Publications de
l'Institut de Mathématiques
de Toulouse

Statistique Descriptive Multidimensionnelle

(pour les nuls)

(version de mai 2010)

Alain Baccini

Institut de Mathématiques de Toulouse — UMR CNRS 5219
Université Paul Sabatier — 31062 – Toulouse cedex 9.

Table des matières

Avant-propos

Ce document est consacré aux trois méthodes les plus courantes de la statitique descriptive multidimensionnelle : l'Analyse en Composantes Principales (chapitre 1), l'Analyse Factorielle des Correspondances (chapitre 2) et l'Analyse des Correspondances Multiples (chapitre 3).

Il a été conçu pour des personnes souhaitant avoir quelques connaissances sur ces méthodes sans avoir la moindre culture scientifique (d'où son sous-titre...). Les connaissances exposées ici sont donc, nécessairement, superficielles mais, nous l'espérons, suffisantes pour comprendre les grandes lignes de ces techniques.

La statistique multidimensionnelle (et principalement l'Analyse des Correspondances Multiples) est aujourd'hui couramment utilisée pour analyser des résultats d'enquêtes, y compris par des personnes n'ayant pas de formation mathématique ou statistique. Ce document leur est donc particulièrement destiné et fait suite au document intitulé "Statistique Descriptive Élémentaire", disponible sur le même site et désigné sous l'appellation "cours SDE" par la suite.

D'autre part, un autre cours sur la statistique multidimensionnelle, plus complet et destiné à des étudiants des filières universitaires de mathématiques appliquées, est également disponible sur ce site sous le titre "Exploration Statistique".

Chapitre 1

Analyse en Composantes Principales

Ce chapitre est consacré à l'Analyse en Composantes Principales (ou A.C.P.), méthode fondamentale en statistique descriptive multidimensionnelle. Cette méthode permet de traiter simultanément un nombre quelconque de variables, toutes quantitatives.

Dans un premier paragraphe, nous donnerons tout d'abord quelques indications sur ce que sont les méthodes de la statistique descriptive multidimensionnelle. Ensuite, nous présenterons en détail un exemple très simple (un exemple d'école, artificiel), pour bien comprendre comment fonctionne une A.C.P., à quoi ça sert, comment on l'interprète... Enfin, dans un dernier paragraphe, nous donnerons quelques indications générales sur cette méthode.

1.1 La statistique descriptive multidimensionnelle

On désigne par statistique descriptive multidimensionnelle l'ensemble des méthodes de la **statistique descriptive** (ou exploratoire) permettant de traiter simultanément un **nombre quelconque de variables** (il s'agit d'aller au-delà de l'étude d'une seule ou de deux variables). Ces méthodes sont purement descriptives, c'est-à-dire qu'elles ne supposent, a priori, aucun modèle sous-jacent, de type probabiliste. (Ainsi, lorsqu'on considère un ensemble de variables quantitatives sur lesquelles on souhaite réaliser une A.C.P., il n'est pas nécessaire de supposer que ces variables sont distribuées selon des lois normales.)

Dans chaque méthode que nous allons développer, les variables considérées seront de même nature : toutes **quantitatives** (Analyse en Composantes Principales) ou toutes **qualitatives** (Analyses des Correspondances).

Les méthodes les plus classiques de la statistique descriptive multidimensionnelle sont les méthodes factorielles. Elles consistent à rechercher des **facteurs** (cette notion sera précisée ultérieurement) en nombre restreint et résumant le mieux possible les données considérées.

Elles aboutissent à des représentations graphiques des données (des individus comme des variables) par rapport à ces facteurs, représentés comme des axes. Ces représentations graphiques sont du type **nuage de points** (ou diagramme de dispersion).

Nous allons développer 3 méthodes, chacune correspondant à un chapitre : l'Analyse en Composantes Principales (A.C.P.), dans ce chapitre 1, l'Analyse Factorielle des Correspondances (A.F.C.), dans le chapitre 2 et l'Analyse des Correspondances Multiples (A.C.M.), dans le chapitre 3.

Nous laisserons de côté l'Analyse Factorielle Discriminante et l'Analyse Canonique (méthodes factorielles plus particulières), ainsi que les méthodes non factorielles (principalement la classification).

La logique des trois chapitres consacrés à la statistique descriptive multidimensionnelle est la suivante : l'objectif, pour les étudiants, est de maîtriser, au moins partiellement, l'Analyse des Correspondances Multiples, méthode souvent utilisée dans les **dépouillements d'enquêtes**, lorsqu'on souhaite aller au-delà des simples **tris à plat** (analyses unidimensionnelles) ou **tris croisés** (analyses bidimensionnelles). On commence donc par introduire l'A.C.P., méthode centrale, indispensable pour bien comprendre le fonctionnement de toute technique factorielle. On développe

ensuite l'A.F.C., cas particulier de l'A.C.M. lorsqu'on ne considère que deux variables qualitatives. On généralise enfin à l'A.C.M.

1.2 Exemple illustratif pour l'A.C.P.

Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique approprié, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus telle qu'elle est définie par l'ensemble des variables initiales (ainsi remplacées par les facteurs).

C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d'indicateurs numériques, appelés aides à l'interprétation, qui sont là pour aider l'utilisateur à faire l'interprétation la plus juste et la plus objective possible.

Sur le plan théorique, l'Analyse en Composantes Principales est une méthode relativement complexe, dans la mesure où elle fait appel à des notions mathématiques non élémentaires : celles de matrices, d'éléments propres... Fort heureusement, il n'est pas nécessaire de connaître ces notions pour comprendre le mécanisme d'une A.C.P. et donc pour l'utiliser correctement. Pour faciliter la tâche du lecteur, nous avons choisi de présenter l'A.C.P. à travers son déroulement sur un exemple fictif, très simple, et qui parlera à tout le monde : les notes obtenues par des élèves dans diverses disciplines.

1.2.1 Présentation

Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

On sait comment analyser séparément chacune de ces 4 variables, soit en faisant un **graphique**, soit en calculant des **résumés numériques**. Nous savons également qu'on peut regarder les **liaisons entre 2 variables** (par exemple mathématiques et français), soit en faisant un graphique du type nuage de points, soit en calculant leur **coefficient de corrélation linéaire**, voire en réalisant la **régression** de l'une sur l'autre (pour tout cela, se reporter au cours SDE).

Mais, comment faire une étude simultanée des 4 variables, ne serait-ce qu'en réalisant un graphique ? La difficulté vient de ce que les individus (les élèves) ne sont plus représentés dans un

plan, espace de dimension 2, mais dans un espace de dimension 4 (chaque élève étant caractérisé par les 4 notes qu'il a obtenues). L'objectif de l'Analyse en Composantes Principales est de revenir à un espace de dimension réduite (par exemple, ici, 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent des données initiales.

Par analogie, on peut penser au photographe qui cherche le meilleur angle de vue pour transcrire en dimension 2 (le plan de sa photo) une scène située en dimension 3 (notre espace ambiant). La méthode mathématique va se charger de trouver l'"angle de vue" optimal, se substituant ainsi au coup d'œil du photographe...

Nous présentons ci-dessous quelques résultats de l'A.C.P. réalisée, avec le logiciel SAS, sur ces données. Cela va permettre de se rendre compte des possibilités de la méthode. On notera que l'on s'est limité à 2 décimales dans les résultats, bien que les logiciels en fournissent, en général, beaucoup plus (mais elles sont rarement utiles).

1.2.2 Résultats préliminaires

Le logiciel fournit tout d'abord la moyenne (*mean*), l'écart-type (*standard deviation*), le minimum et le maximum de chaque variable. Il s'agit donc, pour l'instant, d'**études univariées**.

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Notons au passage la grande homogénéité des 4 variables considérées : même ordre de grandeur pour les moyennes, les écarts-types, les minima et les maxima.

Le tableau suivant est la **matrice des corrélations**. Elle donne les coefficients de corrélation linéaire des variables prises deux à deux. C'est une succession d'**analyses bivariées**, constituant un premier pas vers l'**analyse multivariée**.

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Remarquons que toutes les corrélations linéaires sont positives (ce qui signifie que toutes les variables varient, en moyenne, dans le même sens), certaines étant très fortes (0.98 et 0.95), d'autres moyennes (0.65 et 0.51), d'autres enfin plutôt faibles (0.40 et 0.23).

1.2.3 Résultats généraux

Continuons l'examen des sorties de cette analyse par l'étude de la **matrice des variances-covariances**, matrice de même nature que celle des corrélations, bien que moins "parlante" (nous verrons néanmoins plus loin comment elle est utilisée concrètement). La diagonale de cette matrice fournit les variances des 4 variables considérées (on notera qu'au niveau des calculs, il est plus commode de manipuler la variance que l'écart-type ; pour cette raison, dans de nombreuses méthodes statistiques, comme l'A.C.P., on utilise la variance pour prendre en compte la dispersion d'une variable quantitative).

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Les **valeurs propres** (éléments mathématiques dont la signification peut être laissée de côté pour l'instant) données ci-dessous sont celles de la matrice des variances-covariances.

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
	40.30	1.00	

Interprétation

Chaque ligne du tableau ci-dessus correspond à une variable virtuelle (voilà les **facteurs** !) dont la colonne VAL. PR. (valeur propre) fournit la variance (en fait, chaque valeur propre représente la variance du facteur correspondant). La colonne PCT. VAR, ou pourcentage de variance, correspond au pourcentage de variance de chaque ligne par rapport au total. La colonne PCT. CUM., ou pourcentage cumulé, représente le cumul de ces pourcentages.

Additionnons maintenant les variances des 4 variables initiales (diagonale de la matrice des variances-covariances) : $11.39 + 8.94 + 12.06 + 7.91 = 40.30$. La dispersion totale des individus considérés, en dimension 4, est ainsi égale à 40.30.

Additionnons par ailleurs les 4 valeurs propres obtenues : $28.23 + 12.03 + 0.03 + 0.01 = 40.30$. Le nuage de points en dimension 4 est toujours le même et sa dispersion globale n'a pas changé. C'est la répartition de cette dispersion, selon les nouvelles variables que sont les facteurs, ou composantes principales, qui se trouve modifiée : les 2 premiers facteurs restituent à eux seuls la quasi-totalité de la dispersion du nuage, ce qui permet de négliger les 2 autres.

Par conséquent, les graphiques en dimension 2 présentés ci-dessous résument presque parfaitement la configuration réelle des données qui se trouvent en dimension 4 : l'objectif (résumé pertinent des données en petite dimension) est donc atteint.

1.2.4 Résultats sur les variables

Le résultat fondamental concernant les variables est le tableau des **corrélations variables-facteurs**. Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les facteurs. Ce sont ces corrélations qui vont permettre de donner un sens aux facteurs (de les interpréter).

Corrélations variables-facteurs

FACTEURS -->	F1	F2	F3	F4
MATH	0.81	-0.58	0.01	-0.02
PHYS	0.90	-0.43	-0.03	0.02
FRAN	0.75	0.66	-0.02	-0.01
ANGL	0.91	0.40	0.05	0.01

Les deux premières colonnes de ce tableau permettent, tout d'abord, de réaliser le **graphique des variables** donné par la Fig. 1.1.

Mais, ces deux colonnes permettent également de donner une signification aux facteurs (donc aux axes des graphiques).

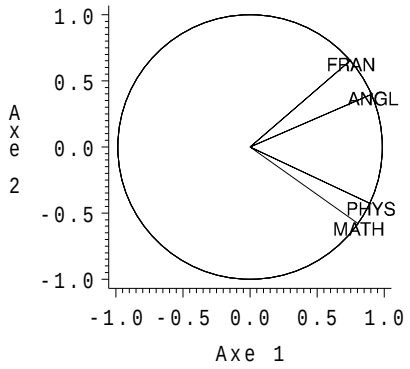

FIG. 1.1 – *Représentation des variables*

On notera que les deux dernières colonnes ne seront pas utilisées puisqu'on ne retient que deux dimensions pour interpréter l'analyse.

Interprétation

Ainsi, on voit que le premier facteur est correlé positivement, et assez fortement, avec chacune des 4 variables initiales : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score élevé sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif ; l'axe 1 représente donc, en quelques sortes, le résultat global (dans l'ensemble des 4 disciplines considérées) des élèves. En ce qui concerne l'axe 2, il oppose, d'une part, le français et l'anglais (corrélations positives), d'autre part, les mathématiques et la physique (corrélations négatives). Il s'agit donc d'un axe d'opposition entre disciplines littéraires et disciplines scientifiques, surtout marqué par l'opposition entre le français et les mathématiques. Cette interprétation, qui est déjà assez claire, peut être précisée avec graphiques et tableaux relatifs aux individus. Nous les présentons maintenant.

1.2.5 Résultats sur les individus

Le tableau donné ci-dessous contient tous les résultats importants de l'A.C.P. sur les individus.

Coordonnées des individus ; contributions ; cosinus carrés

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	0.11	-8.61	-1.41	20.99	29.19	1.83	0.97	0.03
alan	0.11	-3.88	-0.50	4.22	5.92	0.23	0.98	0.02
anni	0.11	-3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	0.11	6.41	-2.05	12.48	16.15	3.87	0.91	0.09
andr	0.11	-3.03	-4.92	9.22	3.62	22.37	0.28	0.72
pier	0.11	-1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	0.11	1.95	-4.20	5.93	1.50	16.29	0.18	0.82
evel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73

On notera que chaque individu représente 1 élément sur 9, d'où un poids (une pondération) de $1/9 = 0.11$, ce qui est fourni par la première colonne du tableau.

Les 2 colonnes suivantes fournissent les coordonnées des individus (les élèves) sur les deux premiers axes (les facteurs) et ont donc permis de réaliser le **graphique des individus**. Ce dernier (Fig. 1.2) permet de préciser la signification des axes, donc des facteurs.

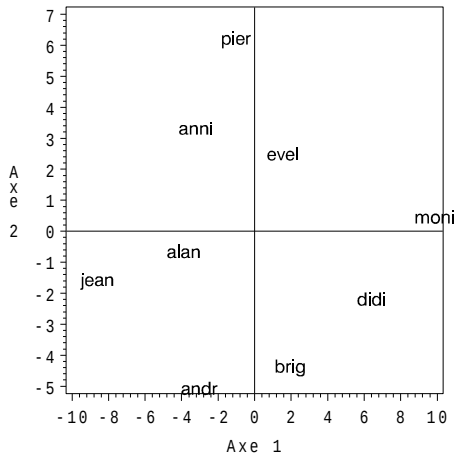

FIG. 1.2 – *Représentation des individus*

La signification et l'utilisation des dernières colonnes du tableau seront explicitées un peu plus loin.

Interprétation

On confirme ainsi que l'axe 1 représente le résultat d'ensemble des élèves : si on prend leur score – ou coordonnée – sur l'axe 1, on obtient le même classement que si on prend leur moyenne générale. Par ailleurs, l'élève "le plus haut" sur le graphique, celui qui a la coordonnée la plus élevée sur l'axe 2, est Pierre dont les résultats sont les plus contrastés en faveur des disciplines littéraires (14 et 11.5 contre 7 et 5.5). C'est exactement le contraire pour André qui obtient la moyenne dans les disciplines scientifiques (11 et 10) mais des résultats très faibles dans les disciplines littéraires (7 et 5.5). On notera que Monique et Alain ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines (mais à des niveaux très distincts, ce qu'a déjà révélé l'axe 1). L'axe 2 oppose bien les "littéraires" (en haut) aux "scientifiques" (en bas).

Les 3 colonnes suivantes du tableau fournissent des **contributions** des individus à diverses dispersions : CONT1 et CONT2 donnent les contributions (en pourcentages) des individus à la variance selon les axes 1 et 2 (rappelons que l'on utilise ici la variance pour mesurer la dispersion) ; CONTG donne les contributions générales, c'est-à-dire à la dispersion en dimension 4 (il s'agit de ce que l'on appelle l'**inertie** du nuage des élèves ; la notion d'inertie généralise celle de variance en dimension quelconque, la variance étant toujours relative à une seule variable). Ces contributions sont fournies en pourcentages (chaque colonne somme à 100) et permettent de repérer les individus les plus importants au niveau de chaque axe (ou du nuage en dimension 4). Elles servent en général à affiner l'interprétation des résultats de l'analyse.

Ainsi, par exemple, la variance de l'axe 1 vaut 28.23 (première valeur propre). On peut la retrouver en utilisant la formule de définition de la variance :

$$Var(C^1) = \frac{1}{9}\sum_{i=1}^{9}(c_i^1)^2$$

(il faut noter que, dans une A.C.P., les variables étant centrées, il en va de même pour les facteurs ; ainsi, la moyenne de C^1 est nulle et n'apparaît pas dans la formule de la variance). La coordonnée de Jean (le premier individu du fichier) sur l'axe 1 vaut $c_1^1 = -8.61$; sa contribution est donc :

$$\frac{\frac{1}{9}(-8.61)^2}{28.23} \times 100 = 29.19\ \%.$$

À lui seul, cet individu représente près de 30 % de la variance : il est prépondérant (au même titre que Monique) dans la définition de l'axe 1 ; cela provient du fait qu'il a le résultat le plus faible, Monique ayant, à l'opposé, le résultat le meilleur.

Enfin, les 2 dernières colonnes du tableau sont des cosinus carrés qui fournissent la qualité de la représentation de chaque individu sur chaque axe. Ces quantités s'additionnent axe par axe, de

sorte que, en dimension 2, Évelyne est représentée à 98 % (0.25 + 0.73), tandis que les 8 autres individus le sont à 100 %.

Précisons un peu cette notion. Lorsqu'on considère les données initiales, chaque individu (chaque élève) est représenté par un vecteur dans un espace de dimension 4 (les éléments – ou coordonnées – de ce vecteur sont les notes obtenues dans les 4 disciplines). Lorsqu'on résume les données en dimension 2, et donc qu'on les représente dans un plan, chaque individu est alors représenté par la projection du vecteur initial sur le plan en question. Le cosinus carré relativement aux deux premières dimensions (par exemple, pour Évelyne, 0.98 ou 98 %) est celui de l'angle formé par le vecteur initial et sa projection dans le plan. Plus le vecteur initial est proche du plan, plus l'angle en question est petit et plus le cosinus, et son carré, sont proches de 1 (ou de 100 %) : la représentation est alors très bonne. Au contraire, plus le vecteur initial est loin du plan, plus l'angle en question est grand (proche de 90 degrés) et plus le cosinus, et son carré, sont proches de 0 (ou de 0 %) : la représentation est alors très mauvaise. On utilise les carrés des cosinus parce qu'ils s'additionnent suivant les différentes dimensions, contrairement à leurs racines.

1.3 Présentation générale de la méthode

Dans ce paragraphe, on expose de façon plus générale ce qu'est l'Analyse en Composantes Principales. Nous sommes donc amenés à faire quelques développements techniques rendant ce paragraphe plus délicat à suivre que le précédent. Une parfaite assimilation de son contenu n'est pas indispensable pour le lecteur, surtout s'il n'est que peu familiarisé avec les aspects mathématiques abordés dans le point 1.3.1. Toutefois, une bonne compréhension des idées directrices de la méthode nous semble nécessaire.

Le principe général de l'A.C.P. est de réduire la dimension des données initiales (qui est p si l'on considère p variables quantitatives), en remplaçant les p variables initiales par q facteurs appropriés ($q < p$).

Les données, toujours centrées, doivent en plus être réduites lorsque les variables sont hétérogènes. Les q facteurs cherchés sont des moyennes pondérées des variables initiales. Leur choix se fait en maximisant la dispersion des individus selon ces facteurs (autrement dit, les facteurs retenus doivent être de variance maximum). Des techniques mathématiques appropiées permettent de réaliser tout cela de façon automatique et optimale.

Lorqu'on a obtenu les résultats d'une A.C.P., il faut être capable de les interpréter. Pour cela, on dispose de graphiques, à la fois pour les variables et pour les individus, ainsi que d'indicateurs numériques, appelés aides à l'interprétation. Ces indicateurs permettent, en association avec les graphiques, de comprendre les éléments clés de la structure des données initiales, et donc d'en faire une interprétation correcte.

Le premier point ci-desous est consacré aux aspects techniques, mathématiques, de l'A.C.P. Autrement dit, on essaye d'y expliquer ce que contient la "boîte noire" qu'est cette méthode. Le second point décrit les résultats obtenus, autrement dit les sorties de la "boîte noire", et les lignes directrices que l'on doit suivre pour les interpréter correctement.

1.3.1 Les principes

Les données à analyser

On considère p **variables quantitatives**, notées $X^1, \ldots, X^j, \ldots, X^p$, observées sur n individus, notés $1, \ldots, i, \ldots, n$. L'observation de la variable X^j sur l'individu i, $X^j(i)$, sera plus simplement notée x_i^j. Les données se présentent ainsi sous la forme d'un tableau du type suivant :

	X^1	$\cdots$	X^j	$\cdots$	X^p
1	x_1^1	$\cdots$	x_1^j	$\cdots$	x_1^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_i^1	$\cdots$	x_i^j	$\cdots$	x_i^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_n^1	$\cdots$	x_n^j	$\cdots$	x_n^p

Noter que le nombre p de variables d'une A.C.P. vaut au moins 2 ; le plus souvent, p est de l'ordre de 10 (ou de quelques dizaines). De son côté, le nombre n d'individus est au moins égal à p ; le plus souvent, il vaut plusieurs dizaines (voire plusieurs centaines).

Le problème à traiter

On cherche à extraire l'information pertinente contenue dans le tableau des données. Pour cela, on va le **résumer** en extrayant l'essentiel de sa structure en vue de faire des **représentations graphiques** à la fois fidèles aux données initiales et commodes à interpréter. Ces représentations devront se faire en dimension réduite : le nuage initial, situé dans un espace de dimension p (puisqu'on dispose, au départ, de p variables quantitatives), sera résumé (réduit, projeté) en dimension q (grâce à l'obtention de q **facteurs** : voir la définition de ce terme plus bas). Le nombre q de facteurs retenus sera compris entre 1 et p ; le plus souvent, il vaudra 2 ou 3.

Le critère utilisé

Les q facteurs que l'on va définir, pour résumer l'information contenue dans le tableau initial, doivent **maximiser la dispersion** du nuage des observations. Rappelons que la dispersion d'une variable quantitative se mesure, en général, par sa **variance** (ou par son **écart-type**, racine carrée positive de la variance). Plus généralement, lorsqu'on dispose d'un nuage d'observations en plusieurs dimensions, on parle d'**inertie** (somme des variances des variables considérées). Le principe de l'A.C.P. consiste donc à rechercher, pour une dimension q restreinte (2 ou 3), les q facteurs maximisant l'inertie du nuage lorsqu'on le projette (le résume) dans le sous-espace de dimension q engendré par ces facteurs : en passant de la dimension initiale p à la dimension réduite q, on perd, obligatoirement, de la dispersion, de l'inertie. L'idée est d'en perdre le moins possible en choisissant convenablement les facteurs.

La méthode

On cherche des **combinaisons linéaires** des variables initiales, appelées **facteurs**, ou encore **composantes principales**, s'écrivant sous la forme suivante (penser à la moyenne pondérée des notes d'un groupe d'élèves à l'issue du bac ; c'est la même chose, en plus général) :

$$C^1 = a_1^1 X^1 + a_1^2 X^2 + \cdots + a_1^p X^p$$
$$C^2 = a_2^1 X^1 + a_2^2 X^2 + \cdots + a_2^p X^p$$
$$\cdots$$

telles que :

C^1 doit contenir un maximum d'"information", c'est-à-dire disperser le plus possible les individus.
L'idée est la suivante : si on dispose d'un nuage de points dans le plan (autrement dit, en dimension $p = 2$) et qu'on souhaite le projeter sur une droite (donc en dimension $q = 1$), la droite la plus "fidèle" à la configuration initiale est celle qui rend maximum la dispersion – la variance – du nuage après sa projection (essayer de faire un dessin).
Le critère choisi est, de façon naturelle, $\mathrm{var}(C^1)$ maximum. Pour des raisons techniques, on doit rajouter la contrainte $\sum_{j=1}^p (a_1^j)^2 = 1$.

On fait la même chose pour C^2, en imposant, en plus, que C^1 et C^2 soient non correlées (pour que l'information apportée par C^2 soit complètement nouvelle par rapport à l'information contenue dans C^1).

Et ainsi de suite ...

On pourra ainsi se contenter d'un petit nombre de facteurs (2 ou 3) pour réaliser des graphiques faciles à lire et à interpréter.

Centrage ou réduction des données ?

Tout d'abord, il faut noter que le centrage des variables d'un tableau soumis à une A.C.P. (on retranche à chaque observation la moyenne de la variable correspondante) ne modifie en rien les résultats de l'A.C.P. En effet, on utilise comme critère la maximisation de la dispersion (de l'inertie) et la dispersion d'une variable n'est pas modifiée par son centrage. Comme il est plus commode de travailler avec des données centrées (les expressions manipulées sont plus simples à écrire), les A.C.P. sont systématiquement réalisées après centrage de chaque variable.

Dans la pratique, on peut ainsi faire soit une **A.C.P. centrée** (les variables X^j considérées sont seulement centrées), soit une **A.C.P. réduite** (les variables sont centrées et réduites : on divise chaque donnée centrée par l'écart-type de la variable correspondante).

On recommande l'A.C.P. seulement centrée lorsque les variables sont **homogènes** : même signification, même unité de mesure, même ordre de grandeur... C'est le cas de l'exemple traité au paragraphe précédent. Au contraire, on recommande l'A.C.P. réduite lorsque les variables sont **hétérogènes**, c'est-à-dire dans les autres cas.

Les outils mathématiques (pour lecteur averti !)

Il s'agit des outils de l'algèbre linéaire, essentiellement les notions de **vecteurs propres** et de **valeurs propres**. Notons $\mathbf{S}$ la matrice $p \times p$ des variances-covariances des variables X^j et $\mathbf{R}$ la matrice $p \times p$ de leurs corrélations linéaires. Dans une A.C.P. seulement centrée, C^1 est le vecteur propre normé de $\mathbf{S}$ associé à la plus grande valeur propre ($\mathbf{S}C^1 = \lambda_1 C^1$ et $\|C^1\| = 1$), C^2 est le vecteur propre normé de $\mathbf{S}$ associé à la seconde plus grande valeur propre, et ainsi de suite. De plus, les différents vecteurs C^k sont orthogonaux (à la non corrélation des variables centrées correspond l'orthogonalité des vecteurs qui les représentent). Dans une A.C.P. réduite, les C^k sont les vecteurs propres orthonormés de la matrice $\mathbf{R}$.

Commentaires

On notera que les différents calculs permettant d'obtenir les résultats d'une A.C.P. (définition des facteurs, calcul de leur variance – les valeurs propres –, détermination des corrélations variables-facteurs, des coordonnées des individus...) ne sont en général pas réalisables "à la main" (pas plus qu'avec une calculette d'ailleurs). Seul l'usage d'un ordinateur et d'un logiciel spécialisé, utilisant un algorithme approprié, peut permettre d'obtenir ces résultats.

1.3.2 Les résultats

Résultats généraux

Avant d'analyser les résultats proprement dits d'une A.C.P., il est bon d'en regarder les **résultats préliminaires**. Tout d'abord, pour chaque variable considérée, son minimum, son maximum, sa moyenne et son écart-type. Cela permet d'avoir une première connaissance des données étudiées et, le cas échéant, de décider si l'A.C.P. doit être réduite on non.

Il est également intéressant d'étudier la **matrice des corrélations** entre variables initiales, dans la mesure où elle permet d'avoir une première idée de la structure de corrélation entre ces variables.

Ensuite, le premier tableau de résultats à regarder est le **tableau des pourcentages d'inertie** correspondants aux différentes valeurs propres, contenant aussi les pourcentages cumulés associés : ce tableau va permettre de choisir la dimension q retenue pour interpréter l'A.C.P.

Résultats sur les variables

La technique de l'A.C.P. permet de calculer les **corrélations variables-facteurs**, autrement dit les coefficients de corrélation linéaire entre chaque variable initiale et chaque facteur retenu.

Dans un premier temps, ces quantités permettent un début d'interprétation des facteurs, dans la mesure où elles indiquent comment ils sont liés aux variables initiales. À ce stade, il est recommandé d'utiliser aussi la **matrice des corrélations** entre variables initiales, pour compléter cette interprétation.

Dans un second temps, les corrélations variables-facteurs permettent de réaliser les **graphiques des variables** dont l'étude détaillée conduit à préciser la signification des axes, c'est-à-dire des facteurs. On doit considérer uniquement le graphique selon les axes 1 et 2 si l'on a choisi $q = 2$; on doit au contraire considérer les 3 graphiques selon les axes 1 et 2, 1 et 3, 2 et 3, si l'on a choisi $q = 3$.

Résultats sur les individus

Là encore, la technique de l'A.C.P. permet de calculer les **coordonnées des individus sur les axes**, leurs **contributions à la dispersion** selon chacun de ces axes (ainsi que leurs contributions à la dispersion globale, selon les p dimensions) et les **cosinus carrés**.

Les coordonnées permettent de réaliser les **graphiques des individus** (1 ou 3 graphiques, selon que l'on a choisi $q = 2$ ou $q = 3$). Concernant ces graphiques, il faut tout d'abord noter que leurs axes s'interprètent de la même manière que les axes des graphiques des variables : les uns comme les autres sont associés aux facteurs.

En associant à ces graphiques les contributions des individus aux axes, on peut affiner l'interprétation de ces axes : chacun d'entre eux est surtout déterminé par les quelques individus présentant les plus fortes contributions ; ce sont en général ceux situés en position extrême sur l'axe, c'est-à-dire y ayant les plus fortes coordonnées, soit positives soit négatives. Bien sûr, avant d'utiliser un tel individu pour affiner l'interprétation d'un axe, il faut s'assurer que cet individu est bien représenté sur cet axe, autrement dit que le cosinus carré correspondant est grand (proche de 1).

Chapitre 2

Analyse Factorielle des Correspondances

L'Analyse Factorielle des Correspondances (A.F.C.) est une méthode factorielle de Statistique Descriptive Multidimensionnelle (voir la première section du chapitre 1).

Son objectif est d'analyser la liaison existant entre deux variables qualitatives (si on dispose de plus de deux variables qualitatives, on aura recours à l'Analyse des Correspondances Multiples, méthode exposée dans le chapitre 3). Ainsi, avant de mettre en œuvre une A.F.C., il faut s'assurer que cette liaison existe bien. Pour cela, il existe des graphiques (diagrammes en barres de profils) et des caractéristiques numériques (indice khi-deux et ses dérivés) permettant de mettre en évidence une telle liaison lorsqu'elle existe (voir le cours de statistique descriptive élémentaire, ici noté SDE). On notera qu'on dispose aussi d'un test statistique, le test du khi-deux d'indépendance, basé sur l'indice khi-deux, permettant de tester s'il existe ou non une liaison significative entre deux variables qualitatives. Ce test est très simple à mettre en œuvre mais ne relève pas de la statistique descriptive.

L'A.F.C. est, en fait, une Analyse en Composantes Principales (A.C.P. ; voir le chapitre 1) particulière, réalisée sur les profils associés à la table de contingence croisant les deux variables considérées. Plus précisemment, l'A.F.C. consiste à réaliser une A.C.P. sur les profils-lignes et une autre sur les profils-colonnes. Les résultats graphiques de ces deux analyses sont ensuite superposés pour produire un graphique (éventuellement plusieurs) de type nuage de points, dans lequel sont réunies les modalités des deux variables considérées, ce qui permet d'étudier les correspondances entre ces modalités, autrement dit la liaison entre les deux variables.

2.1 Principe général de l'A.F.C.

L'A.F.C. étant une A.C.P. particulière, nous ne donnons pas trop de détails techniques sur cette méthode. On en donne juste les grandes lignes dans ce paragraphe. Ensuite, dans le paragraphe 2, on illustre en détails la méthode sur un exemple.

2.1.1 Les données

On considère deux variables qualitatives : X à r modalités notées $x_1, \ldots, x_\ell, \ldots, x_r$; Y à c modalités notées $y_1, \ldots, y_h, \ldots, y_c$; on les observe simultanément sur n individus (ayant ici obligatoirement tous le même poids $\frac{1}{n}$). On sait que ces données peuvent être présentées sous la forme d'une table de contingence, ou tableau à double entrée :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Des précisions sur une telle table de contingence se trouvent dans le chapitre 3 du cours SDE. En particulier, on y trouve les définitions des effectifs conjoints (les $n_{\ell h}$) et des effectifs marginaux (les $n_{\ell +}$ et les n_{+h}).

2.1.2 Le problème

On suppose qu'il existe une liaison entre X et Y, et on cherche à décrire, à expliciter, cette liaison.

Pour cela, on se base sur l'étude des profils-lignes et des profils-colonnes. Rappelons la définition du $\ell^{\text{ième}}$ profil-ligne

$$\{\frac{n_{\ell 1}}{n_{\ell +}}, \ldots, \frac{n_{\ell h}}{n_{\ell +}}, \ldots, \frac{n_{\ell c}}{n_{\ell +}}\},$$

et celle du $h^{\text{ième}}$ profil-colonne

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\}.$$

Rappelons encore que la liaison entre les deux variables est d'autant plus grande que les profils (lignes ou colonnes) sont différents. C'est donc par l'étude des ressemblances et des différences entre profils-lignes et entre profils-colonnes que l'on peut analyser la liaison entre les deux variables. Cette analyse va se faire au moyen de deux Analyses en Composantes Principales (A.C.P.) un peu particulières.

2.1.3 La méthode

On réalise l'A.C.P. du tableau des profils-lignes (les individus de cette A.C.P. sont les lignes de la table de contingence, c'est-à-dire les modalités de X) et l'on fait la représentation graphique des individus, donc des modalités de X (dans cette A.C.P. particulière, on ne s'intéresse pas au graphique des variables). On a un seul graphique si on ne conserve que deux dimensions, plusieurs dans le cas contraire.

On réalise d'autre part l'A.C.P. du tableau des profils-colonnes (les individus de cette A.C.P. sont maintenant les colonnes de la table de contingence, c'est-à-dire les modalités de Y) et l'on fait la représentation graphique des individus, donc des modalités de Y.

On montre que ces deux A.C.P. se correspondent (ce qui est normal, puisque leurs données sont extraites de la même table de contingence) et qu'il est donc légitime de superposer les deux représentations graphiques. On obtient ainsi un graphique de type nuage de points (ou un ensemble de graphiques si on conserve plus de deux dimensions), représentant à la fois les modalités de X et celles de Y.

C'est l'interprétation de ce(s) graphique(s), pour laquelle on dispose d'un certain nombre d'indicateurs, qui permet d'expliciter la liaison entre les deux variables considérées. En particulier, on s'attache à étudier les correspondances entre les modalités de X et celles de Y, d'où le nom de la méthode.

Signalons que la distance entre profils (lignes ou colonnes), utilisée pour réaliser chaque A.C.P., est un peu particulière : ce n'est pas la distance usuelle, mais la distance dite "du khi-deux". Elle est expliquée dans le point 2.2.2, avec la notion d'inertie.

2.2 Exemple illustratif

L'exemple considéré dans ce paragraphe est relatif aux exploitations agricoles de la région Midi-Pyrénées. Les données proviennent des "Tableaux Économiques de Midi-Pyrénées", publiés par la Direction Régionale de Toulouse de l'INSEE, en 1996 (données relatives à l'année 1993 ; chiffres arrondis à la dizaine près).

2.2.1 Les données

Elles sont reproduites ci-dessous.

Exemple 1 *Répartition des exploitations agricoles de la région Midi-Pyrénées selon le département et la S.A.U. (en 1993).*

	INF05	S0510	S1020	S2035	S3550	SUP50
ARIE	870	330	730	680	470	890
AVER	820	1260	2460	3330	2170	2960
H.G.	2290	1070	1420	1830	1260	2330
GERS	1650	890	1350	2540	2090	3230
LOT	1940	1130	1750	1660	770	1140
H.P.	2110	1170	1640	1500	550	430
TARN	1770	820	1260	2010	1680	2090
T.G.	1740	920	1560	2210	990	1240

Les 73 000 exploitations agricoles de la région Midi-Pyrénées ont été ventilées dans cette table de contingence selon le département (en lignes, 8 modalités) et la S.A.U. (Surface Agricole Utilisée, en colonnes, 6 classes).

Codes des départements : ARIE = Ariège ; AVER = Aveyron ; H.G. = Haute-Garonne ; GERS = Gers ; LOT = Lot ; H.P. = Hautes-Pyrénées ; TARN = Tarn ; T.G. = Tarn-et-Garonne.

Codes des classes de S.A.U. : INF05 = moins de 5 hectares ; S0510 = entre 5 et 10 hectares... ; SUP50 = plus de 50 hectares.

On notera que la deuxième variable n'est pas qualitative, mais quantitative continue. En fait, la méthode la considère comme qualitative, ce qui signifie que l'ordre naturel sur les classes n'est pas du tout pris en compte. On pourra toujours essayer de retrouver cet ordre lorsqu'on interprètera le graphique, mais ce sera un complément par rapport à l'A.F.C. proprement dite.

Remarque 1 *En statistique, on parle en général de* variable catégorielle *pour désigner soit une variable qualitative (nominale ou ordinale), soit une variable quantitative (discrète ou continue), lorsque les modalités, valeurs ou classes sont considérées comme des* catégories*, sans aucune structure (structure d'ordre entre les modalités ou les classes, structure numérique – celle de l'ensemble des nombres réels – entre les valeurs). Toute variable prise en compte dans une A.F.C. est systématiquement considérée comme catégorielle. C'est à l'utilisateur de regarder ensuite, sur le(s) graphique(s) obtenu(s), si la structure initiale, lorqu'elle existe, peut être, d'une façon ou d'une autre, retrouvée.*

2.2.2 L'A.F.C. des données de l'exemple 1 avec le logiciel SAS

Nous présentons et commentons ci-après les principaux résultats de l'A.F.C. des données de l'exemple 1 mise en œuvre avec le logiciel SAS. On notera que ces résultats sont comparables à ceux fournis par la plupart des logiciels de statistique (en particulier SPSS, S-plus ou R).

Le tableau initial

Le premier résultat fourni par le logiciel est la table initiale, avec ses marges.

Contingency Table

	INF05	S0510	S1020	S2035	S3550	SUP50	Sum
ARIE	870	330	730	680	470	890	3970
AVER	820	1260	2460	3330	2170	2960	13000
H.G.	2290	1070	1420	1830	1260	2330	10200
GERS	1650	890	1350	2540	2090	3230	11750
LOT	1940	1130	1750	1660	770	1140	8390
H.P.	2110	1170	1640	1500	550	430	7400
TARN	1770	820	1260	2010	1680	2090	9630
T.G.	1740	920	1560	2210	990	1240	8660
Sum	13190	7590	12170	15760	9980	14310	73000

Les contributions au khi-deux

Le second résultat est la valeur de l'indice khi-deux (5375.49) qu'on obtient en faisant la somme, sur l'ensemble des cellules – des cases – de la table de contingence, des quantités

$$\frac{(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n})^2}{\frac{n_{\ell +} n_{+h}}{n}}$$

(voir le chapitre 3 du cours SDE).

En fait, le tableau des contributions au khi-deux fournit les quantités ci-dessus dans chaque cellule, ce qui permet de déceler facilement les cellules (autrement dit les croisements d'un département et d'une surface) contribuant le plus au khi-deux, donc à la définition de la liaison.

Contributions to the Total Chi-Square Statistic

	INF05	S0510	S1020	S2035	S3550	SUP50	Sum
ARIE	32.50	16.60	7.02	36.59	9.75	16.05	118.51
AVER	995.17	6.21	39.54	97.62	86.79	66.49	1291.82
H.G.	108.42	0.08	46.26	62.87	12.97	54.64	285.24
GERS	105.40	90.05	189.25	0.00	145.61	372.82	903.14
LOT	118.62	76.11	88.22	12.64	123.92	154.86	574.38
H.P.	446.82	208.58	133.83	5.96	210.68	718.07	1723.94
TARN	0.52	32.81	74.33	2.29	100.34	21.67	231.96
T.G.	19.63	0.43	9.36	61.97	31.77	123.35	246.51
Sum	1827.07	430.88	587.82	279.95	721.82	1527.95	5375.49

Considérons, par exemple, la cellule (1,1), soit ARIE x INF05 ; on obtient :

$$\frac{[870 - (3970 \times 13190)/73000]^2}{(3970 \times 13190)/73000} \simeq 32.50.$$

Cette valeur est relativement faible (par rapport aux autres valeurs du tableau), ce qui signifie que les très petites exploitations (moins de 5 hectares) n'ont rien de très particulier en Ariège.

Considérons maintenant la cellule (2,1), soit AVER x INF05 ; on obtient :

$$\frac{[820 - (13000 \times 13190)/73000]^2}{(13000 \times 13190)/73000} \simeq 995.17.$$

Cette valeur est la plus grande du tableau des contributions, ce qui signifie qu'en Aveyron, les très petites exploitations présentent une particularité très marquée : elles sont soit très nombreuses, soit très peu nombreuses (le carré intervenant dans l'expression du khi-deux supprime le signe et ne permet pas de dire quelle est celle des deux situations qui se présente). C'est le tableau des profils-lignes, ci-dessous, qui va permettre de lever cette ambiguïté : alors que ce type d'exploitations représente entre 14 % et 29 % de l'ensemble des exploitations dans les autres départements, elles ne sont que 6,3 % en Aveyron, autrement dit très peu nombreuses. Ce phénomène est un élément constitutif très important de la liaison existant entre les départements et les surfaces.

Les tableaux de profils

Il s'agit des deux tableaux donnant les profils-lignes pour le premier et les profils-colonnes pour le second. Le logiciel SAS ne les exprime pas en pourcentages, mais en fréquences, de sorte que les sommes (en lignes pour le premier et en colonnes pour le second) valent 1.

Row Profiles

	INF05	S0510	S1020	S2035	S3550	SUP50	
ARIE	0.219144	0.083123	0.183879	0.171285	0.118388	0.224181	--> 1
AVER	0.063077	0.096923	0.189231	0.256154	0.166923	0.227692	--> 1
H.G.	0.224510	0.104902	0.139216	0.179412	0.123529	0.228431	--> 1
GERS	0.140426	0.075745	0.114894	0.216170	0.177872	0.274894	--> 1
LOT	0.231228	0.134684	0.208582	0.197855	0.091776	0.135876	--> 1
H.P.	0.285135	0.158108	0.221622	0.202703	0.074324	0.058108	--> 1
TARN	0.183801	0.085151	0.130841	0.208723	0.174455	0.217030	--> 1
T.G.	0.200924	0.106236	0.180139	0.255196	0.114319	0.143187	--> 1

On a déjà signalé plus haut l'intérêt des profils dans l'analyse de la table de contingence. Il est clair que ce sont les variations de profils, d'une ligne à l'autre ou d'une colonne à l'autre, qui définissent la liaison entre les deux variables considérées. Elles doivent donc nécessairement être prises en compte dans l'analyse de cette liaison.

Column Profiles

	INF05	S0510	S1020	S2035	S3550	SUP50
ARIE	0.065959	0.043478	0.059984	0.043147	0.047094	0.062194
AVER	0.062168	0.166008	0.202136	0.211294	0.217435	0.206848
H.G.	0.173616	0.140975	0.116680	0.116117	0.126253	0.162823
GERS	0.125095	0.117260	0.110929	0.161168	0.209419	0.225716
LOT	0.147081	0.148880	0.143796	0.105330	0.077154	0.079665
H.P.	0.159970	0.154150	0.134758	0.095178	0.055110	0.030049
TARN	0.134193	0.108037	0.103533	0.127538	0.168337	0.146052
T.G.	0.131918	0.121212	0.128184	0.140228	0.099198	0.086653
TOTAL	1	1	1	1	1	1

La notion d'inertie en A.F.C.

Le tableau qui suit dans les sorties du logiciel SAS est relatif à la notion d'inertie. Avant de le détailler, nous allons essayer de préciser cette notion dans le contexte particulier de l'A.F.C.

Rappelons tout d'abord que la notion d'inertie, ou de dispersion, est fondamentale en statistique. Elle se ramène à la notion de variance dans le cas unidimensionnel (voir le chapitre 2 du cours SDE) et a déjà joué un rôle central en A.C.P. (voir le chapitre 1). C'est encore le cas en A.F.C. où son expression a une signification particulière (elle représente l'indicateur phi-deux, c'est-à-dire le khi-deux divisé par n, le nombre total d'observations).

Tout ceci est expliqué ci-dessous, le plus simplement possible... Malheureusement pour les lecteurs non mathématiciens, ces explications ne peuvent contourner une certaine technicité mathématique.

Que les lecteurs rebutés par ce qui suit ne s'inquiètent pas et retiennent essentiellement le dernier alinéa.

Un profil-ligne est un élément comportant c termes (c est le nombre de colonnes de la table analysée) dont la somme vaut 1. D'un point de vue mathématique, on peut donc représenter chaque profil-ligne par un *vecteur* dans un espace vectoriel de dimension c (en pratique, on considère $\mathbb{R}^c$ muni de la base canonique). Les coordonnées de ce vecteur sont les termes du profil-ligne correspondant. On obtient ainsi, dans l'espace considéré, un nuage de r vecteurs (r est le nombre de lignes de la table analysée) dont on peut déterminer le barycentre, c'est-à-dire le point moyen (chacune des coordonnées du barycentre est la moyenne pondérée des coordonnées correspondantes de l'ensemble des profils-lignes ; les pondérations sont les effectifs marginaux des lignes). Le barycentre

est le vecteur représentant le profil-ligne moyen, autrement dit, dans notre exemple, la répartition des exploitations agricoles selon les classes de S.A.U. dans l'ensemble de la région Midi-Pyrénées, tous départements confondus.

On peut faire le même raisonnement sur les profils-colonnes. L'espace considéré est alors de dimension r, on obtient dans cet espace un nuage de c points dont on peut déterminer le barycentre, représentant le profil-colonne moyen, autrement dit, dans notre exemple, la répartition des exploitations agricoles selon les départements de la région Midi-Pyrénées, indépendamment de la S.A.U.

Dans chacun des espaces vectoriels considérés ci-dessus, on peut déterminer l'inertie du nuage de points par rapport à son barycentre. C'est la somme pondérée des carrés des distances des profils à leur barycentre (formule analogue à celle définissant la variance). Les pondérations prises en compte sont encore les effectifs marginaux (des lignes ou des colonnes selon le cas). Quant aux distances, ce sont les distances définies dans chacun des deux espaces vectoriels considérés (qui sont donc, d'un point de vue mathématique, des espaces euclidiens). En fait, il ne s'agit pas de la distance usuelle, mais d'une distance spécifique à l'A.F.C., appelée distance, ou encore métrique, du khi-deux. Elle est construite à partir des inverses des fréquences des colonnes (dans $\mathbb{R}^c$) et de celles des lignes (dans $\mathbb{R}^r$).

On peut alors vérifier que l'inertie du nuage des profils-lignes, dans l'espace de dimension c, et celle du nuage des profils-colonnes, dans l'espace de dimension r, sont égales et ont pour expression la valeur de l'indicateur phi-deux calculé sur la table de contingence considérée.

Les pourcentages d'inertie des différentes dimensions

Comme en A.C.P., le tableau donnant la part d'inertie restituée par chaque dimension (chaque axe) permet de connaître la qualité globale des résultats (en particulier des graphiques) lorqu'on conserve seulement deux ou trois dimensions.

Sur l'exemple des exploitations agricoles, ce tableau est donné ci-dessous.

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	15 30 45 60 75
0.23455	0.05501	4015.91	74.71	**********
0.12210	0.01491	1088.29	20.25	*******
0.04894	0.00239	174.83	3.25	*
0.02792	0.00078	56.90	1.06	
0.02328	0.00054	39.55	0.74	
	0.07364	5375.49		

Les inerties totales des deux nuages (celui des profils-lignes et celui des profils-colonnes) sont identiques et se décomposent de la même manière selon les différents axes factoriels (ou axes principaux, ou axes principaux d'inertie) obtenus dans l'analyse.

Il n'y a donc qu'un seul tableau de résultats qui, dans la colonne "Principal Inertias" (inerties principales, c'est-à-dire selon les axes principaux), donne les valeurs de l'inertie restituée par chaque axe (c'est l'inertie du nuage, celui des profils-lignes ou celui des profils-colonnes, projeté sur cet axe). La somme de ces inerties est égale au phi-deux (ici 0.07364).

Comme en A.C.P., le premier axe est celui qui restitue la plus grande quantité d'inertie ; le second est celui qui, tout en étant orthogonal au premier (au sens de la métrique du khi-deux), en restitue aussi le maximum ; et ainsi de suite.

Les valeurs singulières ("Singular Values"), racines carrées positives des inerties principales, n'ont pas d'intérêt pratique et ne sont pas utilisées.

Les quantités figurant dans la colonne "Chi-Squares" (khi-deux) sont égales aux inerties principales multipliées par l'effectif de la table de contingence. C'est la raison pour laquelle leur somme est égale au khi-deux (on rappelle que $\chi^2 = n \times \Phi^2$). On peut encore considérer que chaque axe de l'analyse restitue une part du khi-deux, donc de la liaison entre les deux variables initiales, la plus importante pour l'axe 1 et ainsi de suite.

Les pourcentages ("Percents") représentent les pourcentages du khi-deux restitués par chaque axe. Comme en A.C.P., on se sert des pourcentages cumulés pour choisir la dimension à retenir. Dans notre exemple, les deux premières dimensions représentent quasiment 95 % de l'inertie totale. On ne retiendra donc que deux dimensions, ce qui permettra de ne réaliser qu'un seul graphique.

Remarque 2 *Lorqu'on réalise l'A.F.C. d'une table de contingence comportant r lignes et c colonnes, avec par exemple $r \geq c$, la dimension de l'espace dans lequel se trouve l'ensemble des résultats est $c-1$ (si l'on a $r \leq c$, cette dimension est $r-1$; de façon générale, elle vaut $inf(r-1, c-1)$). Ainsi, dans l'exemple considéré, on a $r = 8$ et $c = 6$, ce qui explique que le tableau ci-dessus fournisse seulement 5 dimensions. La diminution de un par rapport à la plus petite des deux dimensions provient du fait que la méthode opère sur des pourcentages dont le dernier peut toujours se déduire des précédents.*

Les coordonnées des lignes et des colonnes

Ce sont ces coordonnées qui permettent de réaliser le graphique représentant simultanément, selon les dimensions 1 et 2, les départements et les S.A.U. Leur détermination se fait selon le même principe qu'en A.C.P.

Nous donnons ci-dessous ces coordonnées. Le graphique correspondant est donné par la Figure 1.

Row Coordinates

	Dim1	Dim2
ARIE	0.037168	-.109849
AVER	-.236684	0.206059
H.G.	0.023759	-.157132
GERS	-.261525	-.089482
LOT	0.255187	0.032261
H.P.	0.478228	0.052226
TARN	-.102814	-.087061
T.G.	0.123568	0.068447

Column Coordinates

	Dim1	Dim2
INF05	0.322690	-.183979
S0510	0.215688	0.069874
S1020	0.147020	0.149383
S2035	-.047693	0.106435
S3550	-.257888	-.011834
SUP50	-.304488	-.103492

L'interprétation du graphique est donnée plus bas.

Les contributions à l'inertie selon chaque axe

On a vu que les inerties de chaque nuage (celui des profils-lignes et celui des profils-colonnes) se décomposaient, de la même façon, selon les différents axes. Ici, puisqu'on ne conserve que deux dimensions, on ne s'intéresse qu'aux inerties selon les deux premiers axes.

Pour chacun des deux axes retenus, les tableaux ci-dessous donnent les parts d'inertie dues d'abord à chaque ligne (ou département), ensuite à chaque colonne (ou classe de S.A.U.). Ces part sont exprimées en fréquences et somment donc à 1.

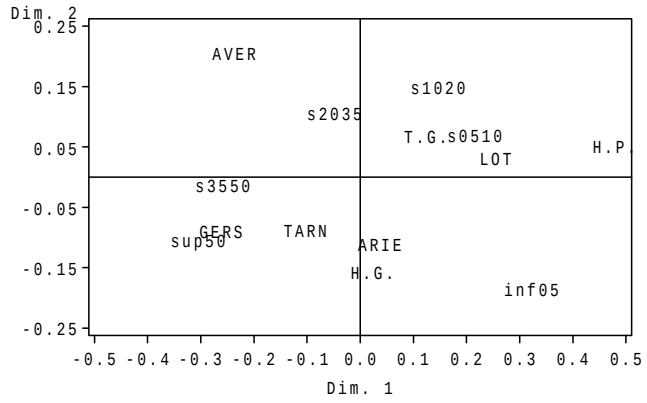

FIG. 2.1 – *Résultats de l'A.F.C. sur les exploitations agricoles de Midi-Pyrénées*

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
ARIE	0.001366	0.044019
AVER	0.181341	0.507201
H.G.	0.001434	0.231410
GERS	0.200115	0.086450
LOT	0.136049	0.008024
H.P.	0.421421	0.018546
TARN	0.025348	0.067070
T.G.	0.032927	0.037281
	1	1

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
INF05	0.342003	0.410237
S0510	0.087925	0.034051
S1020	0.065503	0.249544
S2035	0.008926	0.164051
S3550	0.165276	0.001284
SUP50	0.330367	0.140833
	1	1

Comment détermine-t-on ces contributions ? Si on désigne par c_ℓ^k la coordonnée du département numéro ℓ ($\ell = 1, \ldots, 8$) sur l'axe k ($k = 1, 2$), l'inertie selon l'axe k vaut :

$$\mathcal{I}_k = \sum_{\ell=1}^{r} \frac{n_{\ell+}}{n} (c_\ell^k)^2.$$

La part du département ℓ vaut donc :

$$\frac{\frac{n_{\ell+}}{n}(c_\ell^k)^2}{\mathcal{I}_k}.$$

Prenons l'exemple de l'Aveyron ($\ell = 2$) sur l'axe 1 ($k = 1$). Le tableau des inerties fournit : $\mathcal{I}_1 = 0.05501$. Celui des coordonnées fournit : $c_2^1 = -0.236684$. Enfin, la table de contingence initiale permet d'écrire : $\frac{n_{2+}}{n} = \frac{13}{73}$. On en déduit que la contribution de l'Aveyron à l'inertie du

nuage des départements selon l'axe 1 vaut :

$$\frac{\frac{13}{73} \times (0.236684)^2}{0.05501} \simeq 0.1813,$$

valeur donnée dans le tableau ci-dessus.

Les contributions aux inerties servent à la fois à sélectionner les lignes et les colonnes les plus importantes dans l'analyse (c'est-à-dire dans la définition de la liaison) et, le cas échéant, à interpréter les axes des graphiques.

Signalons néanmoins, qu'en A.F.C., l'interprétation concrète des axes n'est pas aussi fondamentale qu'en A.C.P. On ne fait cette interprétation que si elle est simple à faire et si elle facilite la compréhension des résultats. Pour la faire, on utilise bien sûr le graphique, mais aussi les contributions des lignes et celles des colonnes à l'inertie de leur nuage. Dans l'exemple considéré nous pouvons sans difficulté interpréter les axes (en particulier le premier).

On voit ainsi que les départements les plus importants dans la définition de l'axe 1 (ceux qui contribuent le plus à son inertie) sont les Hautes-Pyrénées, le Gers et l'Aveyron. Du point de vue des tailles de S.A.U., il s'agit des très petites exploitations (INF05), des très grandes (SUP50) et des assez grandes (S3550).

L'axe 2, concernant les départements, est surtout déterminé par l'Aveyron et la Haute-Garonne ; pour la S.A.U., il s'agit surtout des très petites exploitations et de celles de surface comprise entre 10 et 20 hectares, puis, dans une moindre mesure, des surfaces S2035 et SUP50. Nous verrons dans le point 2.3 comment ces éléments interviennent dans l'interprétation des résultats.

Les cosinus carrés

Ces quantités indiquent, comme en A.C.P., la qualité de la représentation sur chaque axe (autrement dit sur chaque dimension) de chaque modalité (ligne ou colonne).

Dans chacun des deux espaces de représentation des modalités (celui des lignes et celui des colonnes, chacun de dimension $\inf(r-1, c-1)$), les angles dont on considère le cosinus sont les angles entre chaque vecteur représentant une modalité et l'axe considéré. Plus cet angle est petit, plus son cosinus (donc son carré) est proche de 1, et plus la qualité de la représentation de la modalité sur cet axe est bonne. Plus cet angle est grand (proche d'un angle droit), plus son cosinus (donc son carré) est proche de 0, et plus la qualité de la représentation de la modalité sur cet axe est mauvaise.

On utilise les carrés des cosinus car on peut les additionner selon les différentes dimensions (propriété géométrique classique).

Squared Cosines for the Row Points

	Dim1	Dim2
ARIE	0.046279	0.404245
AVER	0.563739	0.427291
H.G.	0.020186	0.882916
GERS	0.889835	0.104173
LOT	0.951223	0.015203
H.P.	0.981701	0.011708
TARN	0.438847	0.314675
T.G.	0.536412	0.164587

Squared Cosines for the Column Points

	Dim1	Dim2
INF05	0.751725	0.244357
S0510	0.819488	0.086004
S1020	0.447511	0.462010
S2035	0.128051	0.637744
S3550	0.919524	0.001936
SUP50	0.868303	0.100310

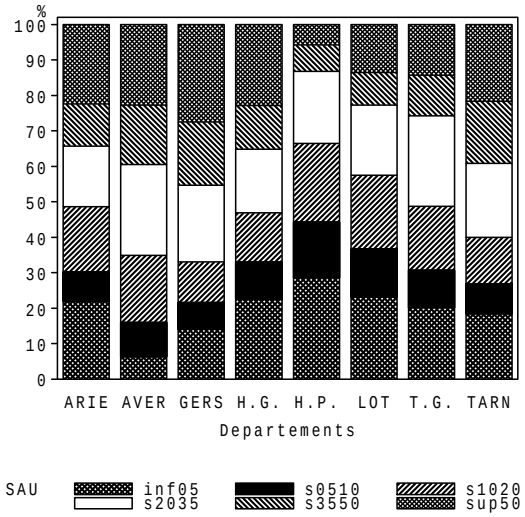

FIG. 2.2 – *Profils-lignes des départements*

Prenons deux exemples.

Le cosinus carré de l'angle entre le vecteur représentant l'Aveyron et le plan du graphique vaut : $0.5637 + 0.4273 = 0.9910$; l'angle correspondant est de 5.4 degrés, autrement dit, très petit. L'Aveyron est donc très bien représenté dans le plan. Ce n'est pas le cas de l'Ariège dont le cosinus carré avec le même plan vaut $0.0463 + 0.4042 = 0.4505$, ce qui correspond à un angle de 47.8 degrés (plus de la moitié d'un angle droit).

On pourra donc interpréter sans réserve la proximité, dans le plan, de l'Aveyron avec tout autre département ou toute autre surface bien représentée. Il faudra par contre être très prudent en ce qui concerne l'Ariège.

2.2.3 Interprétation des résultats

Précisons tout d'abord que cette interprétation se basera uniquement sur les résultats en dimension 2, puisque 95 % de l'information utile (celle exprimée par la dispersion, c'est-à-dire l'inertie) est contenue dans ces deux seules dimensions. On va d'ailleurs voir que les phénomènes les plus marquants sont ceux révélés par la dimension 1.

La figure 2.2 fournit le diagramme en barres des profils-lignes (les départements) qui permet de mieux comprendre les explications qui vont suivre (on notera que le diagramme en barres des profils-colonnes contient la même information statistique, mais que celui des profils-lignes nous paraît plus commode, dans cet exemple, pour aider l'interprétation).

Puisque les surfaces de S.A.U. sont naturellement ordonnées, commençons par étudier leurs positions dans le plan. La première chose remarquable est que leur ordre (rappelons le, non pris en compte dans l'analyse) est strictement respecté sur l'axe 1 qui est donc très structurant : il ordonne, de la droite vers la gauche, les surfaces, des plus petites aux plus grandes. Par conséquent, plus un département se trouve situé à droite, plus il comporte de petites exploitations et réciproquement.

Ainsi, les Hautes-Pyrénées se caractérisent par la présence de nombreuses petites exploitations et la relative rareté des grandes exploitations : près de 45 % des exploitations y ont moins de 10 hectares (le Lot, qui vient juste derrière, en a moins de 37 %) ; seulement un peu plus de 13 % y ont plus de 35 hectares (là encore le Lot, juste derrière, en a déjà près de 23 %). Ce profil traduit le fait qu'il s'agit du département le plus "montagnard" de la région, comme son nom l'indique d'ailleurs. À l'opposé, l'Aveyron et le Gers se caractérisent par la présence de grandes exploitations et la rareté des petites : les exploitations de plus de 35 hectares représentent près de 40 % en Aveyron et plus de 45 % dans le Gers ; celles de moins de 10 hectares représentent seulement 16 % en Aveyron et 21.6 % dans le Gers. Les raisons géographiques en sont différentes : région de plateaux, de causses, pour l'Aveyron et de plaines et de collines pour le Gers ; dans les deux cas, la géographie favorise la présence de grandes exploitations.

On notera que la qualité de représentation en dimension 2 des départements cités est excellente (plus de 0.99 pour l'Aveyron, le Gers et les Hautes-Pyrénées ; 0.97 pour le Lot) ; il en va de même pour les surfaces citées (0.99 pour INF05 ; 0.91 pour S0510 ; 0.92 pour S3550 ; 0.97 pour SUP50).

En ce qui concerne les contributions des départements à l'axe 1, les quatre départements cités sont les seuls à avoir des contributions supérieures à 10 %, et ce de façon très nette. Même chose pour les surfaces INF05, S3550 et SUP50 (S0510 est un peu en dessous de 10 %).

Pour ce qui est des contributions au khi-deux, on pourra vérifier que les phénomènes déjà signalés correspondent à la presque totalité des fortes contributions (supérieures à 100).

La question qui se pose ensuite est de savoir ce que l'on peut dire de plus. En particulier, que représente l'axe 2 ? Ce n'est pas vraiment très clair, et c'est un phénomène courant que l'essentiel ayant été dit sur l'axe 1, le reste ne soit pas simple à interpréter. Essayons néanmoins. Pour les départements, les seules contributions un peu importantes sont celles de la Haute-Garonne et de l'Aveyron, qui s'opposent nettement sur l'axe 2. Pour ce qui est des surfaces, les contributions importantes sont celles de INF05 et S1020 et, dans une moindre mesure, S2035 et SUP50. Le très petit nombre, en Aveyron, d'exploitations de surface inférieure à 5 hectares a déjà été signalé (très forte contribution au khi-deux). D'un autre côté, il faut également signaler, dans ce département, le grand nombre d'exploitations moyennes, de S.A.U. comprise entre 20 et 35 hectares. Ceci permet donc d'affiner le profil, assez particulier, de l'Aveyron : beaucoup de très grandes exploitations (SUP50) et de moyennes (S2035) ; une proportion proche de la moyenne de la région pour les surfaces S1020 et S3550 ; très peu de petites exploitations de moins de 10 hectares. Qu'en est-il pour la Haute-Garonne ? C'est le seul département (avec l'Ariège, mal représenté dans le plan du graphique) à avoir plus de 20 % d'exploitations de moins de 5 hectares et, en même temps, plus de 20 % d'exploitations de plus de 50 hectares. C'est aussi un département où il y a relativement peu d'exploitations moyennes. L'ensemble de ces particularités provient de sa situation géographique, étirée selon l'axe nord-sud, avec, au sud, une zone de montagne (le Comminges) et, au nord, une zone de plaines et de collines (la plaine de la Garonne et le Lauragais).

Pour conclure, précisons que nous avons fait ici, à dessein, une interprétation très détaillée de cette A.F.C. Il n'est pas toujours nécessaire d'entrer autant dans le détail. On retiendra essentiellement que l'interprétation s'appuie sur le (ou les) graphique(s), nécessite le recours à différents indicateurs (contributions aux axes, contributions au khi-deux, cosinus carrés) et qu'il ne faut jamais oublier qu'on analyse les profils (lignes et colonnes) et que c'est donc eux qu'il faut regarder avant d'avancer tout élément d'interprétation. Enfin signalons que, lorsque certains effectifs de la table de contingence initiale sont très faibles (ce qui n'est pas du tout le cas ici), il faut éviter de tirer des conclusions hâtives concernant les modalités correspondantes.

Chapitre 3

Analyse des Correspondances Multiple

Le chapitre 3 était consacré à l'Analyse Factorielle des Correspondances (A.F.C.), méthode factorielle de Statistique Descriptive Multidimensionnelle qui permet d'analyser la liaison entre deux variables qualitatives (éventuellement catégorielles). Dans la mesure où elle ne peut prendre en compte que deux variables, l'A.F.C. est naturellement limitée (elle est d'ailleurs parfois appelée Analyse des Correspondances Binaire, ou encore Analyse des Correspondances Simple).

Dans la pratique, en particulier dans le domaine du traitement d'enquêtes (ou de questionnaires), il est rare qu'on se limite à deux variables (deux questions). Le problème statistique que pose alors ce type de données est l'analyse de la liaison pouvant exister entre un nombre quelconque de variables qualitatives. L'Analyse des Correspondances Multiple (A.C.M.) est la méthode factorielle de Statistique Descriptive Multidimensionnelle qui permet de traiter ce problème.

Dans son principe, l'A.C.M. est une A.F.C. particulière. Ce qui change est le tableau des données sur lequel on applique la méthode. Le problème fondamental est en effet de savoir quel tableau statistique, croisant un nombre quelconque de variables qualitatives, peut généraliser la table de contingence. En fait, la réponse a déjà été donnée dans le cours SDE : c'est le tableau de Burt. Ainsi, l'A.C.M. est une A.F.C. réalisée sur un tableau de Burt relatif à au moins trois variables qualitatives.

La façon d'interpréter les résultats d'une A.C.M. sera donc analogue à la façon d'interpréter ceux d'une A.F.C. Malheureusement, certains indicateurs d'aide à l'interprétation utilisés en A.F.C ne sont plus valables dans le contexte de l'A.C.M. De plus, la présence d'un nombre plus important de variables rend l'interprétation plus délicate. Une bonne maîtrise de l'A.C.M. nécessite donc une grande pratique de cette méthode (plus que de vastes connaissances mathématiques).

Dans le cadre de ce cours, notre ambition se limitera à présenter rapidement la méthode et à en exposer le mécanisme d'interprétation sur un exemple réel relativement simple.

3.1 Rappels sur le tableau de Burt

Nous reprenons, dans ce paragraphe, des notions déjà introduites dans le paragraphe 3 du chapitre 3 du cours SDE.

3.1.1 Les données considérées

Les données avec lesquelles on est amené à construire un tableau de Burt sont précisemment celles considérées dans une Analyse des Correspondances Multiple (A.C.M.).

Soit donc un nombre quelconque (noté p, $p \geq 3$) de variables qualitatives, observées sur un ensemble de n individus (l'échantillon considéré), chacun affecté du même poids $\frac{1}{n}$. Les variables seront notées $X^1, \ldots, X^p$, le nombre de modalités de X^j sera noté c_j $(j = 1, \ldots, p)$, et on posera $c = \sum_{j=1}^{p} c_j$ (nombre total de modalités considérées, toutes variables confondues).

Remarque 3 *Comme en A.F.C., on peut utiliser en A.C.M. des variables catégorielles (variables qualitatives, à modalités ordonnées ou non, ou variables quantitatives, discrètes ou continues). On parle alors de catégories pour désigner soit les modalités, soit les valeurs, soit les classes, étant entendu que la structure de ces catégories (structure d'ordre ou structure numérique) n'est pas prise en compte par l'analyse. Cela rend très souple l'utilisation de l'A.C.M. car c'est une méthode susceptible de traiter n'importe quelle nature de variable.*

3.1.2 Définition du tableau de Burt

Nous redonnons ici la définition du tableau de Burt (sa compréhension est facilitée par l'exemple donné plus bas). Rappelons qu'un tableau de Burt est une généralisation particulière de la table de contingence pour un nombre quelconque p de variables qualitatives.

Le tableau de Burt est en fait une matrice carrée (un tableau carré) $c \times c$, constituée de p^2 sous-matrices. Chacune des p sous-matrices diagonales est relative à l'une des p variables ; la $j^{\text{ième}}$ d'entre elles est carrée d'ordre c_j, diagonale, et comporte sur la diagonale les effectifs marginaux de X^j. La sous-matrice figurant dans le bloc d'indice $(j, j'), j \neq j'$, est la table de contingence construite en mettant X^j en lignes et $X^{j'}$ en colonnes. Le tableau de Burt est donc symétrique.

3.1.3 Illustration

Reprenons le même exemple que dans le cours SDE : on a considéré un échantillon de 797 étudiants de l'Université Paul Sabatier (Toulouse III) ayant obtenu soit le DEUG A soit le DEUG B (diplômes scientifiques de premier cycle, en deux ans), et uniquement ce diplôme, durant la période 1971–1983. Trois variables ont été prises en compte : la série de bac, à 2 modalités (C, D) ; l'âge d'obtention du bac, à 4 modalités (moins de 18 ans, 18 ans, 19 ans, plus de 19 ans) ; la durée d'obtention du DEUG, à 3 modalités (2 ans, 3 ans, 4 ans).

Dans cet exemple, on a : $n = 797$; $p = 3$; $c_1 = 2, c_2 = 4, c_3 = 3$; $c = 9$. Le tableau de Burt correspondant est donné ci-dessous.

	bacC	bacD	< 18	18ans	19ans	> 19	2ans	3ans	4ans
bacC	583	0	108	323	114	38	324	192	67
bacD	0	214	25	97	68	24	76	82	56
< 18	108	25	133	0	0	0	84	35	14
18ans	323	97	0	420	0	0	224	137	59
19ans	114	68	0	0	182	0	73	75	34
> 19	38	24	0	0	0	62	19	27	16
2ans	324	76	84	224	73	19	400	0	0
3ans	192	82	35	137	75	27	0	274	0
4ans	67	56	14	59	34	16	0	0	123

3.2 Principes de l'A.C.M.

3.2.1 Le problème

Il s'agit d'étudier les liaisons pouvant exister entre les p variables considérées. En fait, dans la mesure où les données se présentent sous forme d'un tableau de Burt, juxtaposition de tables de contingence, seules les liaisons entre variables prises deux à deux sont considérées (il s'agit de ce que l'on appelle en statistique les *interactions d'ordre deux*). Pour étudier ces liaisons, la démarche sera de même nature qu'en A.F.C.

3.2.2 La méthode

L'A.C.M. consiste simplement à réaliser l'A.F.C. du tableau de Burt considéré. On peut en effet montrer d'une part que cela a un sens, d'autre part que, dans le cas où l'on fait l'A.F.C. du tableau de Burt relatif à deux variables qualitatives (cas où $p = 2$), on obtient sensiblement les mêmes résultats qu'en partant de la table de contingence relative à ces deux variables : l'A.C.M. est donc bien une généralisation de l'A.F.C.

L'interprétation d'une A.C.M. sera donc, dans ses grandes lignes, analogue à celle d'une A.F.C. Le problème est que certains indicateurs d'aide à l'interprétation utilisés en A.F.C ne sont plus valables en A.C.M., ce qui rend plus délicate son interprétation. De plus, la présence d'un nombre plus important de variables complique encore les choses. Une bonne maîtrise de l'A.C.M. nécessite donc une grande pratique de cette méthode.

3.3 Un exemple illustratif

Cet exemple concerne des étudiants inscrits pour la première fois à l'Université des Sciences Sociales de Toulouse (Toulouse I) à l'automne 1990, en première année de DEUG de droit, et suivis jusqu'en 1996.

3.3.1 Les données

Il y a 1635 étudiants pris en compte ($n = 1635$) et 5 variables qualitatives ($p = 5$). Les variables sont les suivantes :

- le sexe, à 2 modalités : fille, gars ;
- la série de bac, à 5 modalités : bacA, bacB, bacCouD, bacG, autbac ;
- l'âge d'obtention du bac, à 3 modalités : .18., .19., .20. ;
- la Catégorie Socio-Professionnelle (C.S.P.) des parents, à 6 modalités : art+com (artisans et commerçants), empl (employés), inter (professions intermédiaires), ouvr (ouvriers), prolib (professions libérales), autcsp (autres C.S.P.) ;
- la réussite, au moins au DEUG, à 2 modalités : OUI, NON.

Les données se présentent sous la forme d'un fichier à 1635 lignes et 5 colonnes dont on donne ci-dessous les trois premières et les trois dernières lignes.

```
1 4 3 2 2
1 4 3 2 2
2 1 3 1 1
   ...
1 3 3 2 2
1 5 3 5 2
1 2 2 2 2
```

Remarque 4 *Il faut noter ici une particularité qui est, dans la pratique, presque systématique avec ce type de données (nombreuses variables qualitatives) : les modalités de chacune des variables ont été codées 1,2... C'est, bien entendu, nettement plus commode pour l'enregistrement des données sur support informatique. Mais, cela nécessite un recodage pour faire apparaître explicitement les modalités initiales dans un tableau ou sur un graphique. En effet, si l'on arrive à comprendre, dans le fichier ci-dessus, que le "2" figurant ligne 3 et colonne 1 représente une fille, tandis que le "2" figurant ligne 1 et colonne 4 représente un fils d'employé, cela ne sera plus possible lorsqu'on rencontrera un "2" dans un graphique. Une phase de recodage des données est donc en général nécessaire avant de mettre en œuvre une A.C.M.*

3.3.2 L'A.C.M. des données

Comme dans les chapitres précédents, ces données ont été traitées avec le logiciel SAS.

Le tableau de Burt

Le premier résultat fourni est le tableau de Burt, toujours appelé "Contingency Table" dans SAS. Bien entendu, il est plus compliqué à lire qu'une table de contingence ordinaire croisant seulement deux variables.

Lorsqu'on interprète une liaison entre deux variables (parmi toutes celles considérées), il est en général conseillé de consulter le tableau de Burt pour y lire les effectifs correspondants (il faut toujours s'assurer qu'on ne raisonne pas sur un effectif trop faible). On notera que les effectifs marginaux (ce sont les mêmes en lignes et en colonnes puisqu'un tableau de Burt est symétrique)

ne s'interprètent pas facilement ici : chacun est égal à l'effectif de la modalité correspondante multiplié par le nombre p de variables considérées (ici 5). Enfin, l'effectif total est égal au nombre d'observations n (ici 1635) multiplié par p^2 (ici 25), soit 40875.

Contingency Table

	fille	gars	autbac	bacA	bacB	bacCouD	bacG
fille	1014	0	32	366	339	92	185
gars	0	621	19	126	258	94	124
autbac	32	19	51	0	0	0	0
bacA	366	126	0	492	0	0	0
bacB	339	258	0	0	597	0	0
bacCouD	92	94	0	0	0	186	0
bacG	185	124	0	0	0	0	309
.18.	508	221	6	255	314	117	37
.19.	321	210	9	167	190	54	111
.20.	185	190	36	70	93	15	161
art+com	106	61	2	56	62	15	32
autcsp	232	119	20	107	91	24	109
empl	99	54	4	47	69	6	27
inter	156	98	6	70	120	21	37
ouvr	143	74	10	57	78	9	63
prolib	278	215	9	155	177	111	41
NON	550	390	45	287	265	70	273
OUI	464	231	6	205	332	116	36
Sum	5070	3105	255	2460	2985	930	1545

	.18.	.19.	.20.	art+com	autcsp	empl	inter
fille	508	321	185	106	232	99	156
gars	221	210	190	61	119	54	98
autbac	6	9	36	2	20	4	6
bacA	255	167	70	56	107	47	70
bacB	314	190	93	62	91	69	120
bacCouD	117	54	15	15	24	6	21
bacG	37	111	161	32	109	27	37
.18.	729	0	0	63	125	61	132
.19.	0	531	0	65	115	63	74
.20.	0	0	375	39	111	29	48
art+com	63	65	39	167	0	0	0
autcsp	125	115	111	0	351	0	0
empl	61	63	29	0	0	153	0
inter	132	74	48	0	0	0	254
ouvr	90	62	65	0	0	0	0
prolib	258	152	83	0	0	0	0
NON	311	326	303	97	233	87	143
OUI	418	205	72	70	118	66	111
Sum	3645	2655	1875	835	1755	765	1270

	ouvr	prolib	NON	OUI	!	Sum
fille	143	278	550	464	!	5070
gars	74	215	390	231	!	3105
autbac	10	9	45	6	!	255
bacA	57	155	287	205	!	2460
bacB	78	177	265	332	!	2985
bacCouD	9	111	70	116	!	930
bacG	63	41	273	36	!	1545
.18.	90	258	311	418	!	3645

```
.19.          62      152      326      205   !   2655
.20.          65       83      303       72   !   1875
art+com        0        0       97       70   !    835
autcsp         0        0      233      118   !   1755
empl           0        0       87       66   !    765
inter          0        0      143      111   !   1270
ouvr         217        0      143       74   !   1085
prolib         0      493      237      256   !   2465
NON          143      237      940        0   !   4700
OUI           74      256        0      695   !   3475
-------------------------------------------------------
Sum         1085     2465     4700     3475   !  40875
```

Les pourcentages d'inertie des différentes dimensions

Le tableau suivant donne les valeurs propres, ou inerties selon les axes (Principal Inertias), la décomposition du khi-deux sur les axes et les pourcentages d'inertie restitués par chaque axe.

```
                Inertia and Chi-Square Decomposition

Singular  Principal Chi-
Values    Inertias  Squares Percents    3    6    9   12   15
                                      ----+----+----+----+----+---
0.61285   0.37558   3387.43  14.45% ************************
0.50322   0.25323   2283.88   9.74% ****************
0.48110   0.23145   2087.51   8.90% ***************
0.47320   0.22392   2019.58   8.61% **************
0.45086   0.20328   1833.36   7.82% *************
0.44737   0.20014   1805.07   7.70% *************
0.44171   0.19510   1759.67   7.50% *************
0.43237   0.18694   1686.07   7.19% ************
0.42231   0.17835   1608.55   6.86% ***********
0.40973   0.16788   1514.11   6.46% ***********
0.38679   0.14961   1349.33   5.75% **********
0.36548   0.13358   1204.76   5.14% *********
0.31771   0.10094    910.39   3.88% ******
          -------   -------
          2.60000  23449.71
```

Le problème est que ce tableau ne peut pas s'interpréter comme les tableaux analogues rencontrés en A.C.P. et en A.F.C. En effet, le tableau de Burt contient beaucoup d'informations redondantes (en particulier, il est symétrique et tous les effectifs sont répétés deux fois). Les pourcentages ci-dessus étant relatifs à la totalité de l'information contenue dans le tableau, il sont donc largement sous-estimés. Ainsi, les deux premiers axes de cette analyse ne représentent pas 24.19 % de la dispersion totale ($14.45 + 9.74$), mais davantage. Malheureusement, on ne peut pas savoir quel est le pourcentage réel. Ces pourcentages sont donc à prendre uniquement à titre indicatif.

Les coordonnées des modalités et leurs contributions à l'inertie

Seulement deux ensembles de résultats sont pris en compte ici : les coordonnées des colonnes sur les axes, permettant de réaliser le (ou les) graphique(s), selon le nombre d'axes retenus (deux ou plus) ; les contributions des colonnes à l'inertie (la dispersion) selon chaque axe, qui s'interprettent exactement comme en A.F.C. Les autres quantités utilisées en A.F.C. (les contributions au khi-deux, les profils et les cosinus carrés) n'ont plus d'interprétation directe en A.C.M. et ne sont en général pas utilisées.

Remarque 5 *Le tableau de Burt étant symétrique, ses lignes et ses colonnes sont identiques. Les éléments de l'A.C.M. relatifs aux lignes sont donc identiques à ceux relatifs aux colonnes et, par conséquent, ne sont pas fournis.*

Nous donnons ci-après les coordonnées de l'ensemble des modalités sur les deux premiers axes (par soucis de simplicité, nous n'utiliserons ici que les deux premiers axes), puis leurs contributions à l'inertie de chacun de ces axes.

```
                Column Coordinates

                    Dim1          Dim2

fille         -0.11125      -0.53743
gars           0.18165       0.87754
autbac         1.62701       0.56575
bacA          -0.21630      -0.81059
bacB          -0.40520       0.09334
bacCouD       -0.91295       1.55368
bacG           1.40826       0.08171
.18.          -0.68841      -0.11547
.19.           0.09059      -0.16661
.20.           1.21001       0.46039
art+com        0.05265      -0.36354
autcsp         0.65135      -0.25675
empl          -0.02064      -0.68415
inter         -0.22781      -0.14436
ouvr           0.51077      -0.29683
prolib        -0.58262       0.72329
NON            0.57376      -0.00691
OUI           -0.77603       0.00935

Partial Contributions to Inertia for the Column Points

                    Dim1          Dim2

fille         0.004087      0.141475
gars          0.006674      0.231007
autbac        0.043970      0.007885
bacA          0.007497      0.156160
bacB          0.031923      0.002513
bacCouD       0.050491      0.216889
bacG          0.199587      0.000997
.18.          0.112521      0.004695
.19.          0.001419      0.007120
.20.          0.178820      0.038396
art+com       0.000151      0.010661
autcsp        0.048500      0.011177
empl          0.000021      0.034593
inter         0.004293      0.002557
ouvr          0.018438      0.009236
prolib        0.054504      0.124588
NON           0.100786      0.000022
OUI           0.136315      0.000029
```

Le graphique

Le graphique de l'ensemble des modalités selon les deux premières dimensions est donné par la figure 1.

3.3.3 Interprétation

Nous interprèterons seulement les deux premières dimensions : c'est suffisant ici et, de plus, l'interpétation de toute autre dimension se fait selon le même principe. Le principe général est de repérer les modalités ayant des contributions importantes aux axes et de regarder ensuite leur positionnement sur le graphique.

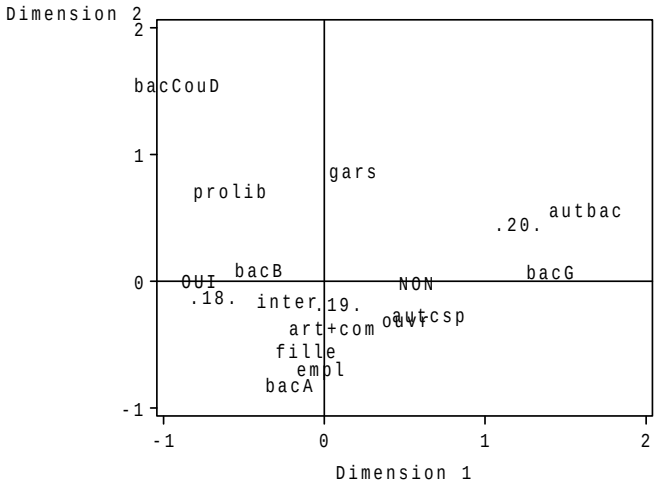

FIG. 3.1 – *Représentation graphique selon les deux premières dimensions*

Sur l'axe 1, ces contributions sont celles du bac G (pratiquement 20 %), des bacheliers de 20 ans ou plus (près de 18 %), de ceux de 18 ans ou moins (un peu plus de 11 %) et de la réussite ou de l'échec (13.6 % et 10 % respectivement). En observant le graphique, on voit que l'axe 1 discrimine la réussite, à gauche, et l'échec, à droite. On peut donc l'interpréter essentiellement comme l'axe d'opposition entre la réussite et l'échec au DEUG de Droit. Les modalités repérées ci-dessus (fortes contributions à l'axe 1) et proches de l'échec sont le bac G et l'obtention tardive du bac ; la modalité proche de la réussite est l'obtention du bac jeune. On voit donc que le facteur prépondérant de la réussite à ce DEUG est l'âge d'obtention du bac (autrement dit, la qualité de la scolarité secondaire). De plus, le bac G semble mal adapté aux études de droit.

Sur l'axe 2, les contributions les plus importantes sont celles des garçons (un peu plus de 23 %) et des filles (un peu plus de 14 %), des bacs C ou D (21.7 %), du bac A (15.6 %) et des professions libérales (environ 12.5 %). On remarque encore une nette discrimination, selon l'axe 2, entre les garçons, en haut, et les filles, en bas. Les garçons sont le plus souvent titulaires d'un bac C ou D et ont souvent des parents appartenant aux professions libérales, tandis que les filles sont plus souvent titulaires d'un bac A, sans que cela soit clairement lié à la réussite ou à l'échec. Il s'agit d'un phénomène bien marqué dans l'enseignement secondaire et que l'on retrouve ici comme sous-produit de notre analyse.

Remarque 6 *Pour terminer, on notera la particularité suivante : dans une A.C.M., toutes les variables prises en compte jouent, a priori, le même rôle : l'analyse ne peut en privilégier aucune. Néanmoins, dans la pratique, il est fréquent qu'une variable joue un rôle spécifique, en ce sens que c'est elle que l'on cherche à expliquer à partir des autres : c'est exactement le cas de la variable "réussite" dans l'exemple ci-dessus. Ce rôle spécifique n'apparaît, éventuellement, qu'au niveau de l'interprétation, autrement dit a posteriori. Lorsque c'est le cas, cela signifie, d'une certaine manière, que l'A.C.M. a bien fonctionné, autrement dit que les variables expliquant le phénomène (ici la variable "réussite") ont bien été prises en compte et ont été mises en évidence par l'analyse.*