soit en nombre supposé.

Si nous cherchons souvent à regrouper des éléments entre eux, c'est afin de mieux interpréter une grand quantité de données.

Les objectifs de la classification sont donc de regrouper les individus décrits par un ensemble de variables, ou regrouper les variables observées sur des individus et d'interpréter ces regroupements par une synthèse des résultats. L'intérêt de regrouper les individus est ici de les classer en conservant leur caractère multidimensionnel, et non pas seulement à partir d'une seule variable. Si les variables sont nombreuses il peut être intéressant de les regrouper afin de réduire leur nombre pour une interprétation plus facile.

Les méthodes de classification sont donc complémentaires des analyses factorielles décrites dans les chapitres précédents.

		Variables			
		1	..... k		K
Individus	1				
	⋮		⋮		
	i	.....	x_{ik}	.....	
	⋮		⋮		
	I				

Tab. 7.1 – Représentation des données pour la classification.

7.1.2 Les données

Les données de départ sont souvent organisées comme une matrice X décrite par le tableau 7.1, où x_{ik} est la valeur de la variable k pour l'individu i, I représente à la fois le nombre d'individus et l'ensemble $I = \{1, ..., I\}$, et K représente à la fois le nombre de variables et l'ensemble $K = \{1, ..., K\}$.

Les variables peuvent être quantitatives continues ou issues de tableaux de contingences, ou binaires issues de tableaux logiques, ou encore qualitatives. Afin de traiter l'ensemble de ces types de variables, c'est la mesure de similarité ou dissimilarité qui doit être adaptée aux types de données. En effet, nous nous doutons qu'il est important de définir une mesure de similarité pour regrouper des éléments ou de dissimilarité pour les éloigner. Une mesure de similarité ou de dissimilarité est une distance à l'exception que l'inégalité triangulaire n'est pas exigée. Ces mesures peuvent être des distances dans le cas de variables quantitatives. Ainsi, il est préférable d'employer une distance euclidienne, de Mahalanobis ou de Minkowsky pour les variables quantitatives continues et une distance du χ^2 pour des tableaux de contingences, distances que nous avons déjà présentées à la section 6.2.2 du chapitre précédent.

Dans le cas de tableaux binaires, un grand nombre de mesures de similarités entre deux éléments ont été définies à partir des quatre quantités. Par exemple pour deux individus x_1 et x_2 elles sont données par :

- soit a le nombre de fois où $x_{1k} = x_{2k} = 1$,
- soit b le nombre de fois où $x_{1k} = 0$ et $x_{2k} = 1$,
- soit c le nombre de fois où $x_{1k} = 1$ et $x_{2k} = 0$,
- soit d le nombre de fois où $x_{1k} = x_{2k} = 0$.

Les similarités suivantes ont été proposées par différents auteurs :

- $\dfrac{a}{a+b+c}$ par Jaccard,
- $\dfrac{a}{a+b+c+d}$ par Russel et Rao,
- $\dfrac{2a}{2a+b+c}$ par Dice,

- $\dfrac{a}{a+2(b+c)}$ par Sokal et Sneath,
- $\dfrac{a+d}{a+b+c+d}$ par Sokal et Michener,
- $\dfrac{a}{a+b}+\dfrac{a}{a+c}$ par Kulzinsky,
- $\dfrac{a+d}{a+d+2(b+c)}$ par Rogers et Tanimoto,
- $\dfrac{ad-bc}{ad+bc}$ par Yule,
- $\dfrac{|ad-bc|}{[(a+b)(c+d)(a+c)(b+d)]^2}$ par Pearson,
- $\dfrac{a}{[(a+b)(c+d)(a+c)(b+d)]^2}$ par Ochiaï.

Dans le cas des variables qualitatives, il suffit de considérer le tableau de contingence associé. En effet, si elles n'ont pas le même nombre de modalités, il est très difficile de définir une distance.

Si le tableau est composé de données mixtes, il suffit de rendre les variables quantitatives en variables qualitatives en choisissant quelques modalités de la même façon que décrite à la section 5.2.4.

7.1.3 Les méthodes

Il existe un grand nombre de méthodes et surtout beaucoup de variantes. Il est possible de les différentier grossièrement soit par leur structure de classification, soit par le type de représentation des classes. Ainsi, nous pouvons distinguer quatre types de représentation [Bro03] :

- Les partitions sont une notion la plus naturelle, chaque individu est affecté à une classe et une seule.
- Les hiérarchies sont un ensemble de partitions emboîtées. Ainsi une classe se divise en sous-classes.
- Les arbres additifs sont une autre vision des hiérarchies ; une structure dont les nœuds terminaux sont les individus classés et les nœuds intérieurs les classes. Une extension des arbres additifs est la notion d'arbre au sens de la théorie de graphes.
- Les pyramides sont une généralisation des hiérarchies car elles permettent des empiétements entre les classes.

Les méthodes de classification cherchent à transformer le tableau de données en un autre tableau ayant de "bonnes propriétés". C'est donc un problème d'optimisation. Cependant ces tableaux se trouvent dans des espaces discrets, ces transformations ne peuvent être décrites par des fonctions issues de calculs formalisés usuels, et il n'y a pas de solutions mathématiques exactes. C'est donc dans le cadre des mathématiques discrètes, que des solutions approximatives sont proposées dans une démarche algorithmique.

Nous nous contentons ici de présenter deux méthodes, deux algorithmes, les plus utilisés et qui se retrouvent dans la plupart des logiciels de statistiques. Nous présentons une méthode conduisant à des partitions, la méthode des centres mobiles à la section 7.2, puis une méthode conduisant à des hiérarchies, la classification hiérarchique à la section 7.3.

7.2 Méthode des centres mobiles

Cette méthode peut être vu comme un cas particulier de l'approche des nuées dynamiques développée par E. Diday [CDG+89]. Cette méthode d'un formalisme très simple n'en est pas moins très efficace pour de vastes tableaux de données. Elle est de plus rapide, mais cependant pas toujours optimale.

La méthode des centres mobiles est fondée sur une méthode de partitionnement directe des individus connaissant par avance le nombre de classes attendues.

7.2.1 Principe de l'algorithme

Nous supposons désirer partitionner le nuage des individus N_I dans l'espace $\mathbb{R}^K$ muni d'une distance appropriée que nous notons par d. Cette distance d doit être choisie en fonction des données (*cf.* section 7.1.2). En pratique, il s'agit souvent de la distance euclidienne ou du χ^2 qui est implémentée. Supposons de plus, que nous souhaitons partitionner N_I en Q classes avec $Q \leq I$.

- Étape 0 : Nous choisissons Q individus dans le nuage N_I qui constituent Q centres provisoires des Q classes. Le choix de ces centres est important pour la rapidité de la convergence, et les connaissances *a priori* doivent ici être mises à profit, s'il y en a. Dans le cas contraire, le plus courant, il suffit de tirer aléatoirement ces centres par un tirage sans remise. Notons par $\left\{C_1^0, ..., C_q^0, ..., C_Q^0\right\}$ ces centres. Ces centres fournissent une première partition $P^0 = \left\{I_1^0, ..., I_q^0, ..., I_Q^0\right\}$ du nuage N_I des individus en Q classes. Un individu i appartient au sous-nuage I_q^0 s'il est plus proche de C_q^0 que de tous les autres centres. Dans un espace à deux dimensions, les sous-nuages sont délimités deux à deux par des droites médiatrices des centres des sous-nuages, c'est ce qui est applelé *diagramme de Voronoï*. Bien sûr à ce niveau, la distance d intervient.
- Étape 1 : Q nouveaux centres $\left\{C_1^1, ..., C_q^1, ..., C_Q^1\right\}$ sont déterminés en prenant les centres de gravité des sous-nuages I_q^0 obtenus par la partition P^0. La distance d intervient de nouveau ici. Ces nouveaux centres induisent une nouvelle partition $P^1 = \left\{I_1^1, ..., I_q^1, ..., I_Q^1\right\}$, suivant le même critère précédent.
- Étape m : Q nouveaux centres $\left\{C_1^m, ..., C_q^m, ..., C_Q^m\right\}$ sont déterminés en prenant les centres de gravité des sous-nuages I_q^{m-1} obtenus par la partition P^{m-1}. Ces nouveaux centres induisent une nouvelle partition $P^m = \left\{I_1^m, ..., I_q^m, ..., I_Q^m\right\}$, suivant le même critère précédent.

La convergence de l'algorithme est garantie [LMP95]. Le critère d'arrêt est celui de deux partitions identiques. D'autres critères permettent d'augmenter la rapidité. Par exemple,

nous pouvons cesser les itérations lorsque la variance intraclasse de toutes les classes est suffisamment faible, ou encore lorsqu'un nombre d'itérations défini *a priori* est atteint.

Cette algorithme est illustré sur la figure 7.1 dans le cas où $Q = 2$. Deux figures

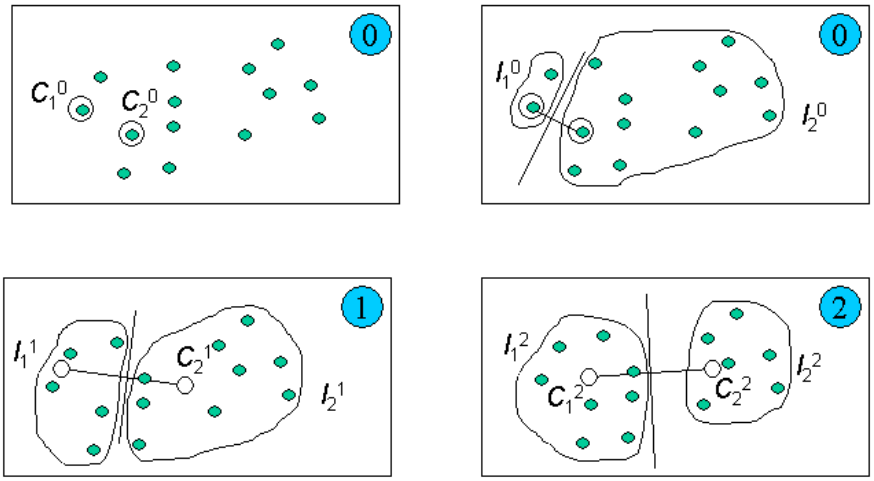


FIG. 7.1 – Illustration de l'algorithme des centres mobiles.

présentent l'étape 0 : le tirage aléatoire des centres provisoires C_1^0 et C_2^0 et la construction de la première partition $P^0 = \{I_1^0, I_2^0\}$ en affectant chaque individu au sous-nuage dont le centre obtenu est le plus proche. L'étape 1 présente les nouveaux centres et les sous-nuages dont ils sont les centres de gravité. De nouveau, l'étape 2 fournit les centres de gravité des nouveaux sous-nuages I_1^2 et I_2^2.

De nombreux algorithmes sont fondés sur un principe similaire. Les deux principaux sont les nuées dynamiques et les *k-means* ou k-moyennes. La différence pour la méthode des nuées dynamiques se situe au niveau de la réaffectation des individus à chaque classe. Après avoir déterminé les centres de gravité, un *noyau* est déterminé pour chaque classe comme étant l'individu le plus proche du centre de gravité de chaque classe. La réaffectation se fait alors en fonction de la distance des autres individus aux noyaux de chaque classe. Ce formalisme a permis plusieurs généralisations de la méthode.

La méthode des *k-means* après avoir choisi une première fois les centres mobiles, recalcule le centre de chaque classe dès lors qu'un individu y est affecté. La position du centre est donc modifiée à chaque affectation, ce qui permet d'avoir une bonne partition en peu d'itérations. D'autres algorithmes sont présentés par exemple dans [Ben80a].

7.3 La classification hiérarchique

Il existe principalement deux familles d'algorithmes de classification hiérarchique :

- les algorithmes ascendants : la construction des classes se fait par des agglomérations successives des éléments deux à deux,
- les algorithmes descendants : la construction des classes se fait par dichotomies successives de l'ensemble des éléments.

Ces deux approches conduisent à une hiérarchie des partitions des éléments. La seconde approche est beaucoup moins employée que la première, nous présentons donc ici la première approche.

7.3.1 Principe de la classification hiérarchique ascendante

Le principe repose donc sur la création à chaque étape d'une partition obtenue en agrégeant deux à deux les éléments (individus ou plus rarement variables) les plus proches. Les différentes façons de créer un nouveau couple constituent autant de différents algorithmes de classification hiérarchique ascendante.

Méthodes d'agrégation

Supposons que le nuage initial, par exemple N_I, à classer est muni d'une distance (ou d'une mesure de similarité ou dissimilarité) d. La façon de regrouper des individus ou des groupes d'individus repose sur des règles de calcul des distances entre ces classes (individus ou groupes d'individus) disjointes, appelées *critère d'agrégation*.

Soit x, y et z trois classes. Si les classes x et y sont regroupées en une seule classe h, plusieurs critères d'agrégation sont possibles :

- distance du *saut minimal* :

$$d(h,z) = \min\{d(x,z), d(y,z)\}, \tag{7.1}$$

- distance du *saut maximal* :

$$d(h,z) = \max\{d(x,z), d(y,z)\}, \tag{7.2}$$

- distance *moyenne* :

$$d(h,z) = \frac{d(x,z) + d(y,z)}{2}, \tag{7.3}$$

- distance *moyenne généralisée*, en notant n_x et n_y le nombre d'individus de x et y :

$$d(h,z) = \frac{n_x d(x,z) + n_y d(y,z)}{n_x + n_y}. \tag{7.4}$$

Ces méthodes d'agrégation ont l'avantage de conduire à des calculs simples et possèdent des propriétés mathématiques intéressantes. Cependant, les résultats ne sont pas toujours bons. En particulier, la distance du saut minimal peut entraîner des *effets de chaîne*, illustrés sur la figure 7.2. Sur le nuage de points représenté sur cette figure, les groupes A et B ne sont pas facilement discernables par la distance du saut minimal. Il

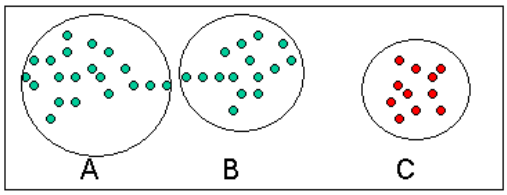


FIG. 7.2 – Illustration de l'effet de chaîne.

est difficile de déterminer au niveau de la chaîne quels points appartiennent à A et quels points appartiennent à B. Le critère de la distance moyenne donne de meilleurs résultats, mais comme nous le voyons sur la figure (les classes ont alors des formes de cercles), elle a tendance à considérer A et B comme deux classes, alors qu'il s'agit d'un seul sous-nuage.

Pour remédier à ce problème, des critères d'agrégation selon la variance sont liés à des calculs d'inertie. Cette méthode est particulièrement facile à mettre en œuvre après une analyse factorielle, les éléments étant donnés par leurs coordonnées sur les premiers axes factoriels.

Agrégation selon l'inertie Cette méthode porte également le nom de la méthode de Ward. La solution au problème évoqué ci-dessus est donc de considérer les éléments - prenons les individus - comme un nuage de points N_I dans $\mathbb{R}^K$. L'idée est ensuite d'agréger les individus en minimisant l'inertie (ou la variance) intraclasse et en maximisant l'inertie interclasse.

Le principe repose sur la formule de décomposition de Huygens présentée par la proposiiton 6.2.1 à la section 6.2. Ainsi l'inertie totale du nuage N_I est égale à la somme de l'inertie interclasse et de l'inertie intraclasse :

$$I = I_{\text{intra}} + I_{\text{inter}}. \tag{7.5}$$

Reprenons la figure 7.3 illustrant cette proposition. Le même nuage est représenté deux fois en reliant les points pour le calcul de l'inertie totale à gauche et de la somme des inerties interclasse et intraclasse à droite. Considérons que chaque individu i est muni

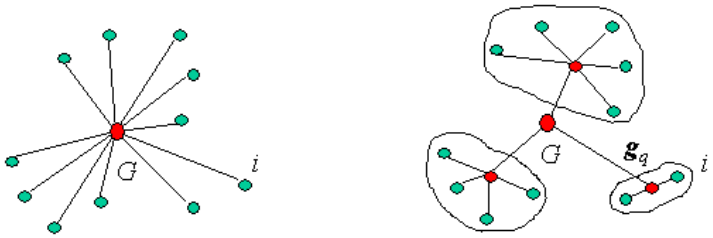


FIG. 7.3 – Illustration de la formule de Huygens.

d'une masse $\mathbf{m}_i$ et chaque classe q est munie d'une masse $\mathbf{m}_q$. Avec les notations de la figure 7.3, la formule de décomposition de Huygens s'écrit :

$$I = \sum_{q \in Q} \mathbf{m}_q d^2(\mathbf{g}_q, \mathbf{G}) + \sum_{q \in Q} \sum_{i \in I_q} \mathbf{m}_i d^2(\mathbf{x}_i, \mathbf{g}_q), \tag{7.6}$$

où d représente la distance choisie initialement, $\mathbf{g}_q$ est le centre de gravité du sous-nuage N_{I_q} et $\mathbf{G}$ le centre de gravité du nuage des individus N_I.

Ainsi la qualité globale d'une partition est liée à l'homogénéité interne des sous-nuages et donc également à l'éloignement des sous-nuages. Par exemple, la figure 7.4 illustre deux partitions en deux sous-nuages, celui de gauche avec une inertie intraclasse faible, celui de droite avec une inertie intraclasse élevée.

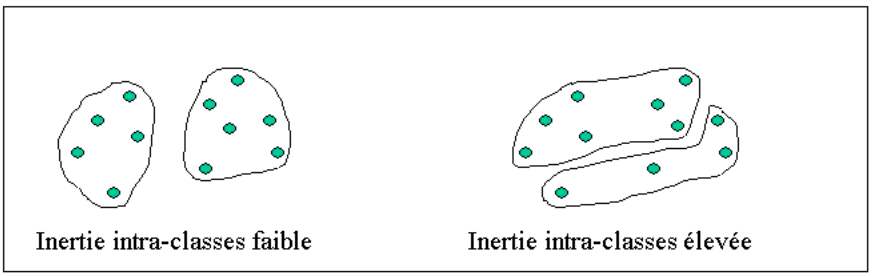


FIG. 7.4 – Illustration d'une inertie intraclasse faible et élevée.

Ainsi pour une agrégation, nous cherchons à faire varier le moins possible l'inertie intraclasse, ce qui est équivalent à rendre minimale la perte d'inertie interclasse résultant de cette agrégation. Considérons une partition P_s à s classes (ou sous-nuages), en associant deux classes $\mathbf{a}$ et $\mathbf{b}$ à P_s, nous obtenons une partition p_{s-1} à $s-1$ classes (*cf.* figure 7.5). L'élément $\mathbf{c}$ obtenu par l'agrégation de $\mathbf{a}$ et $\mathbf{b}$ a pour masse $m_{\mathbf{c}} = m_{\mathbf{a}} + m_{\mathbf{b}}$, et il peut

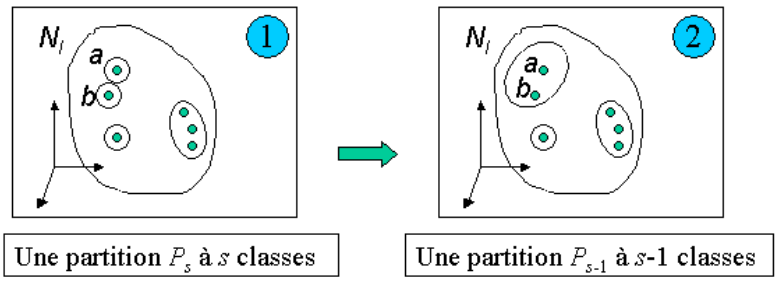


FIG. 7.5 – Illustration du passage d'une partition P_s à une partition p_{s-1}.

être décrit par son centre de gravité donné par :

$$\mathbf{c} = \frac{m_{\mathbf{a}}\mathbf{a} + m_{\mathbf{b}}\mathbf{b}}{m_{\mathbf{a}} + m_{\mathbf{b}}}. \tag{7.7}$$

L'inertie interclasse de $\mathbf{a}$ et de $\mathbf{b}$ peut se décomposer par la formule de Huygens par :

$$I_{\text{inter}(\mathbf{ab})} = m_{\mathbf{a}}d^2(\mathbf{a},\mathbf{G}) + m_{\mathbf{b}}d^2(\mathbf{b},\mathbf{G}) = m_{\mathbf{a}}d^2(\mathbf{a},\mathbf{c}) + m_{\mathbf{b}}d^2(\mathbf{b},\mathbf{c}) + m_{\mathbf{c}}d^2(\mathbf{c},\mathbf{G}), \tag{7.8}$$

or l'inertie de la partition P_s est donnée par :

$$I_s = I_{\text{inter}(\mathbf{ab})} + I_{\text{intra}(\mathbf{a})} + I_{\text{intra}(\mathbf{b})}, \tag{7.9}$$

et celle de la partition P_{s-1} par :

$$I_{s-1} = I_{\text{inter}(\mathbf{c})} + I_{\text{intra}(\mathbf{a})} + I_{\text{intra}(\mathbf{b})} = m_{\mathbf{c}}d^2(\mathbf{c},\mathbf{G}) + I_{\text{intra}(\mathbf{a})} + I_{\text{intra}(\mathbf{b})}. \tag{7.10}$$

Ainsi la perte d'inertie $\Delta I_{\text{inter}(\mathbf{ab})}$ due au passage de la partition P_s à la partition P_{s-1} est donnée par :

$$\Delta I_{\text{inter}(\mathbf{ab})} = I_{\text{inter}(P_s)} + I_{\text{inter}(P_{s-1})} = m_{\mathbf{a}}d^2(\mathbf{a},\mathbf{c}) + m_{\mathbf{b}}d^2(\mathbf{b},\mathbf{c}). \tag{7.11}$$

En remplaçant $\mathbf{c}$ par sa valeur en fonction de $\mathbf{a}$ et $\mathbf{b}$, nous obtenons :

$$\Delta I_{\text{inter}(\mathbf{ab})} = \frac{m_{\mathbf{a}}m\mathbf{b}}{m_{\mathbf{a}} + m_{\mathbf{b}}}d^2(\mathbf{a},\mathbf{b}). \tag{7.12}$$

Cette variation représente un indice de dissimilarité (appelé aussi *indice de niveau*) qui est l'*inertie de l'haltère* (ou *variance du dipôle*) (a, b). Il est aisé de vérifier que la somme des indices de dissimilarité entre toutes les partitions est l'inertie totale du nuage N_I.

Le principe de la méthode de Ward est donc de déterminer les éléments $\mathbf{a}$ et $\mathbf{b}$ d'une partition P_s qui ont un indice de dissimilarité minimal.

Algorithme

L'algorithme de classification hiérarchique ascendante est simple et facile à programmer. Son déroulement suit les étapes suivantes :

- Étape 1 : Nous considérons le nuage N_I comme une partition P_I de I éléments.
- Étape 2 : Une transformation des données s'effectue par la construction à partir de la matrice X décrite par le tableau 7.1 d'une matrice de distances entre les I individus, à partir de la distance retenue initialement. Nous recherchons ensuite les deux éléments à agréger (*i.e.* les deux éléments les plus "proches" en terme de distance ou d'indice de dissimilarité). L'agrégation des deux éléments fournit une partition P_{I-1} à $I-1$ individus.
- Étape 3 : Nous construisons la nouvelle matrice $((I-1)\times(I-1))$ des distances, puis nous recherchons les deux nouveaux éléments à agréger. L'agrégation des deux éléments fournit une partition à $I-2$ individus.

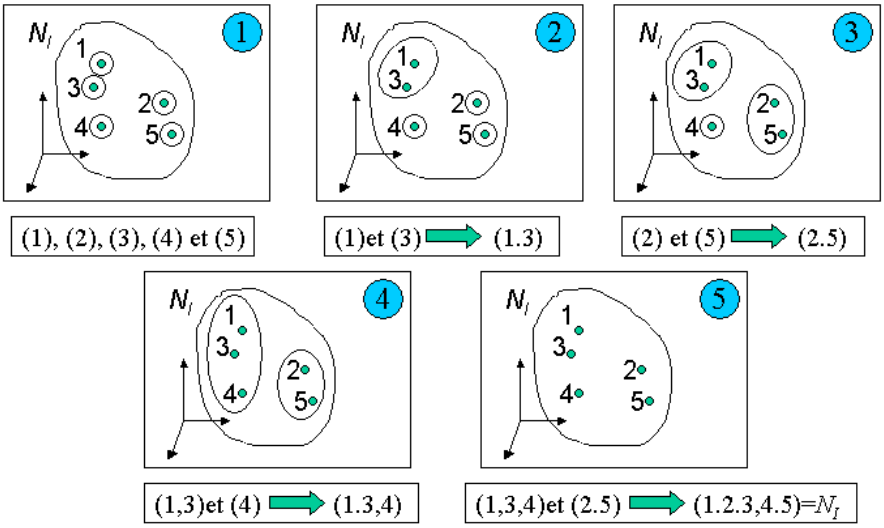


FIG. 7.6 – Illustration de l'algorithme de classification avec avec un nuage de $I = 5$ individus.

Nœud	(6)	(7)	(8)	(9)
benjamin	(1)	(2)	(6)	(8)
aîné	(3)	(5)	(4)	(7)
effectif	2	2	3	5

TAB. 7.2 – Relation entre les nœuds de l'arbre.

- Étape m : Nous calculons la matrice $((I-(m-1)) \times (I-(m-1)))$ des distances, puis nous cherchons à agréger deux éléments jusqu'à ce qu'il n'en reste plus qu'un qui constitue la dernière partition P_1.

Afin d'illustrer cet algorithme, nous donnons un exemple d'un nuage N_I de cinq individus sur la figure 7.6.

Les étapes successives de cet algorithme peuvent être représentées par un *arbre hiérarchique* également appelé *dendrogramme* où sont représentées en ordonnées les indices de dissimilarité (*cf.* figure 7.7).

Vocabulaire lié au dendrogramme

- Les *éléments terminaux* de l'arbre (ou de la hiérarchie) sont les individus (ou variables selon ce qui est classé).
- Les nœuds de l'arbre correspondent aux regroupements de deux éléments appelés *aîné* et *benjamin*. L'arbre de la figure 7.7 peut ainsi être décrit par le tableau 7.2.
- L'agrégation repose sur les inégalités des distances entre elles. Nous pouvons obtenir le même classement en des couples d'éléments en classant ces couples par ordre croissant des distances. Un tel classement est appelé *ordonnance*.

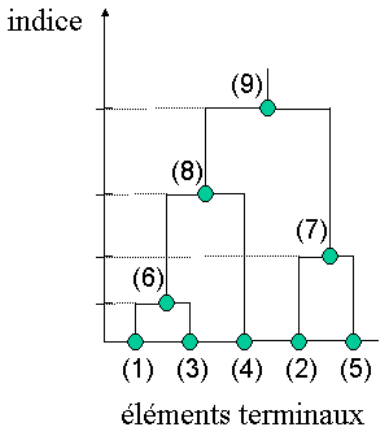


FIG. 7.7 – Exemple de dendrogramme.

- La *hiérarchie* peut être décrite par une famille H d'éléments de I telle que :
  - $I \in H$, $\{i\} \in H \ \forall i \in I$,
  - $\forall A, B \in H$, $A \cap B \in \{A, B, \emptyset\}$ *i.e.* deux classes sont soit disjointes, soit l'une est incluse dans l'autre.
    Ainsi toute classe est la réunion des classes qui sont incluses en elle. La famille des sous-ensembles construits par la classification ascendante hiérarchique forme une hiérarchie. C'est en fait une *hiérarchie binaire*, il en existe d'autres.
- Une *hiérarchie indicée* est une hiérarchie pour laquelle il existe une fonction v de H dans $\mathbb{R}^+$ *i.e.* telle que :

$$A \subset B \Leftrightarrow v(A) \leq v(B), \forall A, B \in H. \tag{7.13}$$

  La hiérarchie est généralement indicée par les valeurs des distances (ou indices de dissimilarité) correspondant à chaque étape d'agrégation.
- En coupant l'arbre par une droite horizontale, nous obtenons une *partition*. Une hiérarchie donne ainsi une chaîne de I partitions de 1 à I classes.

Les hiérarchies indicées ont une propriété particulièrement intéressante, car elle peuvent être vues comme un ensemble muni d'une *ultramétrique* [LMP95]. Une ultramétrique est une distance d particulière. En tant que distance d associée au nuage N_I, elle est une application qui vérifie :

- $x = y \Leftrightarrow d(x, y) = 0, \quad \forall x, y \in N_I$,
- $d(x, y) = d(y, x), \quad \forall x, y \in N_I$ (relation de symétrie),
- $d(x, y) \leq d(x, z) + d(y, z), \quad \forall x, y, z \in N_I$ (inégalité triangulaire).

Cette distance d est une ultramétrique si elle vérifie une condition plus forte que l'inégalité triangulaire donnée par $d(x, y) \leq \max(d(x, z), d(y, z)) \ \forall x, y, z \in N_I$. La distance du saut minimal est la plus grande ultramétrique inférieure à la métrique d initiale.

7.3.2 Interprétation

L'interprétation repose essentiellement sur la lecture du dendrogramme. Elle devient problématique lorsque le nombre d'individus est très important. Elle doit se faire de haut en bas afin d'examiner d'abord les partitions qui possèdent peu de classes, pour ensuite entrer dans des considérations plus détaillées. Nous cherchons, essentiellement la partition qui présente le plus d'intérêt. Pour cela, il faut chercher à construire des classes homogènes. Une bonne partition, *i.e.* une bonne coupure de l'arbre, doit comporter peu de classes avec une inertie intraclasse faible et une inertie interclasse élevée. Pour le choix de la coupure, nous pouvons également nous aider de la courbe des indices. Ainsi nous devons rechercher le nœud après lequel il y a une perte d'indice importante. Ceci peut également se lire sur le dendrogramme.

Exemple 7.3.1 Prenons l'exemple de l'étude des données de granulométrie proposée par Kendall, Stuart et Griffin en 1963. Ces données sont composées d'échantillons de sol décrits par cinq variables sur leur composition : sable, limon, argile, matière organique, pH. La figure 7.8 présente le dendrogramme obtenu par 'approche de Ward, tandis que la figure 7.9 présente la courbe des indices. Nous constatons à partir de ces deux figures qu'une coupure en cinq classes fournit des classes homogènes et éloignées des autres classes. De plus, cette coupure est confortée par le taux de variance intraclasse qui est de 22,3%, alors que le taux de variance interclasse est de 77,7%.

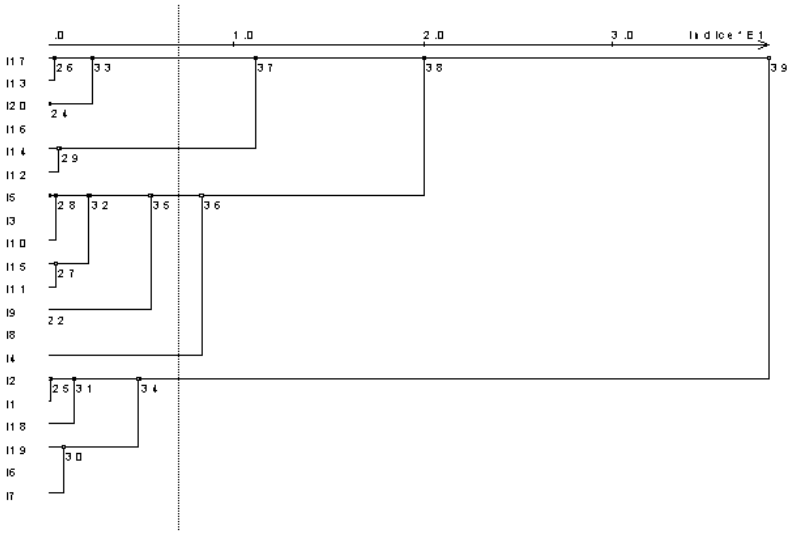


FIG. 7.8 – Dendrogramme sur les données de composition du sol.

Une fois la coupure faite, *i.e.* le choix de la partition à étudier, il faut examiner les classes obtenues. Pour ce faire il faut trouver les variables représentatives de chaque classe, pour ensuite interpréter ces classes à partir des variables explicatives. Deux indicateurs sont essentiellement employés pour cette interprétation :

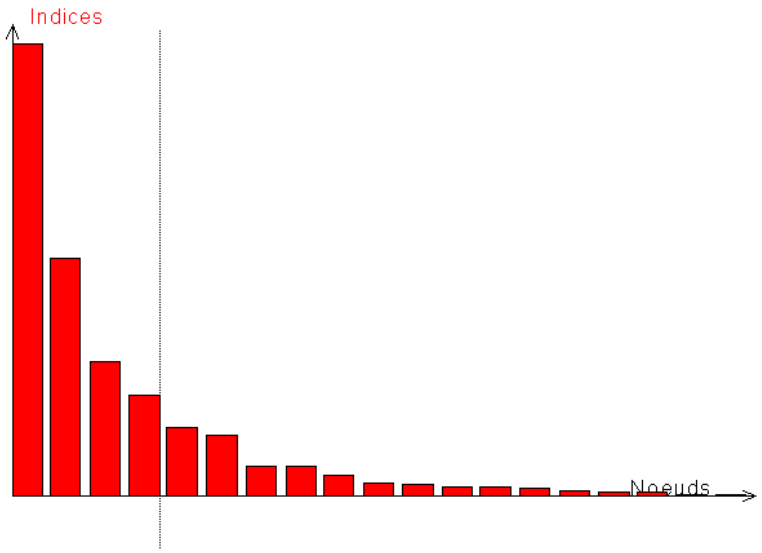


FIG. 7.9 – Courbe des indices sur les données de composition du sol.

- l'*excentricité* de la classe par rapport au centre de gravité général de l'ensemble des individus,
- la *variance du dipôle* constituée par les deux classes aîné et benjamin agrégées au nœud étudié.

Nous avons une excentricité forte pour une classe lorsque celle-ci est éloignée du centre de gravité $\mathbf{G}$ du nuage. Plus l'excentricité est forte, plus la classe diffère de la moyenne et donc plus elle est porteuse de sens et mérite d'être exploitée. La mesure de l'excentricité de la classe q est donnée par : $d^2(\mathbf{g}_q, \mathbf{G})$. Il est intéressant d'étudier la contribution relative de la variable classifiante à l'excentricité de la classe q donnée par :

$$Cor_k(q) = \frac{\left(g_q^k\right)^2}{d^2(\mathbf{g}_q, \mathbf{G})}, \tag{7.14}$$

où g_q^k est la projection du centre de gravité $\mathbf{g}_q$ du sous-nuage I_q sur l'axe représentant la variable k (*cf.* figure 7.10). Ainsi, si la contribution est proche de 1, la variable k explique l'excentricité de la classe. Si le signe est négatif la variable est corrélée négativement à la classe.

Une classe constitue un sous-nuage, qui peut être étudiée par une analyse factorielle. Au lieu d'étudier la classe q, nous pouvons étudier le dipôle (a, b) de l'aîné et benjamin. Cette étude peut se faire par la variance. Ainsi un dipôle allongé dans la direction du premier axe factoriel du sous-nuage I_q, représente une variance élevée dans cette direction (*cf.* figure 7.11). L'indicateur utilisé pour la contribution d'une variable k à la divergence

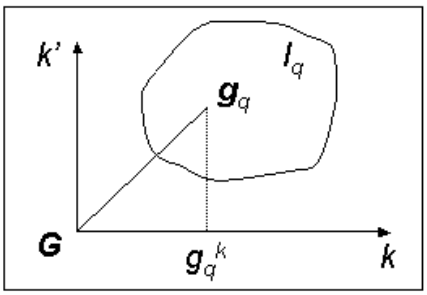


FIG. 7.10 – Représentation d'un sous-nuage I_q dans un plan de projection.

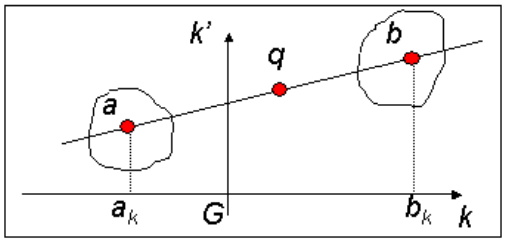


FIG. 7.11 – Caractérisation de la variance du dipôle dans une direction.

entre les deux classes est donné par :

$$Cod_k(q) = \frac{(a_k - b_k)^2}{d^2(a,b)}. \tag{7.15}$$

A l'aide de l'excentricité, nous pouvons ainsi étudier pourquoi les individus sont regroupés dans la classe étudiée, et pourquoi les nœuds aîné et benjamin sont séparés.

Nous proposons de suivre le plan suivant pour l'interprétation de la classification hiérarchique ascendante :

- La première chose est l'étude du dendrogramme et de la courbe des indices afin de déterminer la partition, ou les partitions à étudier. La coupure est réalisée au niveau du ou des sauts importants d'inertie.
- L'étape suivante est l'étude de toutes les classes formées par la ou les partitions plus fines. Il faut donc regarder quels sont les individus qui composent chaque classe. Il est de plus important de comprendre :
  - quelles sont les variables représentatives de chaque classe à l'aide de la contribution relative des variables classifiantes à l'excentricité de chaque classe,
  - quelles sont les variables qui séparent le dipôle formé de l'aîné et du benjamin pour chaque classe, à l'aide de la contribution de chaque variable à la variance du dipôle.

- Il est de plus intéressant de faire un tableau récapitulatif donnant pour chaque classe les individus qui y appartiennent, ainsi que les variables qui caractérisent chaque classe. Ce tableau permet de résumer simplement l'ensemble de l'interprétation.

7.4 Conclusion

Nous avons dans ce chapitre présenté uniquement deux méthodes (ou famille de méthodes) de classification : la méthode des centres mobiles et la classification hiérarchique ascendante. Les méthodes de classification sont cependant très nombreuses. Il existe entre autre une méthode dite de classification mixte (*hybrid classification*) qui est un mélange de la méthode des centres mobiles et de la classification hiérarchique. Elle est particulièrement bien adaptée aux tableaux de données comportant des milliers d'individus, pour lesquels le dendrogramme est difficile à lire. Les étapes de l'algorithme sont les suivantes :

- Une première étape consiste à appliquer la méthode des centres mobiles de façon à obtenir une partition de quelques dizaines, voire quelques centaines de groupes homogènes.
- Ensuite, la classification hiérarchique est appliquée sur ces groupes. Le dendrogramme et la courbe des indices permet de définir le nombre de classes finales à retenir.
- Une fois ce nombre déterminé, afin d'optimiser la classification, la méthode des centres mobiles est de nouveau appliquée à l'ensemble des individus de départ pour obtenir une partition correspondant à celle trouvée par le dendrogramme. Dans cette dernière étape les centres mobiles initiaux peuvent être considérés comme étant les barycentres des classes issues de la partition réalisée par la classification hiérarchique afin d'obtenir une convergence plus rapide.

La classification est une phase importante de l'analyse des données. Il est préférable de l'employer en complément des analyses factorielles (particulièrement la classification ascendante hiérarchique qui utilise la méthode de Ward pour l'agrégation). Il est conseillé d'appliquer la classification après les analyses factorielles. Cependant, les classes peuvent constituer des variables supplémentaires dans l'ACP, l'AFC ou encore l'ACM.

Glossaire

Indications historiques

- Bayes (Thomas) 1702-1761 : mathématicien anglais, il établit la relation liant les probabilités conditionnelles aux probabilités *a priori*.
- Benzécri (Jean-Paul) : mathématicien français, il est l'inventeur de l'analyse factorielle des correspondances (AFC) et le fondateur de l'école française d'analyse des données. Il s'intéressa en particulier aux données textuelles.
- Burt (Sir Cyril) 1883-1971 : psychologue britannique, innovateur certain d'un point de vue méthodologique en analyse de données, il est surtout connu pour ses fraudes scientifiques et ses falsifications d'observations.
- Huygens (Christiaan) 1629-1695 : également orthographié Huyghens, expérimentateur et théoricien néerlandais, il proposa un traité sur le calcul des probabilités. En mécanique, il développa la théorie du pendule qu'il appliqua pour réguler les mouvements d'horloges, et s'intéressa au problème du choc par la quantité de mouvement.
- Mahalanobis (Prasanta Chandra) 1893-1972 : physicien et mathématicien indien, il s'intéressa beaucoup aux statistiques. Il est surtout connu pour la distance qui porte son nom. Il étudia les analyses graphiques des fractiles (quantiles), et les statistiques *D-square*, appliqués à l'économie et à la biométrie. Il est un des premiers a avoir organisé le recueil de données en Inde.
- Minkowsky (Hermann) 1864-1909 : mathématicien allemand, il proposa une représentation de l'espace-temps à quatre dimensions qui fournit une interprétation géométrique de la relativité restreinte de A. Einstein qui fut son élève.
- Pearson (Karl) 1857-1936 : mathématicien anglais, il est un des premiers statisticiens. En particulier ses recherches étaient tournées vers l'hérédité.
- Tchebychev (Pafnouti Lvovitch) 1821-1894 : mathématicien russe, son nom est aussi écrit Chebyshov, Chebyshev, ou Tschebyscheff. Il est connu pour ses travaux dans le domaine de la probabilité et des statistiques, en particulier l'inégalité de Tchebychev qui permet de majorer des probabilités (grossièrement) et de démontrer le théorème de la loi faible des grands nombres.
- Voronoï (Georgi Fedoseevich) 1868-1908 : mathématicien russe, également transcrit Voronoy, il travailla sur la théorie des nombres, en particulier sur les nombres algébriques et la géométrie des nombres. En 1904, il rencontra Minkowski, et ils s'aperçurent qu'ils étudiaient des sujets similaires.

- Ward (Abraham) 1902-1950 : mathématicien, né en Hongrie, il partit à Vienne pour faire ses recherches. Sous l'occupation nazie, d'origine juive, il partit en 1938 aux Etats-Unis. Ses travaux concernèrent les espaces métriques et plus particulièrement les espaces vectoriels à dimension infinie. Il obtint également des résultats en géométrie différentielle.

Rappel de définitions

- Affectation : c'est une étape de classement.
- Caractères : données caractérisant les individus.
  - Caractère qualitatif : le caractère n'est pas mesurable.
    - Caractère qualitatif pur ou variable nominale : les modalités ne possèdent pas de structure d'ordre.
    - Caractère qualitatif ordonné ou variable ordinale : les modalités qualitatives sont ordonnées.
  - Caractère quantitatif : le caractère est mesurable, on y associe le nom de variable statistique (numérique).
    - Variable discrète : les valeurs prises par la variable sont des valeurs ponctuelles.
    - Variable continue : les valeurs prises par la variable sont numériques d'un intervalle donné.
- Classement : attribution d'éléments dans une classe préexistante.
- Classification : construction des classes les plus homogènes possibles dans un échantillon.
- Coefficient de corrélation linéaire : pour un tableau de données de I individus décrits par K variables et x_{ik} une donnée du tableau $i = 1, ..., I$ et $k = 1, ..., K$, le coefficient de corrélation linéaire entre deux variables k et k' est donné par :

$$r_{kk'} = \rho(x_k, x_{k'}) = \frac{\text{cov}(x_k, x_{k'})}{\sigma_k \sigma_{k'}}. \tag{7.16}$$

- Corrélation empirique ou covariance : pour un tableau de données de I individus décrits par K variables et x_{ik} une donnée du tableau $i = 1, ..., I$ et $k = 1, ..., K$, la corrélation empirique entre deux variables k et k' est donnée par :

$$\text{cov}(x_k, x_{k'}) = \frac{1}{I} \sum_{i \in I} \sum_{j \in I} (x_{ik} - \bar{x}_k)(x_{jk'} - \bar{x}_{k'}). \tag{7.17}$$

- Discrimination : la discrimination consiste à déterminer une fonction qui sépare au mieux les données selon un critère prédéfini.
- Dispersion : étalement des points déterminé par une distance. La dispersion d'un nuage sur un axe peut être vu comme l'inertie du nuage sur l'axe. Une mesure de dispersion est la variance.
- Distribution (ou série statistique) : les observations d'un caractère forment une distribution.
- Échantillon : sous-ensemble de la population.
- Effectif vérifiant un critère : nombre d'éléments vérifiant ce critère.
- Fonction de répartition : $F(x)$ est la proportion des individus de la population dont le caractère est inférieur à x.
- Individus ou unités statistiques : éléments de la population.

- Inertie : valeur caractérisant la concentration ou la dispersion de points sur un axe, un plan ou tout espace. L'inertie peut être représentée par une variance.
- Liaison : deux variables sont liées si elles ont un fort coefficient de corrélation linéaire ou encore si elles ne sont pas indépendantes.
- Modalité : les modalités d'un caractère sont les valeurs (mesurable ou non) prises par cette variable.
- Moment d'ordre r : Pour un tableau de données de I individus décrits par K variables et x_{ik} une donnée du tableau $i = 1, ..., I$ et $k = 1, ..., K$, le moment d'ordre r des individus est donné par :

$$x_k^r = \frac{1}{I} \sum_{i \in I} x_{ik}^r. \tag{7.18}$$

- Moyenne : pour un tableau de données de I individus décrits par K variables et x_{ik} une donnée du tableau $i = 1, ..., I$ et $k = 1, ..., K$, la moyenne des individus est donnée par :

$$\bar{x}_k = \frac{1}{I} \sum_{i \in I} x_{ik}. \tag{7.19}$$

- Population : ensemble des données étudiées.
- Quantile : pour un tableau de données de I individus décrits par K variables et x_{ik} une donnée du tableau $i = 1, ..., I$ et $k = 1, ..., K$, le quantile d'ordre α $(0 \leq \alpha \leq 1)$ est la racine de l'équation $F(x) = \alpha$, où F est la fonction de répartition.
- Ressemblance : deux individus se ressemblent, ou sont proches, s'ils possèdent des valeurs proches pour l'ensemble des variables.
- Tableau de contingence : c'est un tableau d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population.
- Taxonomie : littéralement la science des lois de l'ordre, c'est la science de la classification, parfois limitée à la botanique.
- Typologie : ensemble des limites des domaines connexes (patatoïdes) à faire sur chaque plan (pour les individus et les variables).
- Variance : pour un tableau de données de I individus décrits par K variables et x_{ik} une donnée du tableau $i = 1, ..., I$ et $k = 1, ..., K$, la variance des individus est donnée par :

$$\sigma_k^2 = \frac{1}{I} \sum_{i \in I} (x_{ik} - \bar{x}_k)^2. \tag{7.20}$$

Bibliographie

[Ben80a] J.P. Benzecri : *L'analyse de données (Tome 1) La taxinomie.* Dunod, 1980.

[Ben80b] J.P. Benzecri : *L'analyse de données (Tome 2) L'analyse des correspondances.* Dunod, 1980.

[Ber72] C. Berge : *Graphes et hypergraphes.* Dunod, 1972.

[BFRS93] L. Breiman, J.H. Friedman, R.A.Olshen et C.J. Stone : *Classification and regression tree.* Chapman and Hall, 1993.

[Bro03] G. Brossier : *Analyse des données*, chapitre Les éléments fondamentaux de la classification. Hermes Sciences publications, 2003.

[CDG+89] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier et H. Ralambondrainy : *Classification automatique des données.* Dunod, 1989.

[Cel03] G. Celeux : *Analyse des données*, chapitre Analyse discriminante. Hermes Sciences publications, 2003.

[DH97] P. Demartines et J. Hérault : Curvilinear component analysis : A self-organizing neural network for non linear mapping of data set. *IEEE Transactions on Neural Networks*, 8(1):148–154, Janvier 1997.

[EP90] B. Escoffier et J. Pagès : *Analyses factorielles simples et multiples - objectifs, méthodes et interprétations.* Dunod, 1990.

[Goa03] G. Goavert : *Analyse des données.* Hermes Sciences publications, 2003.

[HL03] G. Hébrail et Y. Lechevallier : *Analyse des données*, chapitre Data Mining et analyse des données. Hermes Sciences publications, 2003.

[Jam99a] M. Jambu : *Introduction au Data Mining.* Eyrolles, 1999.

[Jam99b] M. Jambu : *Méthodes de base de l'analyse de données.* Eyrolles, 1999.

[Kun00] M. Kunt : *Reconnaissance des formes et analyse de scènes.* Presses Polytechnique et universitaires romandes, 2000.

[LMP95] L. Lebart, A. Morineau et M. Piron : *Statistique exploratoire multidimensionnelle.* dunod, 1995.

[Mar04] A. Martin : La fusion d'informations, 2004.

[Pag03] J. Pagès : *Analyse des données*, chapitre Analyse factorielle des correspondances. Extensions et applications au traitement statistique des données sensorielles. Hermes Sciences publications, 2003.

[Pha96] D.T. PHAM : Blind separation of instantaneous mixture of sources via independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, 1996.

[Sap90] G. SAPORTA : *Probabilités Analyse des données et Statistique*. Edition Technip, 1990.

[Vap99] V.N. VAPNIK : *The nature of Statistical Learning Theory*. Springer, 1999.

Index