UNIVERSITE
PAUL
SABATIER
TOULOUSE III

Publications du Laboratoire
de
Statistique et Probabilités

Statistique Descriptive Multidimensionnelle

Alain Baccini & Philippe Besse

Laboratoire de Statistique et Probabilités — URA CNRS 745
Université Paul Sabatier — 31062 – Toulouse cedex.

Chapitre I

Introduction

1 Généralités sur la statistique

Définition

Il n'est pas commode, dans cette introduction, de donner une définition précise du concept de statistique, alors que son contenu sera en partie élaboré dans la suite de ce cours. Nous nous contenterons donc, pour fixer les idées, d'en donner une définition volontairement assez vague.

Définition I.1 *On appelle Statistique l'ensemble des méthodes (ou encore des techniques) permettant d'analyser (on dira plutôt de* traiter*) des ensembles d'observations (nous parlerons de* données*).*

Les méthodes en question relèvent le plus souvent des mathématiques (raison pour laquelle, au moins en France, la Statistique fait partie des mathématiques appliquées) et font largement appel à l'outil informatique pour leur mise en œuvre.

On notera la distinction entre *la* Statistique, au sens défini ci-dessus, et *une* statistique, terme parfois utilisé pour désigner des "données statistiques" (voir ce terme plus loin) ; par exemple, on parle de la statistique du commerce extérieur de la France. Dans la suite de ce cours, nous n'utiliserons pas le terme de statistique dans ce dernier sens.

Bref historique

De façon un peu grossière, on peut distinguer trois phases essentielles dans l'évolution de la statistique.

- Depuis l'antiquité et jusqu'à la fin du $19^{\text{ième}}$ siècle, la statistique est restée principalement un ensemble de techniques de dénombrement.

- Entre la fin du $19^{\text{ième}}$ siècle et les années 1960, s'est construit, notamment à la suite de l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman ...), la statistique mathématique (ou statistique inférentielle, voir ci-dessous).

- Depuis les années 1960, avec le développement des outils informatiques et graphiques, la statistique, et surtout la statistique descriptive multidimensionnelle, a connu une expansion considérable.

Statistique descriptive et statistique inférentielle

De manière approximative, il est possible de classer les méthodes statistiques en deux groupes : celui des méthodes descriptives et celui des méthodes inférentielles.

- La statistique **descriptive**. On regroupe sous ce terme les méthodes dont l'objectif principal est la *description* des données étudiées ; cette description des données se fait à travers leur *présentation* (la plus synthétique possible), leur *représentation graphique*, et le calcul de *résumés numériques*. Dans cette optique, il n'est pas fait appel à des modèles probabilistes. On notera que les termes de statistique descriptive, *statistique exploratoire* et *analyse des données* sont quasiment synonymes. C'est essentiellement à ces méthodes qu'est consacré ce cours.

- La statistique **inférentielle**. Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population ; d'une certaine manière, il s'agit donc d'induire (ou encore d'inférer) du particulier au général. Le plus souvent, ce passage ne pourra se faire que moyennant des hypothèses de type probabiliste. Les termes de statistique inférentielle, *statistique mathématique*, et *statistique inductive* sont eux aussi quasiment synonymes.

D'un point de vue méthodologique, on notera que la statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données : les deux aspects de la statistique se complètent bien plus qu'ils ne s'opposent.

2 Terminologie de base

Population Ω (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de *champ de l'étude*.

Individu $\omega \in \Omega$ (ou *unité statistique*) : tout élément de la population.

Échantillon : sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon n : cardinal du sous-ensemble correspondant.

Enquête (statistique) : opération consistant à observer (ou mesurer, ou questionner ...) l'ensemble des individus d'un échantillon.

Recensement : enquête dans laquelle l'échantillon observé est la population tout entière (enquête *exhaustive*).

Sondage : enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population (enquête *non exhaustive*).

Variable (statistique) : $\Omega \xrightarrow{X} \begin{cases} \mathcal{E} \text{ si qualitative} \\ \mathbb{R} \text{ si quantitative} \end{cases}$
caractéristique (âge, salaire, sexe ...), définie sur la population et observée sur l'échantillon ; mathématiquement, il s'agit d'une application définie sur l'échantillon. Si la variable est à valeurs dans $\mathbb{R}$ (ou une partie de $\mathbb{R}$, ou un ensemble de parties de $\mathbb{R}$), elle est dite *quantitative* (âge, salaire, taille ...) ; sinon elle est dite *qualitative* (sexe, catégorie socioprofessionnelle ...).

Données (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Elles sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Lorsqu'un tableau ne comporte que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de *matrice*.

3 Statistique Descriptive Multidimensionnelle

En France, l'expression *"Analyse des Données"* recouvre les techniques ayant pour objectif la *description statistique des grands tableaux* (n lignes, où n varie de quelques dizaines à quelques milliers, p colonnes, où p varie de quelques unités à quelques dizaines). Ces méthodes se caractérisent par une utilisation *intensive* de l'ordinateur et une absence quasi systématique d'hypothèses de nature *probabiliste* sur les phénomènes observés. Depuis la fin des années 1970, de nombreux travaux ont néanmoins permis de sortir du cadre strictement *descriptif*, en introduisant dans des espaces multidimensionnels appropriés les outils probabilistes et la notion de *modèle*, usuelle en statistique *inférentielle*. Les techniques se sont ainsi enrichies de notions telles que l'estimation, la convergence, la stabilité des

résultats, le choix de critères... C'est dans ce cadre élargi, que nous désignerons par l'expression de *Statistique Descriptive Multidimensionnelle*, que se situe ce cours.

On notera que la culture anglo-saxonne ne possède pas l'équivalent de l'expression "Analyse des données". La traduction littérale *"Data Analysis"* a un sens très général irréductible à un ensemble de quelques techniques comme c'est le cas en France. Les méthodes relevant du domaine *"Multivariate Analysis"* poursuivent des objectifs sensiblement différents, un individu n'y étant souvent considéré que pour l'information qu'il apporte sur la connaissance des liaisons entre variables au sein d'un échantillon statistique dont la distribution est le plus souvent soumise à des hypothèses de normalité. Enfin, les techniques décrites par Tukey (1977) sous le titre *"Exploratory Data Analysis"* sont simplement unidimensionnelles ou bidimensionnelles.

La Statistique Descriptive Multidimensionnelle dispose essentiellement des méthodes suivantes :

- Description et réduction de dimension (méthodes factorielles) :
  i. Analyse en Composantes Principales (p variables quantitatives),
  ii. Analyse Factorielle Discriminante (p variables quantitatives, 1 variable qualitative),
  iii. Analyse Factorielle des Correspondances Binaire (2 variables qualitatives) et Multiple (p variables qualitatives),
  iv. Analyse Canonique (p et q variables quantitatives),
  v. "Multidimensional Scaling" (M.D.S.), ou positionnement multidimensionnel, ou analyse factorielle d'un tableau de distances.
- Méthodes de classification :
  i. Classifications Hiérarchiques,
  ii. Classifications non Hiérarchiques,
  iii. Segmentation.
- Autre méthode :
  i. Analyse en Facteurs ("Factor Analysis"), ou analyse en facteurs communs et spécifiques.

Les bases théoriques de ces méthodes sont anciennes : Spearman (1904) et Thurstone (1931, 1947) pour l'Analyse en Facteurs, Hotteling (1935) pour l'Analyse en Composantes Principales et l'Analyse Canonique, Hirschfeld (1935) et Guttman (1941, 1959) pour l'Analyse des Correspondances. Pratiquement, leur emploi ne s'est généralisé qu'avec la diffusion des moyens de calcul.

Les références les plus utiles pour ce cours sont :

- J.M. Bouroche & G. Saporta (1980),
- J.D. Dobson (1991),
- J.J. Droesbeke, B. Fichet & P. Tassi (éditeurs) (1992),
- B.S. Everitt & G. Dunn (1991),
- K.V. Mardia, J.T. Kent & J.M. Bibby (1979),
- G. Saporta (1990).

4 Contenu

Ce document est organisé en trois parties ; la première propose des rappels élémentaires de statistique descriptive unidimensionnelle et bidimensionnelle et ne nécessite aucun outil mathématique spécifique. La deuxième fait largement appel au calcul matriciel dont les éléments fondamentaux sont rappelés en annexe A ; elle expose les fondements des techniques factorielles mises en œuvre dans la plupart des logiciels statistiques.

Les points suivants nécessiteraient également une présentation :

- choix des variables,
- A.C.P. *robuste* en norme L_1 (distance city-bloc),
- rotations varimax, procruste,
- distributions asymptotiques et tests,
- méthodes semi-linéaires,
- ...

Ce cours adopte une optique particulière dans la présentation des techniques multidimensionnelles envisagées : chaque méthode est introduite comme le résultat de l'estimation d'un modèle. Il se démarque donc des approches strictement géométriques ou mécanistes de l'"Analyse des Données" afin de les resituer dans un cadre statistique cohérent sans pour autant adopter un point de vue anglo-saxon : il n'est pas supposé d'hypothèses de normalité sur les distributions, les individus sont étudiés, représentés, décrits pour eux-mêmes, simultanément à l'étude des variables statistiques.

Les graphiques présentés ont été réalisés à l'aide des logiciels SAS ou S+.

Chapitre II

Cas unidimensionnel

Si X est une variable statistique et si ω_i désigne l'individu générique de l'échantillon observé, nous noterons $X(\omega_i)$ la valeur prise par cette variable sur cet individu. C'est dorénavant l'échantillon observé, qu'il soit identique à la population complète ou non, qui sera noté Ω ; nous le supposerons de cardinal n. L'ensemble $\{X(\omega_i)\ ;\ i = 1, \ldots, n\}$ constitue ce que l'on appelle la *série statistique brute* ; c'est en général sous cette forme que se présentent les données dans un fichier informatique. Elles sont alors parfaitement illisibles dès que n est grand ; l'objectif de ce chapitre est d'exposer les outils élémentaires, adaptés au type de variable observée, permettant de présenter une série brute de façon synthétique et d'en résumer les principales caractéristiques.

1 Variable quantitative discrète

1.1 Introduction

En général, on appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs entières (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible (sauf exception, moins d'une vingtaine). Citons, par exemple, le nombre d'enfants dans une population de familles, le nombre d'années d'études après le bac dans une population d'étudiants ...

Exemple II.1 *On a noté l'âge (arrondi à l'année près) des 48 salariés d'une entreprise ; la série statistique brute est donnée ci-dessous (il s'agit de données fictives).*

43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31
62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59
57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

1.2 Présentation des données

Le tableau statistique

C'est un tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X ; ces observations sont rangées par ordre croissant et non répétées ; nous les noterons $\{x_l\ ;\ l = 1, \ldots, r\}$. Dans une seconde colonne, on dispose, en face de chaque valeur x_l, le nombre de réplications qui lui sont associées ; ces réplications sont appelées *effectifs* et notées n_l. Les effectifs n_l sont souvent remplacés par les quantités $f_l = \frac{n_l}{n}$, appelées *fréquences* (rappelons que n désigne le nombre total d'observations, c'est-à-dire le cardinal de Ω : $n = \sum_{l=1}^{r} n_l$).

Les effectifs cumulés et les fréquences cumulées

Il peut être utile de compléter le tableau statistique en y rajoutant soit les effectifs cumulés, soit les fréquences cumulées. Ces quantités sont respectivement définies de la façon suivante :

$$N_l = \sum_{j=1}^{l} n_j \text{ et } F_l = \sum_{j=1}^{l} f_j.$$

On notera que $N_r = n$ et $F_r = 1$.

Illustration

Dans le tableau statistique (II.1), on a calculé, sur les données présentées dans l'exemple II.1, les effectifs, effectifs cumulés, fréquences et fréquences cumulées.

Remarques II.1

- *Comme c'est le cas ci-dessus, les fréquences sont souvent exprimées en pourcentages.*

- *Le choix entre effectifs (resp. effectifs cumulés) et fréquences (resp. fréquences cumulées) est très empirique ; il semble naturel de choisir les effectifs lorsque l'effectif total n est faible et les fréquences lorsqu'il est plus important ; la limite approximative de* 100 *paraît, dans ces conditions, assez raisonnable.*

La présentation tige-et-feuille (ou "stem-and-leaf")

Cette façon particulière de présenter les données est assez commode, dans la mesure où elle préfigure déjà un graphique. Elle est illustrée ci-dessous sur le

x_l	n_l	N_l	$f_l(\%)$	$F_l(\%)$
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,91
52	3	38	6,25	79,16
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,00

TAB. II.1 - Effectifs, effectifs cumulés, fréquences et fréquences cumulées.

même exemple que précédemment.

2	4 6 6 9 9 9
3	1 1 3 3 3 3 7 7 8 8 8 8
4	1 1 1 3 3 3 5 6 6 6 6 6 6 9 9 9
5	0 2 2 2 7 7 7 7 7 9 9
6	0 0 2

1.3 Représentations graphiques usuelles

Pour une variable discrète, on rencontre essentiellement deux sortes de représentations graphiques, qui sont en fait complémentaires : le diagramme en bâtons et le diagramme cumulatif (en escaliers).

Le diagramme en bâtons

Il permet de donner une vision d'ensemble des observations réalisées. La figure II.1 donne le diagramme en bâtons des données de l'exemple II.1.

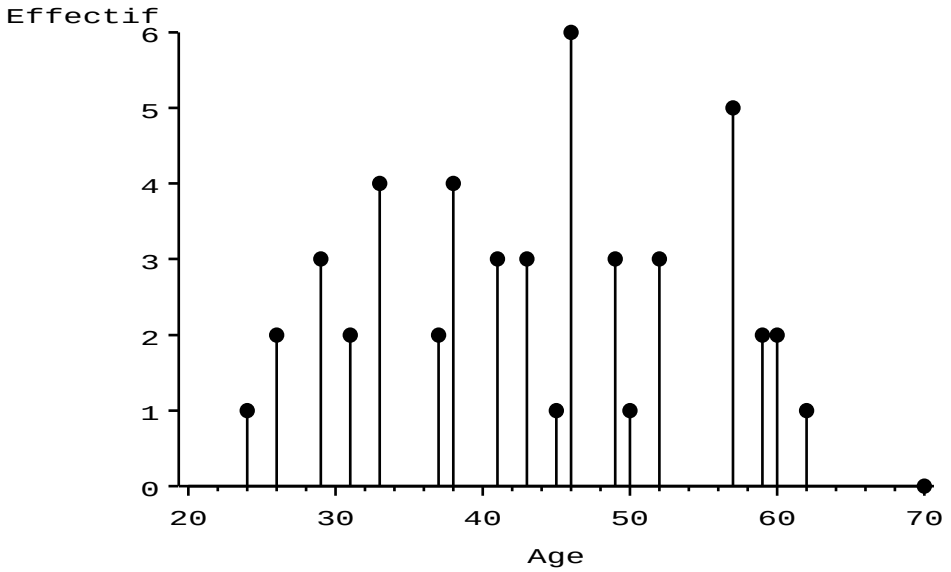


FIG. II.1 - Diagramme en bâtons

Le diagramme cumulatif

Il figure les effectifs cumulés (resp. les fréquences cumulées) et permet de déterminer simplement le nombre (resp. la proportion) d'observations inférieures ou égales à une valeur donnée de la série. Lorsqu'il est relatif aux fréquences, c'est en fait le graphe de la *fonction de répartition empirique* F_X définie de la façon suivante :

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1, \\ F_l & \text{si } x_l \leq x < x_{l+1}, \quad l = 1, \ldots, r-1, \\ 1 & \text{si } x \geq x_r. \end{cases}$$

Le diagramme cumulatif relatif à l'exemple II.1 est donné par la figure II.2.

1.4 Notion de quantile et applications

Définition

La fréquence cumulée F_l ($0 \leq F_l \leq 1$) donne la proportion d'observations inférieures ou égales à x_l. Une approche complémentaire consiste à se donner a priori une valeur α, comprise entre 0 et 1, et à rechercher x_α vérifiant $F_X(x_\alpha) \simeq \alpha$. La valeur x_α (qui n'est pas nécessairement unique) est appelée quantile (ou *fractile*) d'ordre α de la série. Les quantiles les plus utilisés sont associés à certaines valeurs particulières de α.

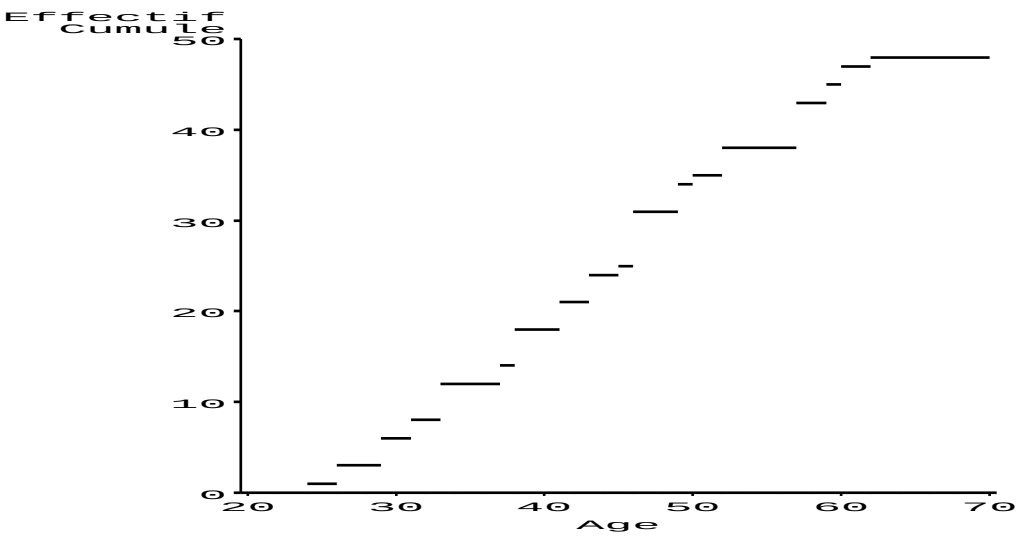


FIG. II.2 - Diagramme cumulatif

La médiane et les quartiles

La médiane est le quantile d'ordre $\frac{1}{2}$; elle partage donc la série des observations en deux ensembles d'effectifs égaux. Le premier quartile est le quantile d'ordre $\frac{1}{4}$, le troisième quartile celui d'ordre $\frac{3}{4}$ (le second quartile est donc confondu avec la médiane).

Les autres quantiles

Les *quintiles*, *déciles* et *centiles* sont également d'usage assez courant.

La boîte-à-moustaches (ou "box-and-whisker plot")

Il s'agit d'un graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane. La figure II.3 donne la boîte-à-moustaches de l'exemple II.1. Dans cet exemple, on a obtenu $x_{\frac{1}{4}} = 35$, $x_{\frac{1}{2}} = 44$ et $x_{\frac{3}{4}} = 52$; on notera que l'obtention, d'une part de $x_{\frac{1}{4}}$ et $x_{\frac{1}{2}}$, d'autre part de $x_{\frac{3}{4}}$, ne s'est pas faite de la même façon.

1.5 Caractéristiques numériques

Les caractéristiques (ou résumés) numériques introduites ici servent à synthétiser la série étudiée au moyen d'un petit nombre de valeurs numériques. On distingue essentiellement les caractéristiques de tendance centrale (ou encore de *position* ou de *localisation*) et les caractéristiques de dispersion.

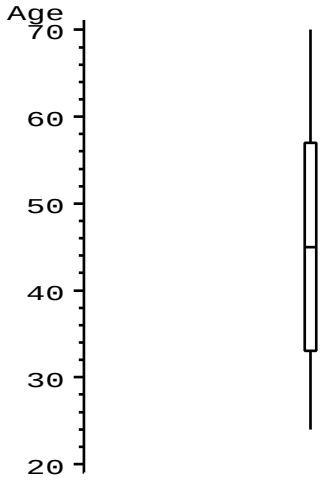


FIG. II.3 - Boîte-à-moustaches

Caractéristiques de tendance centrale

Leur objectif est de fournir un ordre de grandeur de la série étudiée, c'est-à-dire d'en situer le centre, le milieu. Les deux caractéristiques les plus usuelles sont :

- la *médiane*,
- la *moyenne* (ou moyenne arithmétique).

Formule de la moyenne pour une variable quantitative discrète :

$$\overline{x} \;=\; \frac{1}{n}\sum_{l=1}^{r} n_l x_l \;=\; \sum_{l=1}^{r} f_l x_l.$$

Caractéristiques de dispersion

Elles servent à préciser la variabilité de la série, c'est-à-dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

- L'*étendue* $(x_r - x_1)$,
- l'*intervalle interquartiles* $(x_{\frac{3}{4}} - x_{\frac{1}{4}})$,
- l'*écart-moyen à la médiane* $(\frac{1}{n}\sum_{l=1}^{r} n_l \left|x_l - x_{\frac{1}{2}}\right|)$,
- l'*écart-moyen à la moyenne* $(\frac{1}{n}\sum_{l=1}^{r} n_l \left|x_l - \overline{x}\right|)$,

sont des caractéristiques de dispersion que l'on rencontre parfois.

Mais, la caractéristique de loin la plus utilisée est l'*écart-type*, racine carrée positive de la *variance*. Formules de la variance :

$$\begin{aligned} var(X) = \sigma_X^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l(x_l - \overline{x})^2 \\ &= \frac{1}{n}\sum_{l=1}^{r} n_l(x_l)^2 - (\overline{x})^2. \end{aligned}$$

L'écart-type de X sera donc noté σ_X.

Illustration

En utilisant toujours l'exemple II.1, on a calculé :

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{l=1}^{r} n_l x_l = \frac{2094}{48} = 43,625 \simeq 43,6 \text{ ans} ; \\ \sigma_X^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l(x_l)^2 - (\overline{x})^2 = \frac{96620}{48} - (43,625)^2 \simeq 109,7760 ; \\ \sigma_X &= \sqrt{\sigma_X^2} \simeq 10,5 \text{ ans.} \end{aligned}$$

2 Variable quantitative continue

2.1 Généralités

Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels. Cela signifie que, dans ce cas, le sous-ensemble de $\mathbb{R}$ des valeurs possibles de la variable étudiée a été divisé en r intervalles contigus appelés *classes*.

En général, les deux raisons principales qui peuvent amener à considérer comme continue une variable quantitative sont le grand nombre d'observations distinctes (un traitement en discret serait dans ce cas peu commode) et le caractère "sensible" d'une variable (il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis). Deux exemples de variables quantitatives fréquemment considérées comme continues sont l'âge et le revenu (pour un groupe d'individus).

Nous noterons $(b_0\,;b_1)$,...,$(b_{r-1}\,;b_r)$ les classes considérées. Les nombres b_{l-1} et b_l sont appelés les *bornes* de la $l^{\text{ième}}$ classe ; $\frac{b_{l-1}+b_l}{2}$ est le *centre* de cette classe et $(b_l - b_{l-1})$ en est l'*amplitude* (en général notée a_l).

2.2 Présentation des données

On utilise encore un tableau statistique analogue à celui vu au paragraphe précédent, en disposant dans la première colonne les classes rangées par ordre croissant. Les notions d'effectifs, de fréquences, d'effectifs cumulés et de fréquences cumulées sont définies de la même façon que dans le cas discret. On notera que l'on n'utilise pas dans ce cas la présentation tige-et-feuille car les valeurs exactes de la série sont inconnues.

Exemple II.2 *Le tableau ci-dessous donne, pour l'année 1987, la répartition des exploitations agricoles françaises selon la SAU (surface agricole utilisée) exprimée en hectares (Tableaux Economiques de Midi-Pyrénées, INSEE, 1989, p. 77) ; la SAU est ici une variable quantitative continue comportant 6 classes.*

SAU (en ha)	fréquences (%)
moins de 5	24,0
de 5 à 10	10,9
de 10 à 20	17,8
de 20 à 35	20,3
de 35 à 50	10,2
plus de 50	16,8

2.3 Représentations graphiques

Les deux graphiques usuels remplaçant respectivement dans ce cas le diagramme en bâtons et le diagramme cumulatif sont l'histogramme et la courbe cumulative.

La courbe cumulative

C'est encore une fois le graphe de la *fonction de répartition empirique*, cette dernière devant maintenant être précisée au moyen d'*interpolations linéaires.*

On appelle fonction de répartition empirique de la variable continue X la fonction F_X définie par :

$$F_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ F_{l-1} + \frac{f_l}{b_l - b_{l-1}}(x - b_{l-1}) & \text{si } b_{l-1} \leq x < b_l, \quad l = 1, \ldots, r, \\ 1 & \text{si } x \geq b_r \end{cases}$$

(on a supposé $F_0 = 0$).

La courbe cumulative relative à l'exemple II.2 est donnée par la figure II.4. On notera que dans cet exemple, comme c'est souvent le cas avec une variable quantitative continue, il a fallu fixer arbitrairement la borne inférieure de la première classe (il était naturel ici de prendre $b_0 = 0$) ainsi que la borne supérieure de la dernière classe (on a choisi $b_6 = 200$, mais d'autres choix étaient possibles).

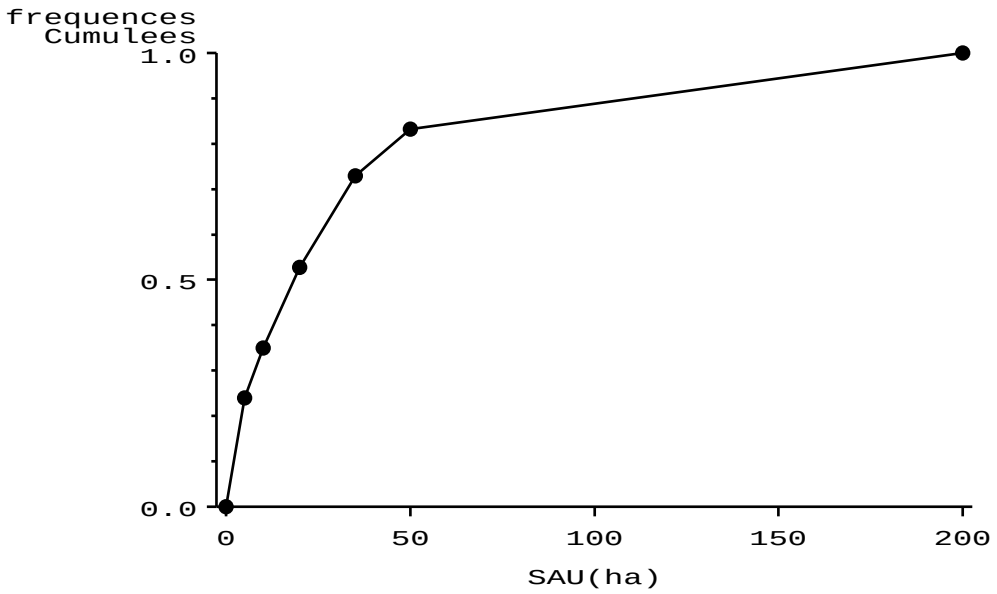


FIG. II.4 - Courbe cumulative

L'histogramme

La fonction de répartition empirique est, dans le cas continu, une fonction dérivable sauf, éventuellement, aux points d'abscisses b_0, b_1, ..., b_r. Sa fonction dérivée, éventuellement non définie en ces points, est appelée *densité empirique* de X et notée f_X. On obtient :

$$f_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ \frac{f_l}{b_l - b_{l-1}} & \text{si } b_{l-1} < x < b_l, \quad l = 1, \ldots, r, \\ 0 & \text{si } x \geq b_r. \end{cases}$$

Le graphe de f_X est alors appelé histogramme de la variable X. Un histogramme est donc la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_l = b_l - b_{l-1}$) et dont les hauteurs sont les quantités $\frac{f_l}{b_l - b_{l-1}}$, appelées *densités de fréquence*. L'aire du $l^{\text{ième}}$ rectangle vaut donc f_l, fréquence de la classe correspondante.

L'histogramme correspondant aux données de l'exemple II.2 est présenté dans la figure II.5.

2.4 Détermination des quantiles

Les quantiles x_α d'une variable continue peuvent être déterminés de façon directe à partir de la courbe cumulative. Cela signifie que, par le calcul, on doit commencer par déterminer la classe dans laquelle se trouve le quantile cherché, puis le déterminer dans cette classe par interpolation linéaire (voir l'illustration plus loin).

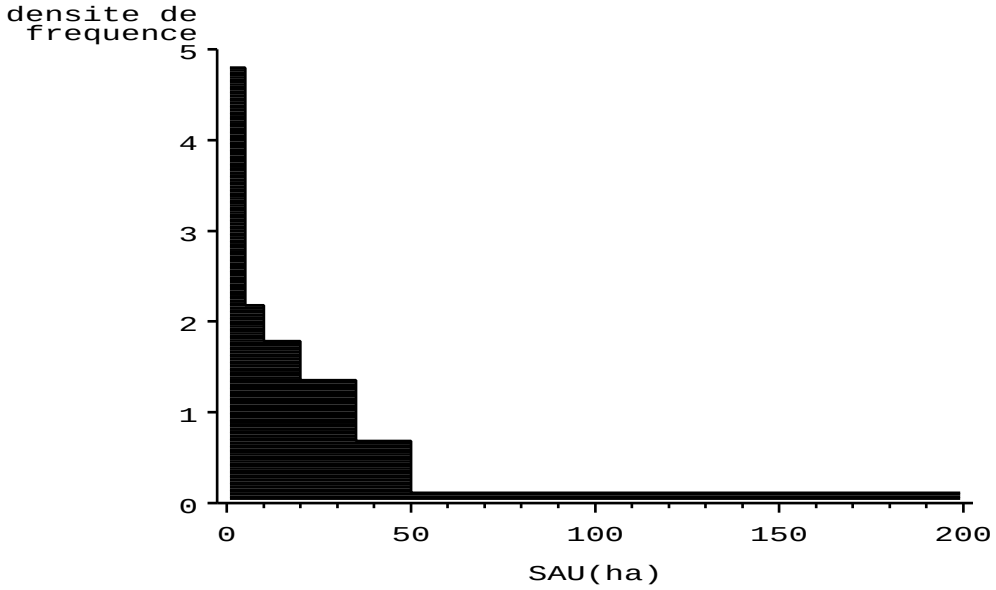


FIG. II.5 - Histogramme

2.5 Détermination des autres caractéristiques numériques

La moyenne, la variance et l'écart-type d'une variable continue se déterminent de la même manière que dans le cas discret ; dans les formules, on doit prendre pour x_l les centres de classes au lieu des observations (qui ne sont pas connues). Les valeurs obtenues pour ces caractéristiques sont donc assez approximatives ; cela n'est pas gênant dans la mesure où le choix de traiter une variable quantitative comme continue correspond à l'acceptation d'une certaine imprécision dans le traitement statistique.

2.6 Illustration

La médiane de la variable présentée dans l'exemple II.2 se situe dans la classe (10 ; 20), puisque la fréquence cumulée de cette classe (52,7) est la première à dépasser 50. On détermine la médiane en faisant l'interpolation linéaire suivante (l'indice l ci-dessous désigne en fait la troisième classe) :

$$\begin{aligned} x_{\frac{1}{2}} &= b_{l-1} + a_l \frac{50 - F_{l-1}}{F_l - F_{l-1}} \\ &= 10 + 10 \frac{15,1}{17,8} \\ &\simeq 18,5 \text{ ha.} \end{aligned}$$

La moyenne vaut :

$$\overline{x} = \sum_{l=1}^{r} f_l x_l = \frac{3080,5}{100} \simeq 30,8 \text{ ha.}$$

Remarques II.2 *Dans cet exemple, il convient de noter trois choses :*

- *tout d'abord, pour le calcul de la moyenne, nous avons choisi* $x_6 = 100$*, plutôt que* 125*, car cette valeur nous a semblé plus proche de la réalité ;*
- *ensuite, il se trouve que, dans ce cas, on peut calculer la* vraie *valeur de la moyenne, connaissant la SAU totale en France (31 285 400 ha) et le nombre total d'exploitations agricoles (981 720) ; on obtient 31,9 ha, ce qui signifie que l'approximation obtenue ici est très correcte ;*
- *enfin, le fait que la médiane soit sensiblement plus faible que la moyenne caractérise les séries fortement concentrées sur les petites valeurs.*

3 Variable qualitative

3.1 Variables nominales et variables ordinales

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées *modalités*. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite *ordinale*. Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite *nominale*.

3.2 Traitements statistiques

Il est clair qu'on ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinale). Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques. Encore faut-il noter que les notions d'effectifs cumulés et de fréquences cumulées n'ont de sens que pour des variables ordinales (elles ne sont pas définies pour les variables nominales).

3.3 Représentations graphiques

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont :

- le *diagramme en barre,*

– le *diagramme en colonnes,*

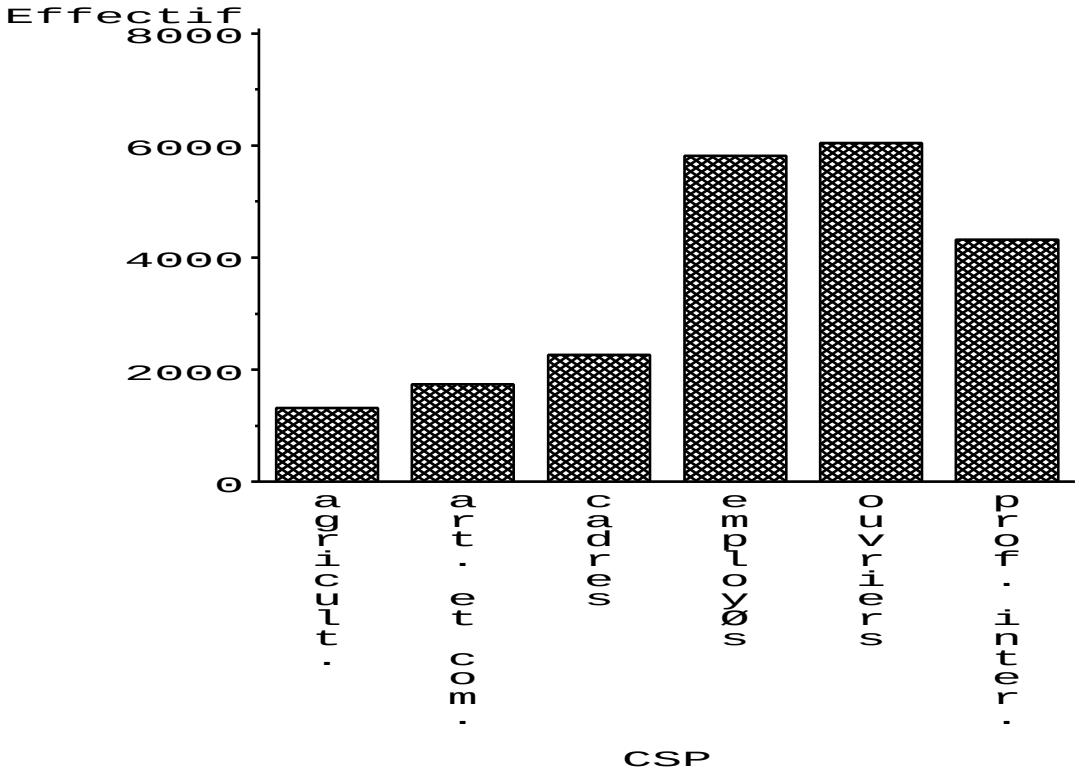
– le *diagramme en secteurs.*

Les figures II.6, II.7 et II.8 présentent chacun de ces trois graphiques sur les données de l'exemple II.3.

Exemple II.3 *Le tableau ci-dessous donne la répartition de la population active occupée (ayant effectivement un emploi) selon la CSP (catégorie socioprofessionnelle), en France, en mars 1988 (Tableaux de l'Economie Française, INSEE, 1989, p. 59).*

CSP	effectifs en milliers	fréquences (%)
1. agriculteurs exploitants	1312	6,1
2. artisans, commerçants, chefs d'entreprises	1739	8,1
3. cadres, professions intellectuelles supérieures	2267	10,6
4. professions intermédiaires	4327	20,1
5. employés	5815	27,0
6. ouvriers	6049	28,1

FIG. II.6 - Diagramme en barre



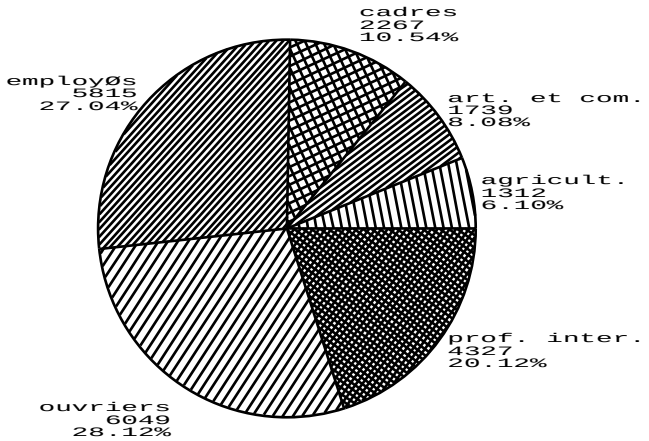


FIG. II.8 - Diagramme en secteurs

Chapitre III

Cas bidimensionnel

Dans tout ce chapitre, on s'intéresse à l'étude simultanée de deux variables, notées X et Y, étudiées sur le même échantillon, toujours noté Ω. L'objectif essentiel des méthodes présentées est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors *liaison*. Dans certains cas, cette liaison peut être considérée a priori comme *causale*, une variable expliquant l'autre ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations.

1 Deux variables quantitatives

1.1 Les données

Les variables quantitatives considérées dans ce chapitre seront toujours discrètes ; en effet, cela n'a pas d'intérêt de regrouper les valeurs identiques de l'une des deux variables, puisqu'elles ne correspondent pas nécessairement à des valeurs identiques de l'autre. La donnée de base est donc constituée par la série statistique brute se présentant sous la forme $\{(X(\omega_i), Y(\omega_i))\ ; i = 1, \ldots, n\}$.

Exemple III.1 *Les données ci-dessous seront utilisées tout au long de ce chapitre ; elles sont extraites des Tableaux de l'Economie Française, INSEE, 1989, p.109. Pour 51 secteurs d'activité (numérotés de 04 à 54, selon la nomenclature NAP 100), on a considéré (au 01.01.1986, en France) le nombre total d'entreprises (variable notée NB), l'effectif salarié (notée EF), et le chiffre d'affaires hors-taxes en millions de francs (notée CA). Dans le tableau ci-dessous, nous donnons les trois premières et les trois dernières lignes de l'ensemble des données.*

code	secteur	NB	EF	CA
04	production de combustibles minéraux solides et cokéfaction	19	49251	14111
05	production de pétrole et de gaz naturel	120	46594	306293
06	production et distribution d'electricité	731	129723	138389
⋮	⋮	⋮	⋮	⋮
52	industrie du caoutchouc	746	87121	37502
53	transformation des matières plastiques	3232	102437	58122
54	industries diverses	10171	84012	38071

Pour mémoire, nous indiquons ci-dessous quelques caractéristiques numériques des trois variables considérées.

variable	minimum	maximum	moyenne	écart-type
NB	11	41866	4135	7435
EF	1701	425082	92591	83832
CA	992	306293	68010	64532

1.2 Représentation graphique : le nuage de points

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y, puis à représenter chaque individu observé ω_i par le point d'abscisse $X(\omega_i)$ et d'ordonnée $Y(\omega_i)$. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables et est appelé *nuage*. On notera qu'on rencontre parfois la terminologie de *diagramme de dispersion*, traduction fidèle du terme anglais *scatter-plot*.

La figure III.1 présente le nuage de points réalisé, avec les données de l'exemple III.1, en utilisant les variables CA (en ordonnées) et EF (en abscisses). De plus, on a tracé la droite de régression de CA sur EF (voir le paragraphe 1.4).

Remarques III.1 *Le choix des échelles à retenir pour réaliser un nuage de points peut s'avérer délicat. D'une façon générale, on distinguera le cas de variables* homogènes *(représentant la même grandeur et exprimées dans la même unité) de celui des variables* hétérogènes. *Dans le premier cas, on choisira la même échelle sur les deux axes (qui seront donc orthonormés) ; dans le second cas, il est recommandé soit de représenter les variables centrées et réduites (voir ci-dessous) sur des axes orthonormés, soit de choisir des échelles telles que ce soit*

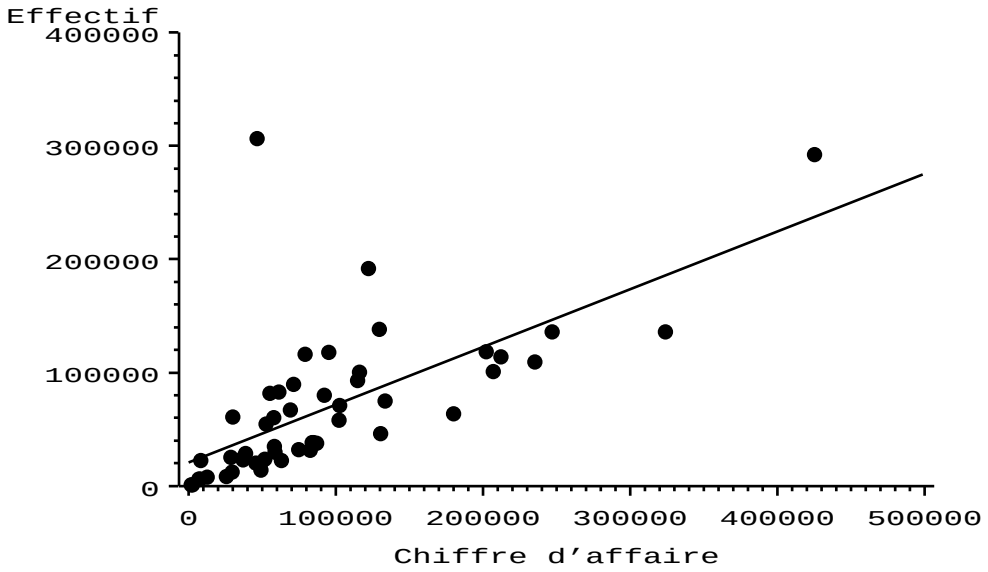


FIG. III.1 - Nuage de points

sensiblement ces variables là que l'on représente (c'est en général cette seconde solution qu'utilisent, de façon automatique, les logiciels statistiques).

Rappel : variables centrées et réduites

Si X est une variable quantitative de moyenne $\overline{x}$ et d'écart-type σ_X, on appelle variable centrée associée à X la variable $X - \overline{x}$ (elle est de moyenne nulle et d'écart-type σ_X), et variable centrée et réduite (ou tout simplement variable réduite) associée à X la variable $\frac{X - \overline{x}}{\sigma_X}$ (elle est de moyenne nulle et d'écart-type égal à un). Une variable centrée et réduite s'exprime sans unité.

1.3 La covariance et le coefficient de corrélation linéaire

L'objectif est maintenant de définir un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Cet indice est le coefficient de corrélation linéaire ; il nécessite la définition préalable de la covariance.

La covariance : définition

La covariance généralise à deux variables la notion de variance ; sa formule de définition est la suivante :

$$cov(X,Y) = c_{XY} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}[X(\omega_i) - \overline{x}][Y(\omega_i) - \overline{y}]$$

$$= \quad [\frac{1}{n}\sum_{i=1}^{n} X(\omega_i)Y(\omega_i)] - \overline{x}\ \overline{y}.$$

La covariance : propriétés

- La covariance est un indice *symétrique*. De façon évidente, on a $c_{XY} = c_{YX}$ (les deux variables jouent donc le même rôle dans la définition de la covariance).

- La covariance peut prendre *toute valeur réelle* (négative, nulle ou positive).

- La covariance est une *forme bilinéaire symétrique* dont la variance est la *forme quadratique* associée. En particulier, on en déduit les deux formules suivantes :

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y),$$

$$[cov(X,Y)]^2 \leq var(X)var(Y)\,;$$

(cette dernière propriété est l'inégalité de Cauchy-Schwarz).

Le coefficient de corrélation linéaire : définition

Il est clair que la covariance dépend des unités de mesure dans lesquelles sont exprimées les variables considérées ; en ce sens, ce n'est pas un indice de liaison "intrinsèque". C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (parfois appelé coefficient de Pearson ou de Bravais-Pearson), rapport entre la covariance et le produit des écarts-types :

$$corr(X,Y) = r_{XY} = \frac{c_{XY}}{\sigma_X \sigma_Y}.$$

Le coefficient de corrélation linéaire : propriétés

- Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à X et Y : $r_{XY} = cov(\frac{X-\overline{x}}{\sigma_X}, \frac{Y-\overline{y}}{\sigma_Y})$. Par conséquent, r_{XY} est indépendant des unités de mesure de X et de Y.

- Le coefficient de corrélation est *symétrique* : $r_{XY} = r_{YX}$.

- $-1 \leq r_{XY} \leq +1$. Les valeurs -1 et $+1$ correspondent à une liaison linéaire parfaite entre X et Y (existence de réels a, b et c tels que : $aX+bY+c=0$).

Illustration

En reprenant les données de l'exemple III.1, nous avons calculé la covariance et le coefficient de corrélation linéaire entre les variables EF et CA ; on a obtenu :

$$cov(EF, CA) = 35769 \times 10^5 \ ; \ corr(EF, CA) = \frac{35769 \times 10^5}{83832 \times 64532} \simeq 0,66 \ .$$

La liaison linéaire entre les deux variables est donc positive et moyenne.

1.4 Regression linéaire entre deux variables

Introduction

Lorsque deux variables quantitatives sont correctement correlées ($|r_{XY}|$ voisin de 1), et que l'on peut a priori considérer que l'une (nous supposerons qu'il s'agit de X) est cause de l'autre (il s'agira donc de Y), il est naturel de chercher, dans un ensemble donné de fonctions, la fonction de X approchant Y "le mieux possible", au sens d'un certain critère ; on dit que l'on fait la regression de Y sur X. Si l'on choisit pour ensemble de fonctions celui des fonctions affines (du type $aX+b$), on parle alors de régression linéaire. C'est le choix que l'on fait le plus fréquemment dans la pratique, le critère le plus usuel étant celui des moindres carrés.

Le critère des moindres carrés

Il consiste à minimiser la quantité suivante :

$$S(a, b) = \sum_{i=1}^{n} \{Y(\omega_i) - [aX(\omega_i) + b]\}^2.$$

On notera que $|Y(\omega_i) - [aX(\omega_i) + b]|$ représente, dans le nuage de points, la distance verticale du point figurant ω_i à la droite d'équation $y = ax + b$.

Solution

La minimisation de S en a et b fournit la solution unique suivante :

$$\widehat{a} = \frac{c_{XY}}{\sigma_X^2} \ ; \ \widehat{b} = \overline{y} - \widehat{a}\overline{x}.$$

Propriétés

- La droite d'équation $y = \widehat{a}x + \widehat{b}$ est appelée *droite de régression* de Y sur X ; elle passe par le *barycentre* du nuage des ω_i, de coordonnées $(\overline{x}, \overline{y})$.
- Les valeurs $\widehat{y}_i = \widehat{a}X(\omega_i) + \widehat{b}$ sont appelées les *valeurs ajustées* ; elles ont la même moyenne $\overline{y}$ que Y.

- Les valeurs $\widehat{e}_i = Y(\omega_i) - \widehat{y}_i$ sont appelées les *résidus*. Ils sont de moyenne nulle et de variance $S(\widehat{a}, \widehat{b})$.

- La variable causale X et la variable résiduelle $\widehat{E}$ sont non correlées :

$$corr(X, \widehat{E}) = 0.$$

Illustration

En considérant toujours l'exemple III.1, nous avons réalisé la régression linéaire de la variable CA sur la variable EF. On a obtenu :

$$\hat{a} = \frac{cov(EF, CA)}{var(EF)} = \frac{35769 \times 10^5}{70278 \times 10^5} \simeq 0,509 \ ;$$

$$\hat{b} = \overline{CA} \ - \ \hat{a} \ \overline{EF} = 68010 - 0,509 \times 92591 \simeq 20884 \ .$$

La droite de régression correspondante a été tracée sur la figure III.1.

2 Une variable quantitative et une qualitative

2.1 Les données

Soit X la variable qualitative considérée, supposée à r modalités notées

$$x_1, \ldots, x_l, \ldots, x_r$$

, et soit Y la variable quantitative de moyenne $\overline{y}$ et de variance σ_Y^2. Désignant toujours par Ω l'échantillon considéré, chaque modalité x_l de X définit une sous-population (un sous-ensemble) Ω_l de Ω : c'est l'ensemble des individus sur lesquels on a observé x_l ; on obtient ainsi une *partition* de Ω en r classes dont nous noterons $n_1, \ldots, n_r$ les cardinaux (avec toujours $\sum_{l=1}^{r} n_l = n$, où $n = card(\Omega)$).

Considérant alors la restriction de Y à Ω_l $(l = 1, \ldots, r)$, on peut définir la moyenne et la variance partielles de Y sur cette sous-population ; nous les noterons respectivement $\overline{y_l}$ et σ_l^2 :

$$\overline{y_l} = \frac{1}{n_l} \sum_{\omega_i \in \Omega_l} Y(\omega_i) \, ;$$

$$\sigma_l^2 = \frac{1}{n_l} \sum_{\omega_i \in \Omega_l} [Y(\omega_i) - \overline{y_l}]^2.$$

Exemple III.2 *Cet exemple est présenté dans Johnson & Wichern (1988), page 218 ; les individus sont 19 chiens prémédiqués au pentobarbital, et l'on étudie l'effet sur leur rythme cardiaque de deux facteurs (variables qualitatives explicatives) croisés. L'effet est mesuré par le temps entre deux battements de cœur successifs*

(variable quantitative Y, mesurée en millisecondes), et les deux facteurs, à deux niveaux chacun, sont la pression d'administration de dioxyde de carbone (CO_2), qui peut être élevée (E) ou faible (F), et la présence (1) ou l'abscence (0) d'halothane ; la variable X est celle obtenue par le croisement de ces deux facteurs ; elle est donc qualitative à 4 modalités : $x_1 = E0, x_2 = F0, x_3 = E1, x_4 = F1$. L'expérience ayant été répétée 4 fois sur chaque chien (une fois dans chacune des 4 conditions ainsi définies), on dispose donc de $n = 4 \times 19 = 76$ individus. Les données se trouvent dans le tableau ci-dessous.

	modalité de X			
numéro du chien	x_1	x_2	x_3	x_4
1	426	609	556	600
2	253	236	392	395
3	359	433	349	357
4	432	431	522	600
5	405	426	513	513
6	324	438	507	539
7	310	312	410	456
8	326	326	350	504
9	375	447	547	548
10	286	286	403	422
11	349	382	473	497
12	429	410	488	547
13	348	377	447	514
14	412	473	472	446
15	347	326	455	468
16	434	458	637	524
17	364	367	432	469
18	420	395	508	531
19	397	556	645	625

Nous indiquons également les moyennes et les écarts-types partiels des 4 sous-populations, ainsi que la moyenne et l'écart-type de la population globale.

	Ω_1	Ω_2	Ω_3	Ω_4	Ω
moyennes	368,2	404,6	479,3	502,9	438,8
écarts-types	51,7	86,9	80,6	68,0	91,1

Remarques III.2 *Ces données sont un peu particulières, dans la mesure où d'une part les individus sont dupliqués 4 fois, et d'autre part les 4 modalités de la variable X correspondent au croisement de deux facteurs. Ainsi, d'autres*

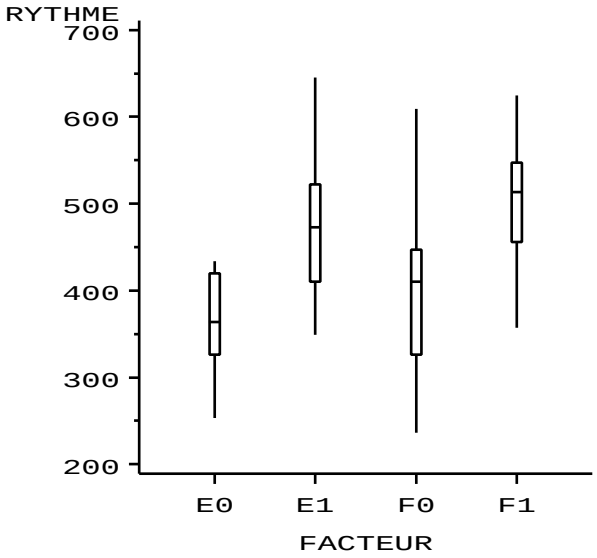


FIG. III.2 - Boîtes parallèles

traitements statistiques que ceux indiqués ici sont envisageables (voir Johnson & Wichern, 1988).

2.2 Représentation graphique : les boîtes parallèles

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des boîtes parallèles ; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour Y une boîte-à-moustaches pour chacune des sous-populations définies par X. La comparaison de ces boîtes donne une idée assez claire de l'influence de X sur les valeurs de Y, c'est-à-dire de la liaison entre les deux variables. La figure III.2 donne les boîtes parallèles de l'exemple III.2.

2.3 Formules de décomposition

Ces formules indiquent comment se décomposent la moyenne et la variance de Y sur la partition définie par X (c'est-a-dire comment s'écrivent ces caractéristiques en fonction de leurs valeurs partielles) ; elles sont nécessaires pour définir un indice de liaison entre les deux variables. Ces formules sont les suivantes :

$$\begin{aligned} \overline{y} &= \frac{1}{n}\sum_{l=1}^{r} n_l \overline{y_l}\,; \\ \sigma_Y^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l(\overline{y_l}-\overline{y})^2 \;+\; \frac{1}{n}\sum_{l=1}^{r} n_l\sigma_l^2 = \sigma_E^2 + \sigma_R^2\ . \end{aligned}$$

Le premier terme de la décomposition de σ_Y^2, noté σ_E^2, est appelé *variance expliquée* (par la partition, c'est-à-dire par X) ; le second terme, noté σ_R^2, est appelé *variance résiduelle*.

On notera qu'une formule de décomposition analogue existe pour la covariance entre deux variables quantitatives.

2.4 Le rapport de correlation

Définition

Il s'agit d'un indice de liaison entre les deux variables X et Y ; il est défini de la façon suivante :

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}} \, .$$

Propriétés

- $s_{Y/X}$ n'est pas symétrique. Cette propriété est évidente, compte-tenu que X et Y ne sont pas de même nature.

- $0 \leq s_{Y/X} \leq 1$. Cet encadrement de $s_{Y/X}$ découle directement de la formule de décomposition de la variance. Les valeurs 0 et 1 ont une signification particulière intéressante.

Illustration

Sur l'exemple III.2 la variance totale vaut 8305,90 , la variance expliquée 2973,94 et la variance résiduelle 5331,96. On en déduit que le rapport de corrélation vaut :

$$s_{Y/X} = \sqrt{\frac{2973,94}{8305,90}} \simeq 0,60.$$

La liaison entre X et Y est donc moyenne.

3 Deux variables qualitatives

3.1 Les données et leur présentation

On considère dans ce paragraphe deux variables qualitatives observées simultanément sur n individus. On suppose que la première, notée X, possède r modalités notées $x_1, \ldots, x_l, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

Le plus souvent, ces données sont présentées dans un tableau à double entrée, appelé *table de contingence*, dans lequel on dispose les modalités de X en lignes

et celles de Y en colonnes. Ce tableau est donc de dimension $r \times c$ et a pour élément générique le nombre n_{lh} d'observations conjointes des modalités x_l de X et y_h de Y ; les quantités n_{lh} sont appelées les *effectifs conjoints*.

Une table de contingence se présente donc sous la forme suivante :

	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_{l1}	$\cdots$	n_{lh}	$\cdots$	n_{lc}	n_{l+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Les quantités $n_{l+} (l = 1, \ldots, r)$ et $n_{+h} (h = 1, \ldots, c)$ sont appelées les *effectifs marginaux*; ils sont définis par $n_{l+} = \sum_{h=1}^{c} n_{lh}$ et $n_{+h} = \sum_{l=1}^{r} n_{lh}$, et ils vérifient $\sum_{l=1}^{r} n_{l+} = \sum_{h=1}^{c} n_{+h} = n$.

De façon analogue, on peut définir les notions de fréquences conjointes et de fréquences marginales.

Exemple III.3 *Dans cet exemple, on a considéré un échantillon de 797 étudiants de l'Université Paul Sabatier (Toulouse) ayant obtenu soit le DEUG A soit le DEUG B (diplômes scientifiques de premier cycle), et uniquement ce diplôme, durant la période 1971-1983. Quatre variables ont été prises en compte, toutes qualitatives : la série de bac, à 2 modalités (C ou E, D), la mention au bac, à 4 modalités (très bien ou bien, assez bien, passable, inconnue), l'âge d'obtention du bac, à 4 modalités (moins de 18 ans, 18 ans, 19 ans, plus de 19 ans), et la durée d'obtention du DEUG, à 3 modalités (2 ans, 3 ans, 4 ans). Dans la table de contingence ci-dessous, on a croisé en lignes la durée d'obtention du DEUG (variable X, à $r = 3$ modalités), et en colonnes l'âge d'obtention du bac (variable Y, à $c = 4$ modalités).*

	$<$ 18 ans	18 ans	19 ans	$>$ 19 ans	sommes
2 ans	84	224	73	19	400
3 ans	35	137	75	27	274
4 ans	14	59	34	16	123
sommes	133	420	182	62	797

3.2 Les représentations graphiques

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'*adapter* les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de

l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas et que l'on appelle les *profils*.

Définition des profils

On appelle $l^{\text{ième}}$ profil-ligne l'ensemble des fréquences de la variable Y conditionnelles à la modalité x_l de X (c'est-à-dire définies au sein de la sous-population Ω_l de Ω associée à cette modalité). Il s'agit donc des quantités :

$$\{\frac{n_{l1}}{n_{l+}}, \ldots, \frac{n_{lh}}{n_{l+}}, \ldots, \frac{n_{lc}}{n_{l+}}\}.$$

On définit de façon analogue le $h^{\text{ième}}$ profil-colonne :

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

La représentation graphique des profils-lignes ou des profils-colonnes, au moyen, par exemple, de diagrammes en barre parallèles, donne alors une idée assez précise de la variation conjointe des deux variables.

Illustration

Nous avons déterminé les profils-colonnes (en pourcentages) relatifs à la table de contingence présentée dans l'exemple III.3 et nous les donnons ci-dessous.

	$<$ 18 ans	18 ans	19 ans	$>$ 19 ans	profil moyen
2 ans	63,2	53,3	40,1	30,6	50,2
3 ans	26,3	32,6	41,2	43,6	34,4
4 ans	10,5	14,1	18,7	25,8	15,4
sommes	100,0	100,0	100,0	100,0	100,0

La figure III.3 donne les diagrammes en barre pour les profils-colonnes ci-dessus, et une liaison entre les deux variables étudiées apparaît très clairement.

3.3 Les indices de liaison : le khi-deux et ses dérivés

Propriété préliminaire

On peut établir l'équivalence des trois propriétés suivantes :

i. Tous les profils-lignes sont égaux .

ii. Tous les profils-colonnes sont égaux.

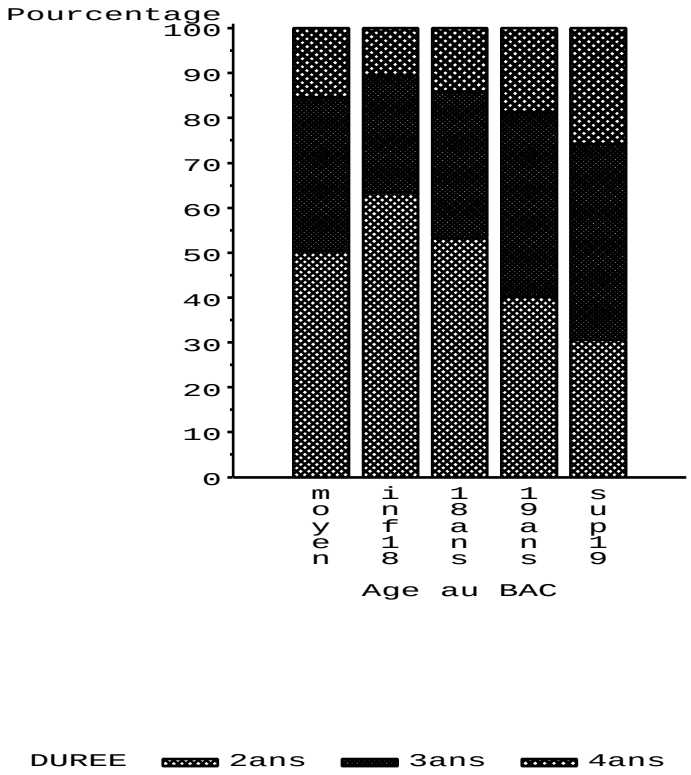


FIG. III.3 - Diagrammes en barre

iii. $\forall (l,h) \in \{1,\ldots,r\} \times \{1,\ldots,c\}$:

$$n_{lh} = \frac{n_{l+} n_{+h}}{n} \ .$$

Si une table de contingence vérifie ces trois propriétés, on peut alors dire qu'il n'existe aucune forme de liaison entre les deux variables considérées X et Y. Par suite, la mesure de la liaison va se faire en évaluant l'écart entre la situation observée et l'état de non liaison que nous définirons par (iii).

Définition du khi-deux

Il est courant en statistique de comparer une table de contingence observée, d'effectif conjoint générique n_{lh}, à une table de contingence donnée a priori (et appelée *standard*), d'effectif conjoint générique s_{lh}, en calculant la quantité

$$\sum_{l=1}^{r} \sum_{h=1}^{c} \frac{(n_{lh} - s_{lh})^2}{s_{lh}} \ .$$

De façon naturelle, pour mesurer la liaison sur une table de contingence, on utilise donc l'indice appelé khi-deux (chi-square) et défini comme suit :

$$\chi^2 = \sum_{l=1}^{r} \sum_{h=1}^{c} \frac{(n_{lh} - \frac{n_{l+} n_{+h}}{n})^2}{\frac{n_{l+} n_{+h}}{n}}.$$

Le coefficient χ^2 est toujours positif ou nul et il est d'autant plus grand que la liaison entre les deux variables considérées est forte. Malheureusement, il dépend aussi des dimensions r et c de la table étudiée, ainsi que de la taille n de l'échantillon observé ; en particulier, il n'est pas majoré. C'est la raison pour laquelle on a défini d'autres indices, liés au khi-deux, et dont l'objectif est de palier ces défauts.

Autres indicateurs liés au khi-deux

Nous en citerons trois.

- Le *phi-deux*: $\Phi^2 = \frac{\chi^2}{n}$. Il ne dépend plus de n, mais dépend encore de r et de c.

- Le coefficient T de Tschuprow :

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}.$$

  On peut vérifier : $0 \leq T \leq 1$.

– Le coefficient C de Cramer :

$$C = \sqrt{\frac{\Phi^2}{d-1}},$$

avec : $d = \inf(r, c)$. On vérifie maintenant : $0 \leq T \leq C \leq 1$.

Illustration

En utilisant les données de l'exemple III.3, on a obtenu :

$$\chi^2 = 28,7\,;\ \Phi^2 = 0,036\,;\ T = 0,12\,;\ C = 0,13.$$

On a en fait affaire à une liaison peu importante.

4 Vers le cas multidimensionnel

L'objectif des prochains chapitres de ce cours est d'exposer les techniques de la statistique descriptive multidimensionnelle. Or, sans connaître ces techniques, il se trouve qu'il est possible de débuter une exploration de données multidimensionnelles en adaptant simplement les méthodes déjà étudiées. C'est cette idée que nous développons dans ce paragraphe.

4.1 Les matrices des variances-covariances et des corrélations

Lorsqu'on a observé simultanément plusieurs variables quantitatives (p variables, $p \geq 3$) sur le même échantillon, il est possible de calculer d'une part les variances de toutes ces variables, d'autre part les $\frac{p(p-1)}{2}$ covariances des variables prises deux à deux. L'ensemble de ces quantités peut alors être disposé dans une matrice carrée ($p \times p$) et symétrique, comportant les variances sur la diagonale et les covariances à l'extérieur de la diagonale ; cette matrice, appelée matrice des variances-covariances (ou encore matrice des covariances) sera notée **S**. Elle sera utilisée par la suite, mais n'a pas d'interprétation concrète. Notons qu'il est possible de vérifier que **S** est semi-définie-positive.

De la même manière, on peut construire la matrice symétrique $p \times p$, comportant des 1 sur toute la diagonale et, en dehors de la diagonale, les coefficients de corrélation linéaire entre les variables prises deux à deux. Cette matrice est appelée matrice des corrélations, elle est également semi-définie-positive, et nous la noterons **R**. Elle est de lecture commode et indique quelle est la structure de corrélation des variables étudiées.

Illustration

Nous avons repris ici encore l'exemple III.1 et calculé les matrices des variances-covariances et des corrélations entre les trois variables intervenant.

$$\mathbf{S} = \begin{bmatrix} 55284 \times 10^3 & 25084 \times 10^4 & 25156 \times 10^3 \\ 25084 \times 10^4 & 70278 \times 10^5 & 35769 \times 10^5 \\ 25156 \times 10^3 & 35769 \times 10^5 & 41644 \times 10^5 \end{bmatrix};$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0,402 & 0,052 \\ 0,402 & 1 & 0,661 \\ 0,052 & 0,661 & 1 \end{bmatrix}.$$

4.2 Les tableaux de nuages (ou graphiques splom)

Notons $X^1, \ldots, X^p$ les p variables quantitatives considérées ; on appelle tableau de nuages le graphique obtenu en juxtaposant, dans une sorte de matrice carrée $p \times p$, p^2 sous-graphiques ; chacun des sous-graphiques diagonaux est relatif à l'une des p variables, et il peut s'agir, par exemple, d'un histogramme ; le sous-graphique figurant dans le bloc d'indice $(j, j'), j \neq j'$, est le nuage de points réalisé avec la variable X^j en abscisses et la variable $X^{j'}$ en ordonnées. Dans certains logiciels anglo-saxons, ces graphiques sont appelés *splom* (Scatter PLOt Matrix). On notera qu'il est possible de ne représenter que la partie supérieure (ou inférieure) du tableau en question. Il est également possible de faire apparaître dans chaque sous-graphique non diagonal une droite de régression. Le tableau de nuages, avec la matrice des corrélations, fournit ainsi une vision globale des liaisons entre les variables étudiées.

Illustration

La figure III.4 présente le tableau de nuages relatif aux données de l'exemple III.1.

4.3 La matrice des coefficients de Tschuprow (ou de Cramer)

Considérons maintenant le cas où l'on étudie simultanément plusieurs variables qualitatives (p variables, $p \geq 3$). La matrice des coefficients de Tschuprow est la matrice carrée d'ordre p, symétrique, comportant des 1 sur la diagonale et, en dehors de la diagonale, les coefficients de Tschuprow entre les variables prises deux à deux. Il s'agit donc d'une matrice du même type que la matrice des corrélations (elle est d'ailleurs, elle aussi, semi-définie-positive), et son utilisation pratique est analogue. Notons que l'on peut, de la même façon, utiliser les coefficients de Cramer au lieu des coefficients de Tschuprow.

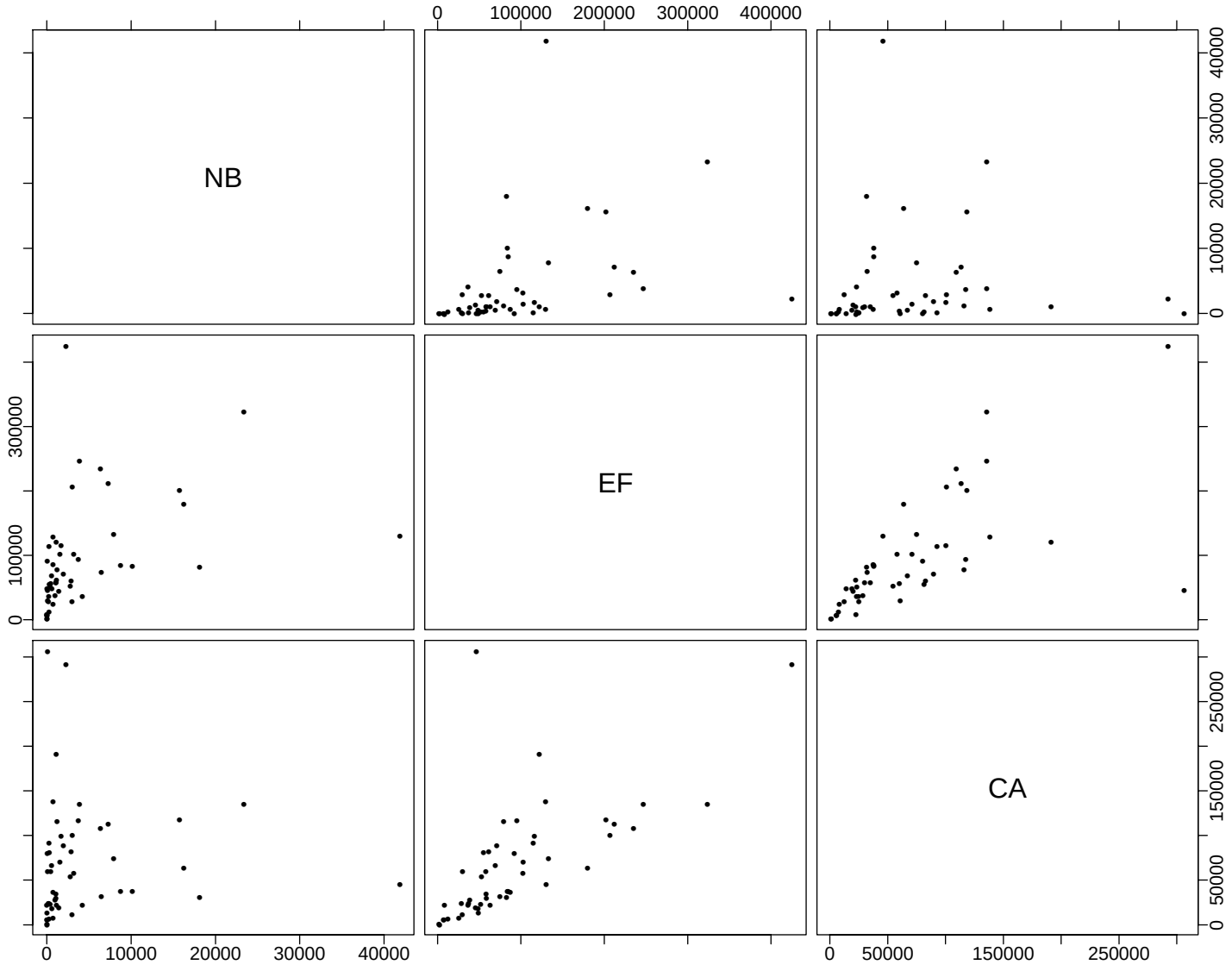


FIG. III.4 - Tableau des nuages

	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

TAB. III.1 - Tableau de Burt

Illustration

La matrice des coefficients de Tchuprow relative aux données de l'exemple III.3 est la suivante :

$$\mathbf{T} = \begin{bmatrix} 1 & 0,08 & 0,13 & 0,18 \\ 0,08 & 1 & 0,11 & 0,16 \\ 0,13 & 0,11 & 1 & 0,12 \\ 0,18 & 0,16 & 0,12 & 1 \end{bmatrix}.$$

4.4 Le tableau de Burt

Le tableau de Burt est une généralisation particulière de la table de contingence dans le cas où l'on étudie simultanément p variables qualitatives. Notons $X^1, \ldots, X^p$ ces variables et appelons c_j le nombre de modalités de $X^j, j = 1, \ldots, p$ et $c = \sum_{j=1}^{p} c_j$. Le tableau de Burt est en fait une matrice carrée $c \times c$, 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. Il apparaît en fait comme l'analogue qualitatif du tableau des nuages.

Illustration

Toujours avec les données de l'exemple III.3, nous avons déterminé le tableau de Burt pour les trois variables série de bac, âge au bac, et durée du DEUG.

Chapitre IV

Analyse en Composantes Principales

L'Analyse en Composantes Principales (A.C.P.) est introduite ici comme l'estimation d'un modèle, ainsi que l'a suggéré Caussinus (1986), afin de préciser la signification statistique des résultats obtenus et rendre possible la discussion de certains problèmes d'optimalité (choix de dimension, de métrique).

Cette technique est illustrée dans ce chapitre à travers l'étude de deux jeux de données :

températures Les données sont constituées des moyennes sur dix ans des températures moyennes mensuelles de 32 villes françaises. La matrice initiale $\mathbf{Y}$ est donc (32×12). Les colonnes sont l'observation à différents instants d'une même variable ; elles sont homogènes et il est inutile de les réduire.

criminalité Le jeu de données suivant est extrait de la documentation de SAS (1989) où il illustre l'utilisation de la procédure `princomp`. Les $p = 7$ variables sont des taux de criminalité, selon différents types de délits, observés dans les $n = 50$ états des USA.

Ce chapitre ne contient que les résultats graphiques nécessaires à l'interprétation ; les tableaux numériques sont reportés en annexe.

1 Introduction

Soit p *variables statistiques* réelles $Y^j, j = 1 \ldots, p$ observées sur n individus $i = 1, \ldots, n$ affectés des poids w_i :

$$\forall i = 1, \ldots, n \ : w_i \qquad 0 \text{ et } \sum_{i=1}^{n} w_i = 1 \ ;$$

$$\forall i = 1, \ldots, n \ : y_i^j \ = \ Y^j(i), \text{ mesure de } Y^j \text{ sur le } i\text{ème individu.}$$

Ces mesures sont regroupées dans une matrice $\mathbf{Y}$ d'ordre $(n \times p)$:

$$\mathbf{Y} = \begin{bmatrix} y_1^1 & \dots & y_1^j & \dots & y_1^p \\ \vdots & & \vdots & & \vdots \\ y_i^1 & \dots & y_i^j & \dots & y_i^p \\ \vdots & & \vdots & & \vdots \\ y_n^1 & \dots & y_n^j & \dots & y_n^p \end{bmatrix}.$$

1.1 Représentation vectorielle de données quantitatives

– A chaque individu i est associé le vecteur y_i contenant la $i^{ème}$ ligne de $\mathbf{Y}$ mise en colonne. C'est un élément d'un espace vectoriel noté E de dimension p ; nous choisissons $\mathbb{R}^p$ muni de la base canonique $\mathcal{E}$ et d'une métrique de matrice $\mathbf{M}$ lui conférant une structure d'espace euclidien :

$$E \text{ est isomorphe à } (\mathbb{R}^p, \mathcal{E}, \mathbf{M}).$$

E est alors appelé *espace des individus*.

– A chaque variable Y^j est associé le vecteur y^j contenant la $j^{ème}$ colonne de $\mathbf{Y}$. C'est un élément d'un espace vectoriel noté F de dimension n ; nous choisissons $\mathbb{R}^n$ muni de la base canonique $\mathcal{F}$ et d'une métrique de matrice $\mathbf{D}$ diagonale des *poids* lui conférant une structure d'espace euclidien :

$$F \text{ est isomorphe à } (\mathbb{R}^n, \mathcal{F}, \mathbf{D}) \text{ avec } \mathbf{D} = \mathrm{diag}(w_1, \dots, w_n).$$

F est alors appelé *espace des variables*.

1.2 Interprétation statistique de la métrique des poids

L'utilisation de la métrique des poids dans l'espace des variables F donne un sens très particulier aux notions usuelles définies sur les espaces euclidiens. Ce paragraphe est la clé permettant de fournir les interprétations en termes statistiques des propriétés et résultats mathématiques.

Moyenne empirique de Y^j :	$\overline{y^j}$	$=$	$\langle y^j, \mathbf{1}_n \rangle_{\mathbf{D}} = y^{j\prime}\mathbf{D}\mathbf{1}_n$.
Barycentre des individus :	$\bar{y}$	$=$	$\mathbf{Y}'\mathbf{D}\mathbf{1}_n$.
Centrage de Y^j :	x^j	$=$	$y^j - \overline{y^j}\mathbf{1}_n$.
Matrice des données centrées :	$\mathbf{X}$	$=$	$\mathbf{Y} - \mathbf{1}_n\overline{y}'$.
Ecart-type de Y^j :	σ_j	$=$	$(x^{j\prime}\mathbf{D}x^j)^{1/2} = \|x^j\|_{\mathbf{D}}$.
Covariance de Y^j et Y^k :	$x^{j\prime}\mathbf{D}x^k$	$=$	$\langle x^j, x^j \rangle_{\mathbf{D}}$.
Matrice des covariances :	$\mathbf{S}$	$=$	$\sum_{i=1}^n w_i x_i x_i' = \mathbf{X}'\mathbf{D}\mathbf{X}$.
Corrélation de Y^j et Y^k :	$\frac{\langle x^j, x^k \rangle_{\mathbf{D}}}{\|x^j\|_{\mathbf{D}} \|x^k\|_{\mathbf{D}}}$	$=$	$\cos\theta_{\mathbf{D}}(x^j, x^k)$.

Ainsi, lorsque les variables sont centrées et représentées par des vecteurs de F :

- la *longueur* d'un vecteur représente un *écart-type*,
- le *cosinus* d'un angle entre deux vecteurs représente une *corrélation*.

1.3 La méthode

Les objectifs poursuivis par une A.C.P. sont :

- la représentation graphique "optimale" des individus (lignes), minimisant les déformations du nuage des points, dans un sous-espace E_q de dimension q ($q < p$),
- la représentation graphique des variables dans un sous-espace F_q en explicitant au "mieux" les liaisons initiales entre ces variables,
- la réduction de la dimension (compression), ou approximation de Y par un tableau de rang q ($q < p$).

Définition IV.1 *Soit $Y^1, \ldots, Y^p$, p variables quantitatives observées sur n individus de poids w_i. L'Analyse en Composantes Principales (A.C.P.) du triplet $(\mathbf{Y}, \mathbf{M}, \mathbf{D})$ est obtenue par la décomposition en valeurs singulières de $(\mathbf{X}, \mathbf{M}, \mathbf{D})$.*

Des arguments de type géométrique dans la littérature francophone, ou bien de type statistique avec une hypothèse de normalité dans la littérature anglo-saxonne, justifient le plus souvent cette technique. Nous adoptons ici une optique intermédiaire en se référant à un modèle "allégé" car ne nécessitant pas d'hypothèse "forte" sur la distribution des observations (normalité).

Plus précisément, l'A.C.P. admet des définitions équivalentes selon que l'on s'attache à la représentation des individus, à celle des variables ou encore à leur représentation simultanée.

2 Point de vue des individus

Les notations sont celles du paragraphe précédent :

- $\mathbf{Y}$ désigne le tableau des données issues de l'observation de p variables *quantitatives* Y^j sur n individus i de *poids* w_i,
- E est l'espace des individus muni de la base canonique et de la métrique de matrice $\mathbf{M}$,
- F est l'espace des variables muni de la base canonique et de la métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

2.1 Modèle

De façon générale, un modèle s'écrit :

$$\textbf{Observation} = \textbf{Modèle} + \textbf{Bruit}$$

assorti de différents types d'hypothèses et de contraintes sur le modèle et sur le bruit.

En ACP, la matrice des données est supposée être issue de l'observation de n vecteurs aléatoires indépendants $\{y_1, \ldots, y_n\}$, de même matrice de covariance $\sigma^2\mathbf{\Gamma}$, mais d'espérances différentes z_i, toutes contenues dans un sous-espace affine de dimension q ($q < p$) de E. Dans ce modèle, $E(y_i) = z_i$ est un paramètre spécifique attaché à chaque individu i et appelé *effet fixe*, le modèle étant dit *fonctionnel*.

Ceci s'écrit en résumé :

$$\begin{array}{l} \{y_i \ ; i = 1, \ldots, n\},\ n \text{ vecteurs aléatoires indépendants de } E, \\ y_i = z_i + \varepsilon_i \ , i = 1, \ldots, n \text{ avec } \left\{ \begin{array}{l} E(\varepsilon_i) = 0,\ \mathrm{var}(\varepsilon_i) = \sigma^2\mathbf{\Gamma}, \\ \sigma > 0 \text{ inconnu, } \mathbf{\Gamma} \text{ régulière et connue,} \end{array} \right. \\ \exists A_q,\ \text{sous-espace affine de dimension } q \text{ de } E \text{ tel que } \forall i, z_i \in A_q \ (q < p). \end{array} \tag{IV.1}$$

Soit $\bar{z} = \sum_{i=1}^{n} w_i z_i$. Les hypothèses du modèle entraînent que $\bar{z}$ appartient à A_q. Soit donc E_q le sous-espace vectoriel de E de dimension q tel que :

$$A_q = \bar{z} + E_q.$$

Les paramètres à estimer sont alors E_q et $z_i, i = 1, \ldots, n$, éventuellement σ ; z_i est la part systématique, ou *effet*, supposée de rang q ; éliminer le bruit revient donc à réduire la dimension.

Si les z_i sont considérés comme *aléatoires*, le modèle est alors dit *structurel* ; on suppose que $\{y_1, \ldots, y_n\}$ est un échantillon statistique i.i.d. Les unités statistiques jouent des rôles symétriques, elles ne nous intéressent que pour l'étude des relations entre les variables. On retrouve alors le principe de l'analyse en facteurs (ou en facteurs communs et spécifiques ou factor analysis).

2.2 Estimation

Proposition IV.1 *L'estimation est fournie par l'A.C.P. de* $(\mathbf{Y}, \mathbf{M}, \mathbf{D})$ *c'est à dire par la décomposition en valeurs singulières de* $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ *:*

$$\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda_k^{1/2} u^k v^{k\prime} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}_q'.$$

Preuve

Sans hypothèse sur la distribution de l'erreur, une estimation par les moindres carrés conduit à résoudre le problème :

$$\min_{E_q, z_i} \left\{ \sum_{i=1}^{n} w_i \, \|y_i - z_i\|^2_{\mathbf{M}} \ ; \ \dim(E_q) = q, z_i - \bar{z} \in E_q \right\}. \qquad \text{(IV.2)}$$

On note $\mathbf{X} = \mathbf{Y} - \mathbf{1}_n \bar{y}'$ la matrice centrée où $\bar{y} = \sum_{i=1}^{n} w_i y_i$ et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(z_i - \bar{z})'$.

$$\sum_{i=1}^{n} w_i \, \|y_i - z_i\|^2_{\mathbf{M}} = \sum_{i=1}^{n} w_i \, \|y_i - \overline{y} + \overline{z} - z_i\|^2_{\mathbf{M}} + \|\overline{y} - \overline{z}\|^2_{\mathbf{M}} \ ;$$

le problème (IV.2) conduit alors à prendre $\widehat{\overline{z}} = \overline{y}$ et devient équivalent à résoudre :

$$\min_{\mathbf{Z}} \left\{ \|\mathbf{X} - \mathbf{Z}\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p}, \ \text{rang}(\mathbf{Z}) = q \right\}. \qquad \text{(IV.3)}$$

La fin de la preuve est une conséquence immédiate du théorème (A.5).

□

- Les u^k sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\mathbf{XMX'D}$ associés aux valeurs propres λ_k rangées par ordre décroissant.

- Les v_k, appelés *vecteurs principaux*, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\mathbf{X'DXM} = \mathbf{SM}$ associés aux mêmes valeurs propres ; ils engendrent des s.e.v. de dimension 1 appelés axes principaux.

Les estimations sont donc données par :

$$\begin{array}{rcl}
\widehat{\overline{z}} & = & \overline{y}, \\
\widehat{\mathbf{Z}_q} & = & \displaystyle\sum_{k=1}^{q} \lambda^{1/2} u^k v^{k'} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}'_q = \mathbf{X} \widehat{\mathbf{P}_q}', \\
\text{où } \widehat{\mathbf{P}_q} & = & \mathbf{V}_q \mathbf{V}'_q \mathbf{M} \text{ est la matrice de projection} \\
 & & \mathbf{M}\text{-orthogonale sur } \widehat{E_q}, \\
\widehat{E_q} & = & \text{vect}\{v^1, \ldots, v^q\}, \\
\widehat{E_2} & \text{est} & \text{appelé plan principal}, \\
\widehat{z_i} & = & \widehat{\mathbf{P}_q} x_i + \overline{y}.
\end{array}$$

Remarques IV.1 *:*

i. Les solutions sont emboîtées pour $q = 1, \ldots, p$:

$$E_1 = \ vect\{v^1\} \subset E_2 = \ vect\{v^1, v^2\} \subset E_3 = \ vect\{v^1, v^2, v^3\} \subset \ldots$$

ii. Les espaces principaux sont uniques sauf, éventuellement, dans le cas de valeurs propres multiples.

iii. Si les variables ne sont pas homogènes (unités de mesure différentes, variances disparates), elles sont préalablement réduites :

$$\widetilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Sigma}^{-1} \text{ où } \boldsymbol{\Sigma} = diag(\mathbf{S})^{1/2} ;$$

$\widetilde{\mathbf{S}}$ *est alors la matrice* $\mathbf{R} = \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}$ *des* corrélations.

Remarques IV.2 *Sous l'hypothèse que la distribution de l'erreur est gaussienne, une estimation par maximum de vraisemblance conduit à la même solution (cf. Anderson, 19xx).*

3 Point de vue des variables

On considère p variable statistiques *centrées* $X^1, \ldots, X^p$. Une *combinaison linéaire* de coefficients f_j de ces variables,

$$c = \sum_{j=1}^{p} f_j x^j = \mathbf{X}f,$$

définit une nouvelle variable centrée C qui, à tout individu i, associe la "mesure"

$$C(i) = x_i' f.$$

3.1 Définition équivalente

Proposition IV.2 *Soient p variables quantitatives centrées $X^1, \ldots, X^p$ observées sur n individus de poids w_i ; l'A.C.P. de $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ est aussi la recherche des q combinaisons linéaires normées des X^j, non corrélées et dont la somme des variances soit maximale.*

3.2 Propriétés

- Les vecteurs $f^k = \mathbf{M}v^k$ sont les *facteurs principaux*. Ils permettent de définir les combinaisons linéaires des X^j optimales au sens ci-dessus.
- Les vecteurs $c^k = \mathbf{X}f^k$ sont les *composantes principales*.
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les *variables principales* ;

$$\begin{aligned} \mathrm{cov}(C^k, C^l) &= (\mathbf{X}f^k)'\mathbf{D}\mathbf{X}f^l = f^{k\prime}\mathbf{S}f^l = \\ &= v^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}v^l = \lambda_l v^{k\prime}\mathbf{M}v^l = \lambda_l \delta_k^l. \end{aligned}$$

- Les f^k sont les vecteurs propres $\mathbf{M}^{-1}$-orthonormés de la matrice $\mathbf{MS}$.

– La matrice
$$\mathbf{C} = \mathbf{XF} = \mathbf{XMV} = \mathbf{U}\mathbf{\Lambda}^{1/2}$$
est la matrice des composantes principales.

– Les axes définis par les vecteurs $\mathbf{D}$-orthonormés u^k sont appelés *axes factoriels*.

4 Représentations graphiques

4.1 Les individus

Les graphiques obtenus permettent de représenter "au mieux" les distances euclidiennes inter-individus mesurées par la métrique $\mathbf{M}$.

Projection

Chaque individu ω_i représenté par x_i est approché par sa projection $\mathbf{M}$-orthogonale $\widehat{z}_i^q$ sur le sous-espace $\widehat{E_q}$ engendré par les q premiers vecteurs principaux $\{v^1, \ldots, v^q\}$. En notant e_i un vecteur de la base canonique de E, la coordonnée de l'individu i sur v^k est donnée par :
$$\left\langle x_i, v^k \right\rangle_{\mathbf{M}} = x_i' \mathbf{M} v^k = e_i' \mathbf{XM} v^k = c_i^k.$$

Proposition IV.3 *Les coordonnées de la projection* $\mathbf{M}$*-orthogonale de* x_i *sur* $\widehat{E_q}$ *sont les* q *premiers élément de la* $i^{ème}$ *ligne de la matrice* $\mathbf{C}$ *des composantes principales.*

Mesures de "qualité"

La "qualité globale" des représentations est mesurée par la *part de dispersion expliquée* :
$$r_q = \frac{\mathrm{tr}\mathbf{SM}\widehat{\mathbf{P}_q}}{\mathrm{tr}\mathbf{SM}} = \frac{\sum_{k=1}^q \lambda_k}{\sum_{k=1}^p \lambda_k}.$$
La qualité de la représentation de chaque x_i est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :
$$[\cos\theta(x_i, \widehat{z}_i^q)]^2 = \frac{\left\| \widehat{\mathbf{P}_q} x_i \right\|_{\mathbf{M}}^2}{\|x_i\|_{\mathbf{M}}^2} = \frac{\sum_{k=1}^q (c_i^k)^2}{\sum_{k=1}^p (c_i^k)^2}.$$

Pour éviter de consulter un tableau qui risque d'être volumineux (n lignes), les étiquettes de chaque individu sont affichées sur les graphiques avec des caractères dont la *taille est fonction de la qualité*. Un individu très mal représenté est à la limite de la lisibilité.

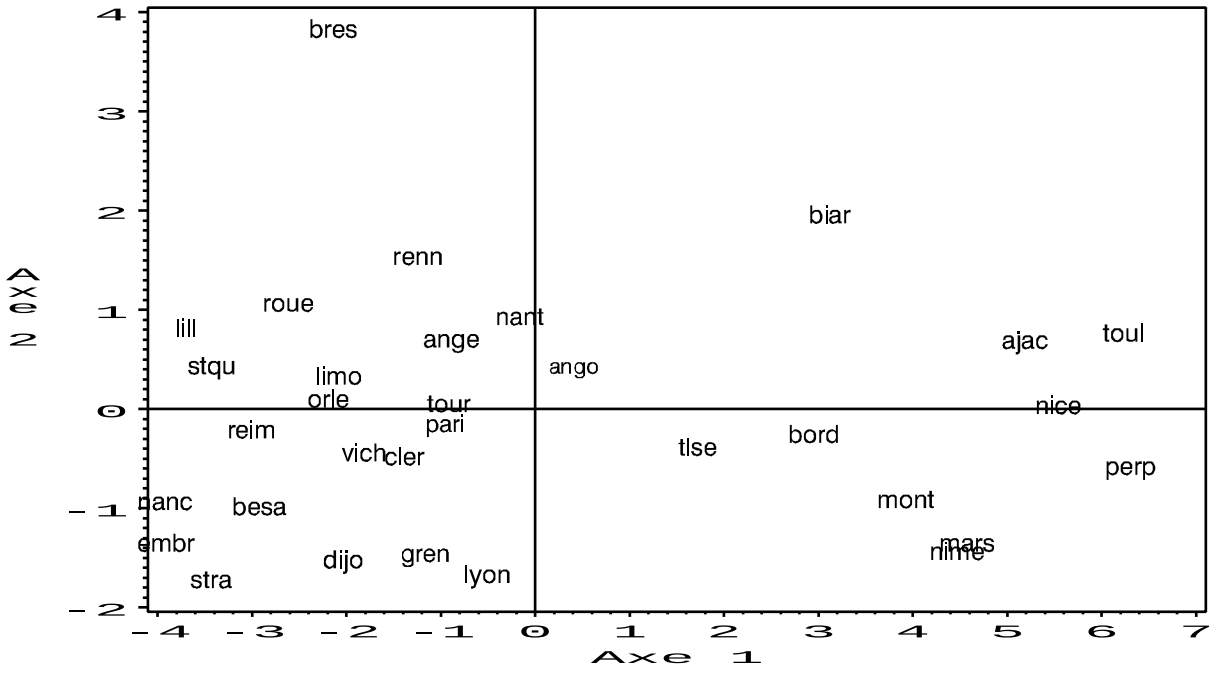


FIG. IV.1 - Températures : premier plan des individus.

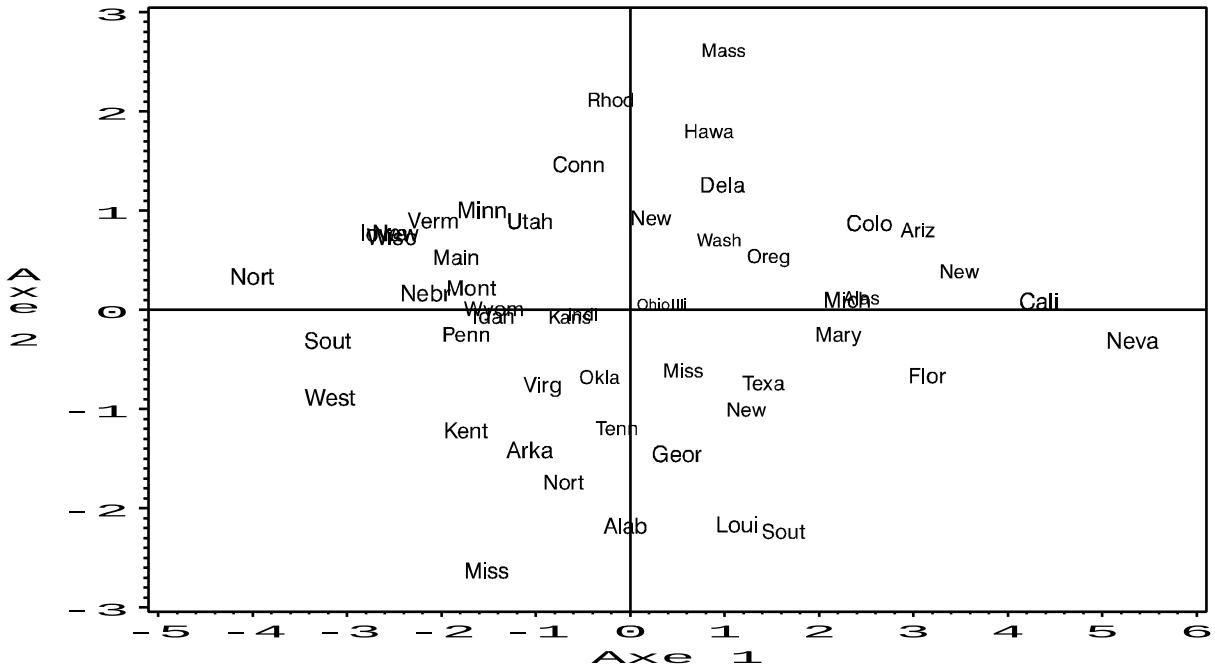


FIG. IV.2 - Criminalité : premier plan des individus.

Contributions

Les contributions de chaque individu à la dispersion globale

$$\gamma_i = \frac{w_i \left\| x_i \right\|^2_{\mathbf{M}}}{\mathrm{tr}\mathbf{SM}} = \frac{w_i \sum_{k=1}^{p} (c_i^k)^2}{\sum_{k=1}^{p} \lambda_k},$$

ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{w_i (c_i^k)^2}{\lambda_k},$$

permettent de déceler les observations les plus *influentes* et, éventuellement, aberrantes. Ces points apparaissent visiblement lors du tracé des boîtes-à-moustaches parallèles des composantes principales qui évitent ainsi une lecture fastidieuse de ce tableau des contributions. En effet, ils se singularisent aussi comme "outliers" hors de la boîte (au delà des moustaches) correspondant à une direction principale. Les individus correspondants, considérés comme *individus supplémentaires*, peuvent être éliminés lors d'une nouvelle analyse.

Individus supplémentaires

Il s'agit de représenter, par rapport aux axes principaux d'une analyse, des individus qui n'ont pas participé aux calculs de ces axes. Soit s un tel vecteur, il doit être centré, éventuellement réduit, puis projeté sur le sous-espace de représentation. Les coordonnées sont fournies par :

$$\left\langle v^k, \mathbf{V}_q \mathbf{V}_q' \mathbf{M}(s - \bar{y}) \right\rangle_{\mathbf{M}} = v^{k\prime} \mathbf{M} \mathbf{V}_q \mathbf{V}_q' \mathbf{M}(s - \bar{y}) = e^{k\prime} \mathbf{V}_q' \mathbf{M}(s - \bar{y}).$$

Les coordonnées d'un individu supplémentaire dans la base des vecteurs principaux sont donc :

$$\mathbf{V}_q' \mathbf{M}(s - \bar{y}).$$

4.2 Les variables

Les graphiques obtenus permettent de représenter "au mieux" les corrélations entre les variables (cosinus des angles) et, si celles-ci ne sont pas réduites, leurs variances (longueurs).

Projection

Une variable X^j (ou Y^j) est représentée par la projection $\mathbf{D}$-orthogonale $\widehat{\mathbf{Q}_q x^j}$ sur le sous-espace F_q engendré par les q premiers axes factoriels. La coordonnée de x^j sur u^k est :

$$\left\langle x^j, u^k \right\rangle_{\mathbf{D}} = x^{j\prime} \mathbf{D} u^k = \frac{1}{\sqrt{\lambda_k}} x^{j\prime} \mathbf{D} \mathbf{X} \mathbf{M} v^k = \frac{1}{\sqrt{\lambda_k}} e^{j\prime} \mathbf{X}' \mathbf{D} \mathbf{X} \mathbf{M} v^k = \sqrt{\lambda_k} v_j^k.$$

Proposition IV.4 *Les coordonnées de la projection* $\mathbf{D}$*-orthogonale de* x^j *sur le sous-espace* F_q *sont les* q *premiers éléments de la* $j^{ème}$ *ligne de la matrice* $\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Mesure de "qualité"

La qualité de la représentation de chaque x^j est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$\left[\cos\theta(x^j, \widehat{\mathbf{Q}_q x^j})\right]^2 = \frac{\left\|\widehat{\mathbf{Q}_q x^j}\right\|^2_{\mathbf{D}}}{\|x^j\|^2_{\mathbf{D}}} = \frac{\sum_{k=1}^q \lambda_k (v_k^j)^2}{\sum_{k=1}^p \lambda_k (v_k^j)^2}.$$

Corrélations variables×facteurs

Ces indicateurs aident à l'interprétation des axes factoriels en exprimant les corrélations entre variables principales et initiales.

$$\mathrm{cor}(X^j, C^k) = \cos\theta(x^j, c^k) = \cos\theta(x^j, u^k) = \frac{\left\langle x^j, u^k \right\rangle_{\mathbf{D}}}{\|x^j\|_{\mathbf{D}}} = \frac{\sqrt{\lambda_k}}{\sigma_j} v_j^k \ ;$$

ce sont les éléments de la matrice $\mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Cercle des corrélations

Dans le cas de variables réduites $\tilde{x}^j = \sigma_j^{-1} x^j, \|\tilde{x}^j\|_{\mathbf{D}} = 1$, les $\tilde{x}^j$ sont sur la sphère unité $\mathcal{S}_n$ de F. L'intersection $\mathcal{S}_n \cap F_2$ est un cercle centré sur l'origine et de rayon 1 appelé *cercle des corrélations*. Les projections de $\tilde{x}^j$ et x^j sont colinéaires, celle de $\tilde{x}^j$ étant à l'intérieur du cercle :

$$\left\|\widehat{\mathbf{Q}_2 \tilde{x}^j}\right\|_{\mathbf{D}} = \cos\theta(x^j, \widehat{\mathbf{Q}_2 x_j}) \leq 1.$$

Ainsi, plus $\widehat{\mathbf{Q}_2 \tilde{x}^j}$ est proche de ce cercle, meilleure est la qualité de sa représentation. Ce graphique est commode à interpréter à condition de se méfier des échelles, le cercle devenant une ellipse si elles ne sont pas égales. Comme pour les individus, la taille des caractères est aussi fonction de la qualité des représentations.

4.3 Représentation simultanée ou "biplot"

A partir de la décomposition en valeurs singulières de $(\mathbf{X}, \mathbf{M}, \mathbf{D})$, on remarque que chaque valeur

$$x_i^j = \sum_{k=1}^p \sqrt{\lambda_k} u_i^k v_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$$

s'exprime comme produit scalaire usuel des vecteurs

$$c_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } v^j \text{ ou encore } u_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j .$$

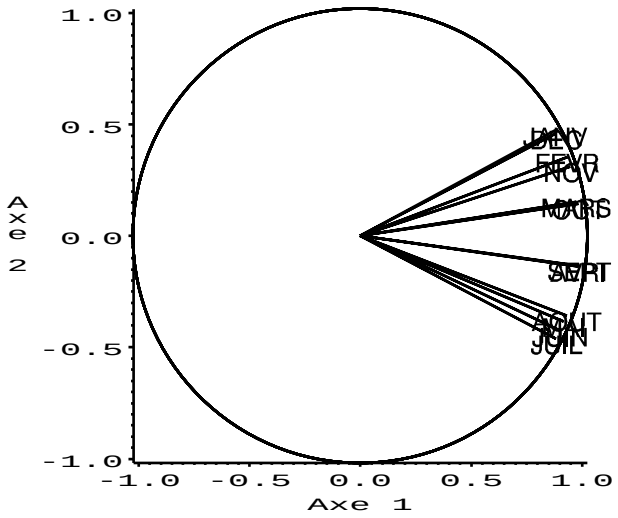


FIG. IV.3 - Températures : premier plan des variables.

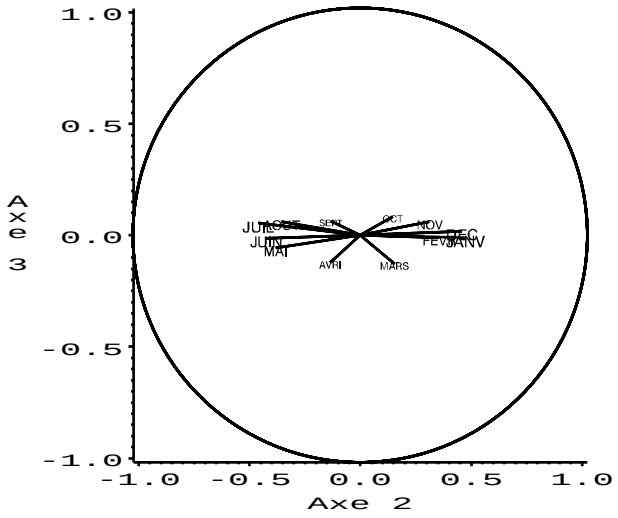


FIG. IV.4 - Températures : deuxième plan des variables.

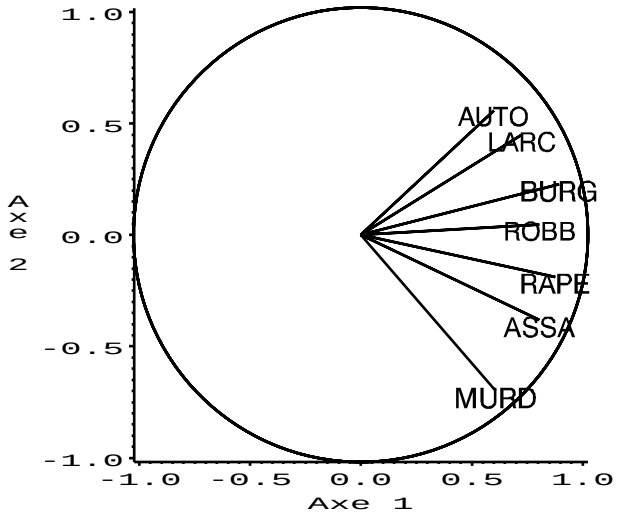


FIG. IV.5 - Criminalité : premier plan des variables.

Pour $q = 2$, la quantité $\widehat{z}_i^j$ en est une approximation limitée aux deux premiers termes.

Cette remarque permet d'interpréter deux autres représentations graphiques en A.C.P. projetant *simultanément* individus et variables.

i. la représentation *isométrique ligne* utilise les matrices $\mathbf{C}$ et $\mathbf{V}$; elle permet d'interpréter les distances entre individus ainsi que les produits scalaires entre un individu et une variable qui sont, dans le premier plan principal, des approximations des valeurs observées $x_i^j = X^j(i)$;

ii. la représentation *isométrique colonne* utilise les matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$; elle permet d'interpréter les angles entre vecteurs variables (corrélations) et les produits scalaires comme précédemment.

Remarques IV.3 .

i. Dans le cas fréquent où $\mathbf{M} = \mathbf{I}_p$ et où les variables sont réduites, le point représentant X^j, en superposition dans l'espace des individus se confond avec un pseudo individu supplémentaire qui prendrait la valeur 1 (écart-type) pour la variable j et 0 pour les autres.

ii. En pratique, ces différents types de représentations (simultanées ou non) ne diffèrent que par un changement d'échelle sur les axes ; elles sont très voisines et suscitent souvent les mêmes interprétations.

5 Choix de dimension

La qualité des estimations auxquelles conduit l'A.C.P. dépend directement du choix de q, c'est-à-dire du nombre de composantes retenues pour reconstituer les données ou encore de la dimension du sous-espace de représentation. De nombreux critères de choix ont été proposés dans la littérature dont Jolliffe (1986) propose une revue.

Certains critères, non explicités ici, s'inspirent des pratiques statistiques décisionnelles ; sous l'hypothèse que l'erreur admet une distribution *gaussienne*, on peut exhiber les lois asymptotiques des valeurs propres et donc construire des tests de nullité ou d'égalité de ces dernières. Malheureusement, outre la nécessaire hypothèse de normalité, ceci conduit à une procédure de tests emboîtés dont le niveau global est incontrôlable. Nous nous intéressons dans cette section aux autres critères qui, dans ce cadre descriptif, ne nécessitent pas d'hypothèse sur la loi de l'erreur.

La présentation de l'A.C.P., comme résultat de l'estimation d'un modèle, permet d'apporter des éléments de réponse plus satisfaisants. La qualité des estimations est évaluée de façon habituelle en statistique par un risque moyen quadratique définissant un critère de stabilité du sous-espace de représentation.

5.1 Part de variance

La "qualité globale" des représentations est mesurée par la *part de variance expliquée* :

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

La valeur de q est choisie de sorte que cette part de variance expliquée r_q soit supérieure à une valeur seuil fixée a priori par l'utilisateur. C'est souvent le seul critère employé.

5.2 Règle de Kaiser

On considère que, si tous les éléments de Y sont indépendants, les composantes principales sont toutes de variances égales (égales à 1 dans le cas de l'A.C.P. réduite). On ne conserve alors que les valeurs propres supérieures à leur moyenne car seules jugées plus "informatives" que les variables initiales ; dans le cas d'une A.C.P. réduite, ne sont donc retenues que celles plus grandes que 1. Ce critère, utilisé implicitement par SAS/ASSIST, a tendance à surestimer le nombre de composantes pertinentes.

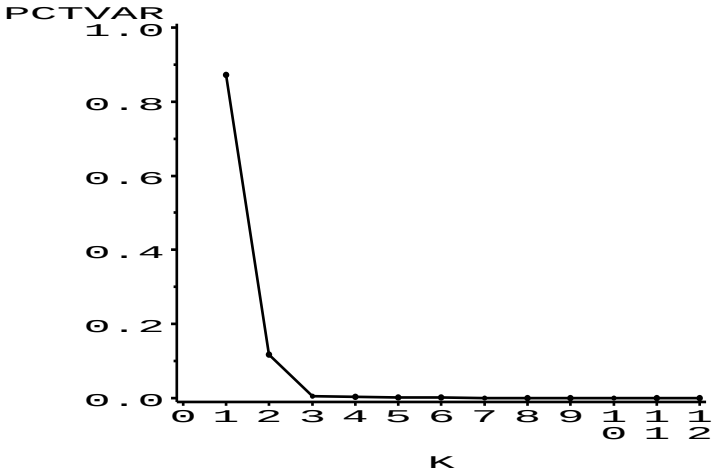


FIG. IV.6 - Températures : éboulis des valeurs propres.

5.3 Éboulis des valeurs propres

C'est le graphique (figures 5.3 et 5.3) présentant la décroissance des valeurs propres. Le principe consiste à rechercher, s'il existe, un "coude" (changement de signe dans la suite des différences d'ordre 2) dans le graphe et de ne conserver que les valeurs propres jusqu'à ce coude. Intuitivement, plus l'écart $(\lambda_q - \lambda_{q+1})$ est significativement grand, par exemple supérieur à $(\lambda_{q-1} - \lambda_q)$, et plus on peut être assuré de la stabilité de $\widehat{E_q}$.

5.4 Boîtes-à-moustaches des variables principales

Un graphique (figure 5.4 et 5.4) présentant, en parallèle, les boîtes-à-moustaches des variables principales illustre bien leurs qualités : stabilité lorsqu'une grande boîte est associée à de petites moustaches, instabilité en présence d'une petite boîte, de grandes moustaches et de points isolés. Intuitivement, on conserve les premières "grandes boîtes". Les points isolés ou "outliers" désignent les points à forte contribution ou influents dans une direction principale.

5.5 Critère de stabilité

L'A.C.P. ayant été introduite comme résultat de l'estimation d'un modèle (IV.1), il est possible de discuter de la qualité de cette estimation. Comme il est d'usage en statistique, la mesure de qualité considérée sera défini comme

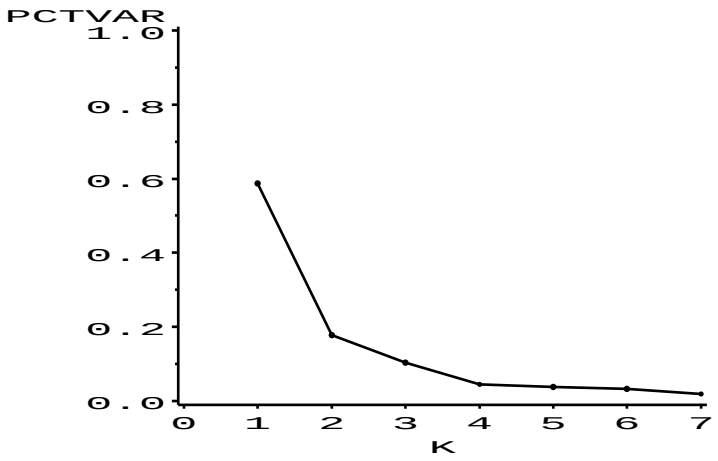


FIG. IV.7 - Criminalité : éboulis des valeurs propres.

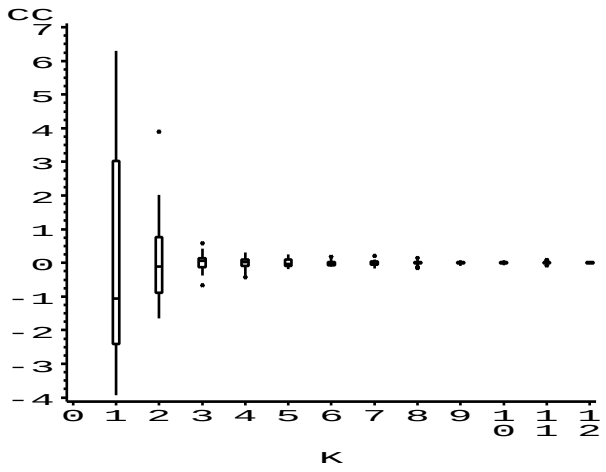


FIG. IV.8 - Températures : composantes en boîtes.

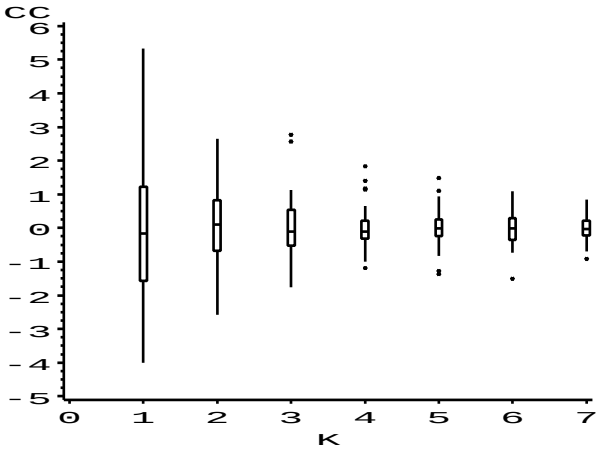


FIG. IV.9 - Criminalité : composantes en boîtes.

un risque quadratique moyen (ou Mean Square Error) défini comme l'espérance d'une distance entre le modèle "vrai" et l'estimation qui en est faite. Besse (1992) propose d'étudier la qualité de l'estimation du sous-espace de représentation $\widehat{E_q}$. On considère la fonction perte :

$$L_q = Q(E_q, \widehat{E_q}) = \frac{1}{2} \left\| \mathbf{P}_q - \widehat{\mathbf{P}_q} \right\|^2_{\mathbf{M},\mathbf{D}} = q - \mathrm{tr}\mathbf{P}_q\widehat{\mathbf{P}_q}, \qquad \text{(IV.4)}$$

où Q mesure la distance entre deux sous-espaces par la distance usuelle entre les matrices de projection qui leur sont associées. C'est aussi la somme des carrés des coefficients de corrélation canonique entre les ensembles de composantes ou de variables principales qui engendrent respectivement E_q et son estimation $\widehat{E_q}$.

Un risque moyen quadratique est alors défini en prenant l'espérance de la fonction perte :

$$R_q = EQ(E_q, \widehat{E_q}). \qquad \text{(IV.5)}$$

Sans hypothèse sur la distribution de l'erreur, seules des techniques de ré-échantillonnage permettent de fournir une estimation de ce risque moyen quadratique. Leur emploi est justifié car le risque est invariant par permutation des observations. On se pose donc la question de savoir pour quelles valeurs de q les représentations graphiques sont fiables, c'est-à-dire stables pour des fluctuations de l'échantillon. Différentes estimations ont été comparées (bootstrap, jackknife) par Besse et Falguerolles (1993) mais, comme elles sont très coûteuses en temps de calcul, Besse (1992) propose d'utiliser une approximation de l'estimateur par

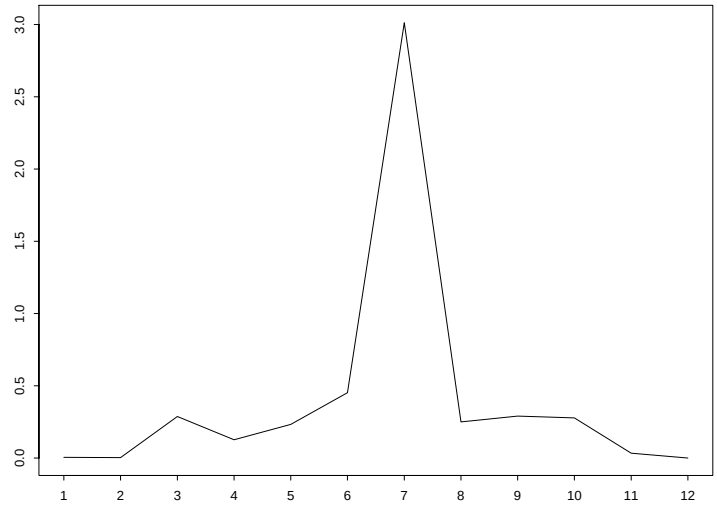


FIG. IV.10 - Températures : stabilité des sous-espaces.

jackknife ; elle fournit, directement à partir des résultats de l'A.C.P. (valeurs propres et composantes principales), une estimation satisfaisante du risque :

$$\widehat{R_{JKq}} = \widehat{R_{\mathbf{P}q}} + O((n-1)^{-2}).$$

$\widehat{R_{\mathbf{P}q}}$ est une approximation analytique de l'estimateur jackknife qui a pour expression :

$$\widehat{R_{\mathbf{P}q}} = \frac{1}{n-1} \sum_{k=1}^{q} \sum_{j=q+1}^{p} \frac{\frac{1}{n}\sum_{i=1}^{n} (c_i^k)^2 (c_i^j)^2}{(\lambda_j - \lambda_k)^2} \qquad \text{(IV.6)}$$

où c_i^j désigne le terme général de la matrice des composantes principales $\mathbf{C}$.

Ce résultat souligne l'importance du rôle que joue l'écart $(\lambda_q - \lambda_{q+1})$ dans la stabilité du sous-espace de représentation. Le développement est inchangé dans le cas d'une A.C.P. réduite ; de plus, il est valide tant que

$$n > \frac{\|\mathbf{S}\|_2^2}{\inf\{(\lambda_k - \lambda_{k+1}); k = 1, \ldots, q\}}.$$

Les figures 5.5 et 5.5 montrent la stabilité du sous-espace de représentation en fonction de la dimension q pour l'A.C.P. des données de criminalité. Comme souvent, le premier axe est très stable tandis que le premier plan (températures) et le premier sous-espace de dimension 3 (criminalité) restent fiables. Au delà, les axes étant très sensibles à toute perturbation des données, ils peuvent être associés à du bruit. Ces résultats sont cohérents avec les deux critères graphiques précédents mais souvent, en pratique, le critère de stabilité conduit à un choix de dimension plus explicite.

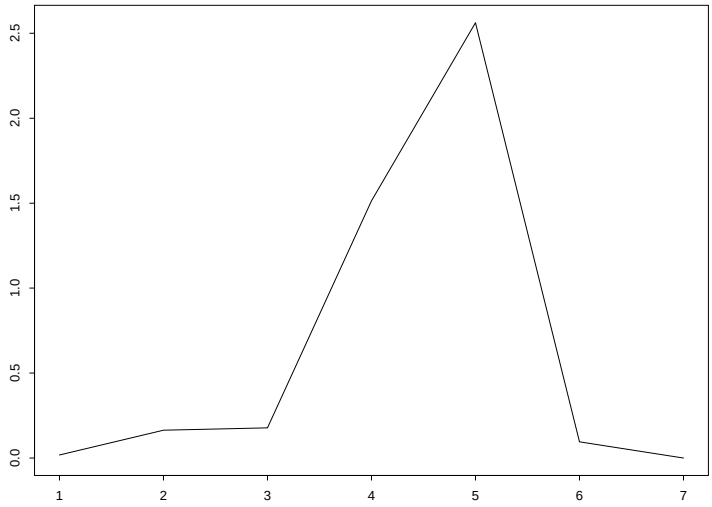


FIG. IV.11 - Criminalité : stabilité des sous-espaces.

6 Pratique de l'A.C.P.

6.1 Préliminaires

Les données se présentent sous la forme d'un fichier dont chaque ligne ou article est découpée en rubriques :

identificateur — var1 —. . . — varp

et nécessite un traitement préalable à l'exécution d'un programme d'A.C.P. afin de :

- vérifier la cohérence et l'exactitude des données,
- éliminer certaines variables,
- procéder à d'éventuelles transformations de variables (racine, log. . .).

Il faut donc considérer successivement et répétitivement les opérations suivantes :

i. formattage, nettoyage du fichier, traitement des données manquantes,

ii. choix des variables (se méfier de l'effet taille),

iii. étude univariée (moyennes, médianes, indices de dispersion, histogrammes, boîtes-à-moustaches. . .),

iv. étude bivariée (matrice des corrélations, des nuages de points...),

v. itérer...

On obtient alors la matrice $\mathbf{Y}_{(n\times p)}$ qui sera centrée par le programme.

6.2 Options

Différents choix sont offerts à l'utilisateur :

- réduction des variables (par défaut) lorsqu'elles ne sont pas homogènes (unités de mesure différentes ou variances disparates),
- pondération des individus (par défaut $\frac{1}{n}$) pour regrouper des données identiques, redresser un échantillon...
- métrique de l'espace des individus : par défaut $\mathbf{M} = \mathbf{I}_p$; pour pondérer les variables : $\mathbf{M} =$ diag$(a_1^2, \ldots, a_p^2)$.

6.3 Simplification de métrique

Une métrique étant donnée dans l'espace E des individus, la matrice $\mathbf{M}$ associée est symétrique, définie-positive ; elle se fractionne en $\mathbf{M} = \mathbf{F}\mathbf{F}'$ où $\mathbf{F}$ $(p \times p)$ est de rang p. Soit $\mathbf{Y}$ un tableau de données et f un vecteur de $\mathbb{R}^p$ alors la combinaison linéaire $\mathbf{Y}f$ des colonnes y^j définit une nouvelle variable. Globalement, le produit $\mathbf{YF}$ construit un nouveau tableau obtenu par transformations linéaires des variables initiales. Les variables transformées ont pour moyenne $\mathbf{F}\overline{y}$ et pour covariance $\mathbf{F}'\mathbf{S}\mathbf{F}$.

Proposition IV.5 *Les A.C.P. de* $(\mathbf{YF}, \mathbf{I}_p, \mathbf{D})$ *et* $(\mathbf{Y}, \mathbf{M}, \mathbf{D})$ *sont équivalentes au sens où elles conduisent aux mêmes matrices des composantes principales.*

Ce résultat élémentaire permet de calculer avec un logiciel usuel une A.C.P. pour une métrique quelconque. Il suffit de pré-traiter les données avant de calculer l'a.c.p. de $(\mathbf{Y}\mathbf{M}^{1/2}, \mathbf{I}_p, \mathbf{D})$

6.4 Interprétation

Les macros SAS décrites en annexe, de même que la plupart des logiciels, proposent, ou autorisent, l'édition des différents indicateurs (contributions, qualités, corrélations) et graphiques définis dans les paragraphes précédents.

- Les *contributions* permettent d'identifier les individus très influents pouvant déterminer à eux-seuls l'orientation de certains axes ; ces points sont vérifiés, caractérisés, puis éventuellement considérés comme *supplémentaires* dans une autre analyse.

- Il faut choisir le nombre de composantes à retenir, c'est-à-dire la dimension des espaces de représentation.
- Les axes factoriels sont interprétés par rapport aux variables initiales bien représentées.
- Les graphiques des individus sont interprétés, en tenant compte des qualités de représentation, en termes de regroupement ou dispersions par rapport aux axes factoriels et projections des variables initiales.

Les quelques graphiques présentés suffisent, dans la plupart des cas, à l'interprétation d'une A.C.P. classique et évitent la sortie volumineuse, lorsque n est grand, des tableaux usuels d'aide à l'interprétation. Ceux-ci sont reportés en annexe. On échappe ainsi à une critique fréquente, et souvent justifiée, des anglo-saxons vis-à-vis de la pratique française de "l'analyse des données" qui, paradoxalement, cherche à "résumer au mieux l'information" mais produit plus de chiffres en sortie qu'il n'y en a en entrée !

Remarques IV.4 *L'A.C.P. est une technique* linéaire *optimisant un critère* quadratique*; elle ne tient donc pas compte d'éventuelles liaisons non linéaires et présente une forte sensibilité aux valeurs extrêmes.*

Chapitre V

Analyse Factorielle Discriminante

1 Métrique optimale en A.C.P.

L'A.C.P. ayant été introduite comme résultat de l'estimation d'un modèle (IV.1), il est alors possible de discuter de la qualité de cette estimation. Comme il est d'usage en statistique, la mesure de qualité considérée est définie comme un risque quadratique moyen (ou Mean Square Error) défini comme l'espérance d'une distance entre le modèle "vrai" et l'estimation qui en est faite. Appliquée au choix de la métrique, cette démarche permet de montrer en quoi une Analyse Factorielle Discriminante (A.F.D.) est une A.C.P. optimale.

1.1 Critère

On considère pour fonction perte, la moyenne (les individus sont de même poids pour simplifier) des carrés des distances entre les effets fixes z_i et leurs estimations :

$$L_q = \frac{1}{n} \sum_{i=1}^{n} \|z_i - \hat{z}_i\|^2_{\mathbf{A}} ,$$

où $\mathbf{A}$ désigne une métrique euclidienne quelconque de $\mathbb{R}^p$.

L'espérance de cette quantité définit un risque quadratique moyen noté :

$$R_q(\mathbf{M}, \mathbf{A}) = E(L_q)$$

dont la minimisation caractérise un choix optimal pour la métrique $\mathbf{M}$ de l'espace des individus.

1.2 Approximation

Seule une approximation par linéarisation permet de tenir compte du critère ainsi défini. En utilisant la théorie des Perturbations, Besse et al. (1988) ont

montré la

Proposition V.1 *En supposant σ^2 "petit", $R_q(\mathbf{M},\mathbf{A})$ admet le développement :*

$$R_q(\mathbf{M},\mathbf{A}) = \sigma^2\left[(q+1)tr\mathbf{\Gamma A} + (n-q-1)tr(\widehat{\mathbf{P}}_q\mathbf{\Gamma P}_q'\mathbf{A})\right] + 0(\sigma_3).$$

Les notations sont celles du chapitre IV.

1.3 Optimisation

Il est alors possible de montrer le

Théorème V.2 *Pour toute métrique $\mathbf{A}$, la partie principale de $R_q(\mathbf{M},\mathbf{A})$ est minimale pour $\mathbf{M} = \mathbf{\Gamma}^{-1}$.*

Ce théorème fournit, dans le cadre de l'ACP, une propriété de type Gauss-Markov : la métrique optimale est donnée par l'inverse de la matrice de covariance de l'erreur.

Cette matrice étant généralement inconnue, la métrique euclidienne usuelle $\mathbf{I}_p$ est naturellement choisie dans l'espace des individus. Mais, dans certaines situations, $\mathbf{\Gamma}$ et son inverse peuvent être estimées. C'est le cas en analyse discriminante lorsqu'on l'introduit comme un type particulier de modèle à effet fixe.

2 Introduction à l'A.F.D.

2.1 Données

Les données sont constituées de

- p variables *quantitatives* $Y^1,\ldots,Y^p$ jouant le rôle de variables explicatives comme dans le modèle linéaire,
- une variable *qualitative* T, à m modalités $\{\mathcal{T}_1,\ldots,\mathcal{T}_m\}$, jouant le rôle de variable à expliquer.

La situation est analogue à celle de la régression linéaire multiple mais, comme la variable à expliquer est qualitative, on aboutit à une méthode très différente. Les variables sont observées sur l'ensemble Ω des n individus affectés des poids $w_i > 0$, $(\sum_{i=1}^n w_i = 1)$, et l'on pose

$$\mathbf{D} = \text{ diag}(w_i\ ; i = 1,\ldots,n).$$

La variable T engendre une partition $\{\Omega_k\ ; k = 1,\ldots,m\}$ de l'ensemble Ω des individus dont chaque élément est d'effectif n_k.

On note $\mathbf{T}$ $(n \times m)$ la matrice des indicatrices des modalités de la variable T ; son terme général est

$$t_i^k = t^k(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_k \\ 0 & \text{sinon} \end{cases} .$$

En posant

$$\overline{w_k} = \sum_{i \in \Omega_k} w_i,$$

il vient

$$\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{ diag}(\overline{w_1}, \ldots, \overline{w_m}).$$

2.2 Objectifs

Deux techniques cohabitent sous la même appellation d'analyse discriminante :

descriptive : cette méthode recherche, parmi toutes les A.C.P. possibles sur les variables Y^j, celle dont les représentations graphiques des individus *discriminent* "au mieux" les m classes engendrées par la variable T (e.g. recherche de facteurs de risque en statistique médicale) ;

décisionnelle : connaissant, pour un individu donné, les valeurs des Y^j mais pas la modalité de T, cette méthode consiste à affecter cet individu à une modalité (e.g. reconnaissance de formes).

Remarques V.1 *Lorsque le nombre et les caractéristiques des classes sont connues, il s'agit d'une* discrimination *; sinon, on parle de* classification *ou encore, avec des hypothèses sur les distributions, de* reconnaissance de mélanges.

2.3 Notations

On note $\mathbf{Y}$ la matrice $(n \times p)$ des données quantitatives, $\overline{\mathbf{Y}}$ la matrice $(m \times p)$ des barycentres des classes :

$$\overline{\mathbf{Y}} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{Y} = \begin{bmatrix} \overline{y_1}' \\ \vdots \\ \overline{y_m}' \end{bmatrix} \text{ où } \overline{y_k} = \frac{1}{\overline{w_k}} \sum_{i \in \Omega_k} w_i y_i,$$

et $\mathbf{Y}_e$ la matrice $(n \times p)$ dont la ligne i est le barycentre $\overline{y_k}$ de la classe Ω_k à laquelle appartient l'individu i :

$$\mathbf{Y}_e = \mathbf{T}\overline{\mathbf{Y}} = \mathbf{P}\mathbf{Y} \ ;$$

$\mathbf{P} = \mathbf{T}\overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}$ est la matrice de projection $\mathbf{D}$-orthogonale sur le sous-espace engendré par les indicatrices de T ; c'est l'espérance conditionnelle sachant T.

Deux matrices "centrées" sont définies de sorte que $\mathbf{X}$ se décompose en

$$\mathbf{X} = \mathbf{X}_r + \mathbf{X}_e$$

avec

$$\mathbf{X}_r = \mathbf{Y} - \mathbf{Y}_e \text{ et } \mathbf{X}_e = \mathbf{Y}_e - \mathbf{1}_n \overline{y}'.$$

On note également $\overline{\mathbf{X}}$ la matrice centrée des barycentres :

$$\overline{\mathbf{X}} = \overline{\mathbf{Y}} - \mathbf{1}_m \overline{y}'.$$

On appelle alors variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \mathbf{X}_r' \mathbf{D} \mathbf{X}_r = \sum_{k=1}^{m} \sum_{i \in \Omega_k} w_i (y_i - \overline{y}_k)(y_i - \overline{y}_k)',$$

et variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{X}}' \overline{\mathbf{D}\mathbf{X}} = \mathbf{X}_e' \mathbf{D} \mathbf{X}_e = \sum_{k=1}^{m} \overline{w_k} (\overline{y}_k - \overline{y})(\overline{y}_k - \overline{y})'.$$

Proposition V.3 *La matrice des covariances se décompose en*

$$\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r.$$

3 Définition

3.1 Modèle

Dans l'espace des individus, le principe consiste à projeter les individus dans une direction permettant de mettre en évidence les groupes. À cette fin, Il faut privilégier la variance interclasse au détriment de la variance intraclasse considérée comme due au bruit.

En A.C.P., pour chaque effet z_i à estimer, on ne dispose que d'une observation y_i ; dans le cas de l'A.F.D. on considère que les éléments d'une même classe Ω_k sont les observations répétées n_k fois du même effet z_k pondéré par $\overline{w_k} = \sum_{i \in \Omega_k} w_i$. Le modèle devient donc :

$$\begin{array}{l} \{y_i \ ; i = 1, \ldots, n\}, \ n \text{ vecteurs indépendants de } E, \\ \forall k, \forall i \in \Omega_k, y_i = z_k + \varepsilon_i \text{ avec } \left\{ \begin{array}{l} E(\varepsilon_i) = 0, \ \mathrm{var}(\varepsilon_i) = \mathbf{\Gamma}, \\ \mathbf{\Gamma} \text{ régulière et inconnue,} \end{array} \right. \\ \exists A_q, \text{ sous-espace affine de de dimension } q \text{ de } E \text{ tel que} \\ \forall k, z_k \in A_q, \ (q < \min(p, m-1)). \end{array} \tag{V.1}$$

Remarques V.2 *Soit $\overline{z} = \sum_{k=1}^{m} \overline{w_k} z_k$. Le modèle entraîne que $\bar{z} \in A_q$. Soit E_q le sous-espace de dimension q de E tel que $A_q = \bar{z} + E_q$. Les paramètres à estimer sont E_q et $\{z_k \ ; k = 1, \ldots, m\}$; $\overline{w_k}$ est un paramètre de nuisance qui ne sera pas considéré.*

3.2 Estimation

L'estimation par les moindres carrés s'écrit ainsi :

$$\min_{E_q, z_k} \left\{ \sum_{k=1}^{m} \sum_{i \in \Omega_k} w_i \, \|y_i - z_k\|^2_{\mathbf{M}} \ ; \ \dim(E_q) = q, z_k - \bar{z} \in E_q \right\}.$$

Comme on a

$$\sum_{k=1}^{m} \sum_{i \in \Omega_k} w_i \, \|y_i - z_k\|^2_{\mathbf{M}} = \sum_{k=1}^{m} \sum_{i \in \Omega_k} w_i \, \|y_i - \overline{y_k}\|^2_{\mathbf{M}} + \sum_{k=1}^{m} \overline{w_k} \, \|\overline{y_k} - z_k\|^2_{\mathbf{M}},$$

on est conduit à résoudre :

$$\min_{E_q, z_k} \left\{ \sum_{k=1}^{m} \overline{w_k} \, \|\overline{y_k} - z_k\|^2_{\mathbf{M}} \ ; \ \dim(E_q) = q, z_k - \bar{z} \in E_q \right\}.$$

La covariance $\sigma^2 \mathbf{\Gamma}$ du modèle (V.1) étant inconnue, il faut l'estimée. Ce modèle stipule que l'ensemble des observations d'une même classe Ω_l suit une loi (inconnue) de moyenne z_l et de variance $\mathbf{\Gamma}$. Dans ce cas particulier, la matrice de covariances intraclasse ou matrice des covariances résiduelles empiriques $\mathbf{S}_r$ fournit donc une estimation "optimale" de la métrique de référence :

$$\mathbf{M} = \hat{\mathbf{\Gamma}}^{-1} = \mathbf{S}_r^{-1}$$

.

Proposition V.4 *L'estimation des paramètres E_q et z_k du modèle V.1 est obtenue par l'A.C.P. de $(\overline{\mathbf{Y}}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$. C'est l'Analyse Factorielle Discriminante (A.F.D.) de $(\mathbf{Y}|\mathbf{T}, \mathbf{D})$.*

4 Réalisation de l'A.F.D.

Les expressions matricielles définissant les représentations graphiques et les aides à l'interprétation découlent de celles de l'A.C.P..

4.1 Matrice à diagonaliser

L'A.C.P. de $(\overline{\mathbf{Y}}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à l'analyse spectrale de la matrice positive $\mathbf{S}_r^{-1}$-symétrique :

$$\overline{\mathbf{X}}' \overline{\mathbf{D}} \overline{\mathbf{X}} \mathbf{S}_r^{-1} = \mathbf{S}_e \mathbf{S}_r^{-1}.$$

Comme $\mathbf{S}_r^{-1}$ est régulière, cette matrice est de même rang que $\mathbf{S}_e$ et donc de même rang que $\overline{\mathbf{Y}}$ qui est de dimension $(m \times p)$. Les données étant centrées lors de l'analyse, le rang de la matrice à diagonaliser est

$$h = \mathrm{rang}(\mathbf{S}_e \mathbf{S}_r^{-1}) \leq \inf(m-1, p),$$

qui vaut en général $m-1$ c'est-à-dire le nombre de classes moins un.

On note $\lambda_1 \geq \cdots \geq \lambda_h > 0$ les valeurs propres de $\mathbf{S}_e\mathbf{S}_r^{-1}$ et $v^1, \ldots, v^h$ les vecteurs propres$\mathbf{S}_r^{-1}$-orthonormés associés. On pose

$$\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h) \text{ et } \mathbf{V} = [v^1, \ldots, v^h].$$

Les vecteurs v^k sont appelés *vecteurs discriminants* et les sous-espaces vectoriels de dimension 1 qu'ils engendrent dans $\mathbb{R}^p$ les *axes discriminants*.

4.2 Représentation des individus

L'espace des individus est ($\mathbb{R}^p$, b. c., $\mathbf{S}_r^{-1}$). Une représentation simultanée des individus y_i et des barycentres $\overline{y}_k$ des classes par rapport aux mêmes axes discriminants est obtenue dans cet espace au moyen des coordonnées :

$$\begin{aligned} \mathbf{C} &= \mathbf{X}\mathbf{S}_r^{-1}\mathbf{V} \text{ pour les individus et} \\ \overline{\mathbf{C}} &= \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C} \text{ pour les barycentres.} \end{aligned}$$

Les individus initiaux sont projetés comme des individus supplémentaires dans le système des axes discriminants. Comme en A.C.P., on peut calculer des cosinus carrés pour préciser la qualité de représentation de chaque individu.

Il est utile de différencier graphiquement la classe de chaque individu afin de pouvoir apprécier visuellement la qualité de la discrimination.

4.3 Représentation des variables

L'espace des variables est ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$). Chaque variable Y^j est représenté par un vecteur dont les coordonnées dans le système des axes factoriels est une ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$.

4.4 Interprétations

Les interprétations usuelles : la norme est un écart-type, un cosinus d'angle est un coefficient de corrélation, doivent être faites en termes d'écarts-types et de corrélations *expliquées* par la partition.

La représentation des variables est utilisée pour interprétée les axes en fonction des variables initiales conjointement avec la matrice des corrélations expliquées variables×facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$. La matrice $\mathbf{\Sigma}_e^{-1}$ étant la matrice diagonale des écarts-types expliqués σ_e^j c'est-à-dire des racines carrées des éléments diagonaux de la matrice $\mathbf{S}_e$.

Le point essentiel est de savoir si la représentation des individus-barycentres et des individus initiaux permet de faire une bonne discrimination entre les classes définies par la variable T. Si ce n'est pas le cas, l'A.F.D. ne sert à rien, les Y^j n'expliquent pas T. Dans le cas favorable, le graphique des individus permet

d'interpréter la discrimination en fonction des axes et, celui des variables, les axes en fonction des variables initiales. La synthèse des deux permet l'interprétation de T selon les Y^j.

4.5 Exemple de règle d'affectation

Soit s un nouvel individu dont on connaît seulement les valeurs :

$$y_s = [Y^1(s), \ldots, Y^p(s)],$$

alors, la classe d'affectation optimale est celle d'indice :

$$l = \arg \min_{k=1,\ldots,m} \{d^2_{\mathbf{S}_r^{-1}}(\overline{y_k}, y_s)\}.$$

5 Variantes de l'A.F.D.

5.1 Individus de mêmes poids

L'A.F.D. peut être définie de différentes façon, lans la littérature anglo-saxonne, et donc dans la version standard d'A.F.D. du logiciel SAS (procédure `candisc`), ce sont les estimations sans biais des matrices de variances "intra" (within) et "inter" (between) qui sont considérées dans le cas d'individus de mêmes poids $1/n$.

Dans ce cas particulier,

$$\mathbf{D} = \frac{1}{n}\mathbf{I}_n \text{ et } \overline{\mathbf{D}} = \frac{1}{n}\text{diag}(n_1, \ldots, n_m) \text{ où } n_k = \text{card}(\Omega_k)$$

et les matrices de covariances empiriques ont alors pour termes généraux :

$$\begin{aligned}
(\mathbf{S})_j^k &= \frac{1}{n}\sum_{i=1}^{n} x_i^j x_i^k, \\
(\mathbf{S}_e)_j^k &= \frac{1}{n}\sum_{l=1}^{m} n_l \overline{y^j}_l \overline{y^k}_l, \\
(\mathbf{S}_r)_j^k &= \frac{1}{n}\sum_{l=1}^{m}\sum_{i\in\Omega_l} (x_i^j - \overline{y^j}_l)(x_i^k - \overline{y^k}_l).
\end{aligned}$$

Du point de vue de le Statistique inférentielle, on sait que les quantités calculées ci-dessus ont respectivement $(n-1)$, $(m-1)$ et $(n-m)$ degrés de liberté. En conséquence, ce point de vue est obtenu en remplaçant dans les calculs

$$\begin{aligned}
\mathbf{S} \quad &\text{par} \quad \mathbf{S}^* = \frac{n}{n-1}\mathbf{S}, \\
\mathbf{S}_e \quad &\text{par} \quad \mathbf{S}_e^* = \mathbf{B} = \frac{n}{m-1}\mathbf{S}_e, \\
\mathbf{S}_r \quad &\text{par} \quad \mathbf{S}_r^* = \mathbf{W} = \frac{n}{n-m}\mathbf{S}_r.
\end{aligned}$$

Les résultats numériques de l'A.F.D. se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e^* \mathbf{S}_r^{*-1}$	$=$	$\frac{n-m}{m-1}\mathbf{S}_e \mathbf{S}_r^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}^*$	$=$	$\frac{n-m}{m-1}\mathbf{\Lambda}$,
– vecteurs propres :	$\mathbf{V}^*$	$=$	$\sqrt{\frac{n}{n-m}}\mathbf{V}$,
– représentation des barycentres :	$\overline{\mathbf{C}}^*$	$=$	$\sqrt{\frac{n-m}{n}}\overline{\mathbf{C}}$,
– représentation des variables :	$\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\sqrt{\frac{n}{m-1}}\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{*-1}\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Ainsi, les représentations graphiques sont identiques à un facteur d'échelle près tandis que les parts de variance expliquée et les corrélations variables-facteurs sont inchangées.

5.2 Métrique de Mahalanobis

L'A.F.D. est souvent introduite dans la littérature francophone comme un cas particulier d'Analyse Canonique entre un ensemble de p variables quantitatives et un ensemble de m variables indicatrices des modalités de T. La proposition suivante établit les relations entre les deux approches :

Proposition V.5 *l'A.C.P. de* $(\overline{\mathbf{Y}}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ *conduit aux mêmes vecteurs principaux que l'A.C.P. de* $(\overline{\mathbf{Y}}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$. *Cette dernière est l'A.C.P. des barycentres des classes lorsque l'espace des individus est muni de la métrique dite de Mahalanobis* $\mathbf{M} = \mathbf{S}^{-1}$ *et l'espace des variables de la métrique des poids des classes* $\overline{\mathbf{D}}$.

Les résultats numériques de l'A.F.D. se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e\mathbf{S}^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}$,
– vecteurs propres :	$\mathbf{V}(\mathbf{I}+\mathbf{\Lambda})^{1/2}$,
– représentation des barycentres :	$\overline{\mathbf{C}}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}$,
– représentation des variables :	$\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Les représentations graphiques des individus (voir ci-dessus) ne diffèrent alors que d'une homothétie et conduisent à des interprétations identiques, les corrélations variables-facteurs ainsi que les représentations des variables sont inchangées.

6 Exemple

Ce chapitre est illustrée par une comparaison des sorties graphiques issues d'une A.C.P. et d'une A.F.D.. Les données décrivent trois classes d'insectes sur

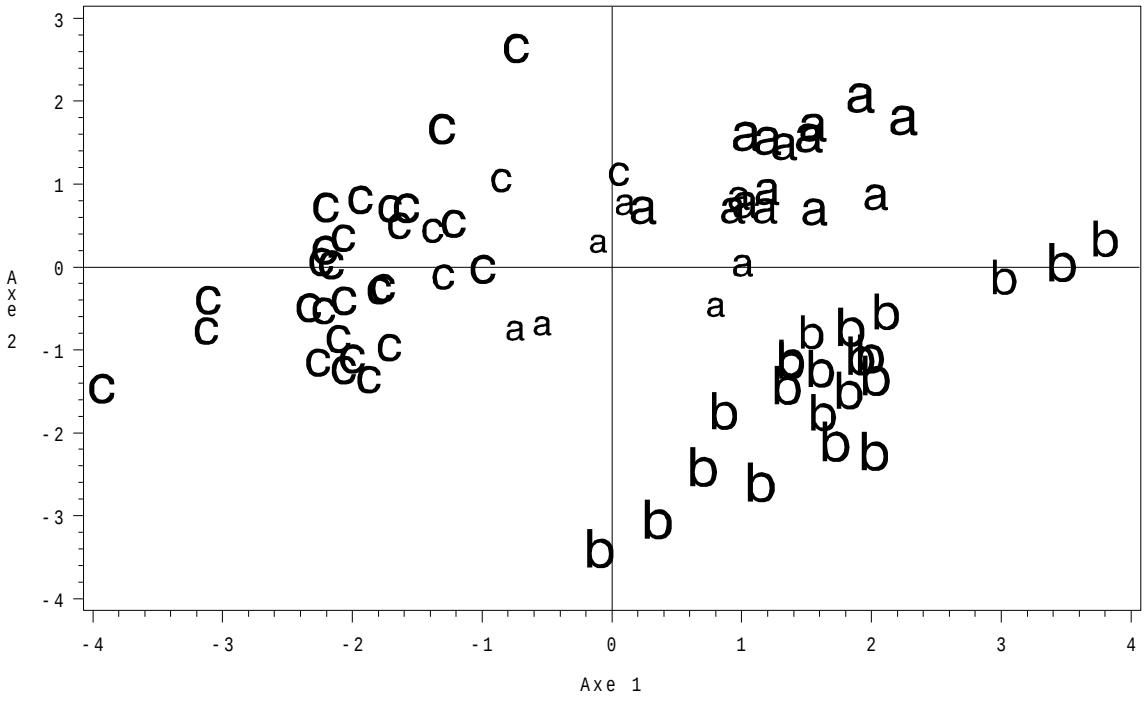


FIG. V.1 - Insectes : premier plan factoriel de l'A.C.P.

lesquels ont été réalisées 6 mesures anatomiques. On cherche à savoir si ces mesures permettent de retrouver la typologie de ces insectes. Ce jeu de données "scolaire" conduit à une bien meilleure discrimination que ce que l'on peut obtenir dans une situation concrète.

7 Projections révélatrices par A.C.P.

L'A.F.D. décrite précédemment fournit une représentation optimale de groupes de points lorsque le découpage en classes est a priori connu. Dans le cas contraire, il n'est pas sûr qu'une A.C.P. usuelle les fasse apparaître ; en effet le critère d'intéret de l'A.C.P. est la recherche d'axes de dispersion ou de variance maximale, il n'est pas de révéler une structure particulière des données surtout si celle-ci est masquée par un bruit de variance imortante.

Ces remarques ont conduit de nombreux auteurs à proposer des techniques cherchant à fournir des représentations privilégiant la manifestation visuelle d'une

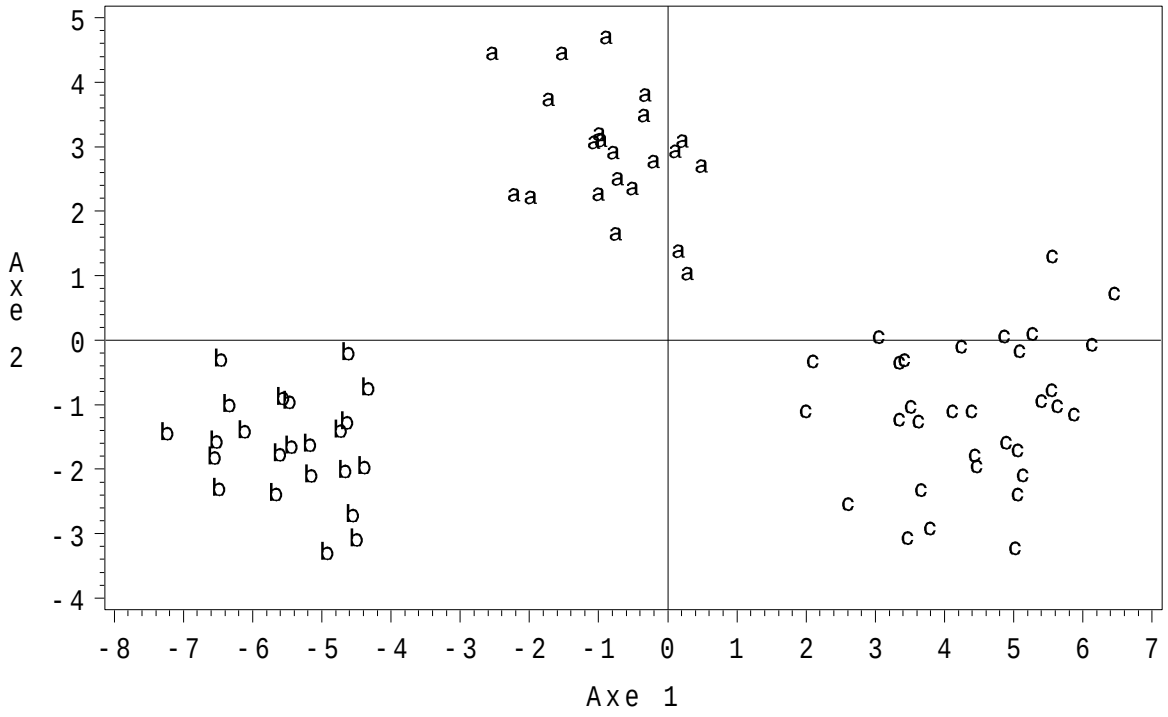


FIG. V.2 - Insectes : premier plan factoriel de l'A.F.D.

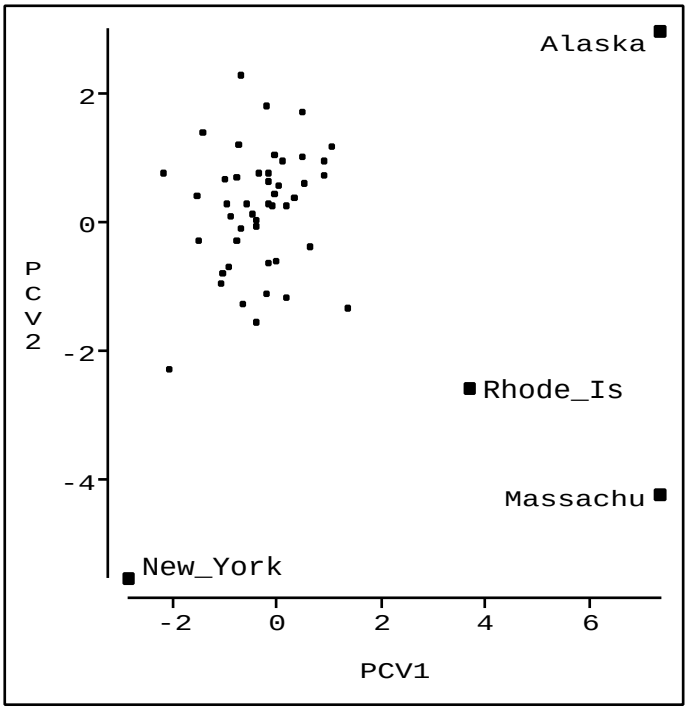


FIG. V.3 - Cartographie des points suspects.

structuration inconnue des données (regroupements, points aberrants,...). Ces techniques, dont Huber (1985) propose un cadre synthétique, recherchent des directions ou plans de projection (projection poursuit) révélateurs en optimisant des critères ne reposant pas sur la variance mais, par exemple, sur la caractérisation de distributions se démarquant au mieux de la normalité (entropie). La mise en œuvre de ces techniques pose des problèmes pratiques :

- l'estimation d'une densité multidimensionnelle nécessite une quantité de données très importante (the curse (la fatalité) of dimensionality),
- les temps de calcul sont considérables,
- les solutions obtenues pour $q = 1, 2, \ldots$ ne sont pas emboîtées.

Dans le même esprit, Caussinus (1992) a proposé, toujours dans le cadre du modèle fonctionnel (IV.1), une adaptation de l'A.C.P. , économique en calcul, dans le but de mettre en évidence des groupements d'individus ou, au contraire, des individus suspects car isolés (outliers) constituant un groupe à eux seul. L'idée, comme pour l'A.F.D., est d'utilisée une métrique spécifique dans l'espace des individus. Celle-ci est une estimation robuste de la matrice de covariance.

Ainsi, des points suspects, globalement pour toutes les variables (multivariate outliers), sont identifiés et cartographiés en exécutant une A.C.P. utilisant pour

métrique l'inverse de la matrice :

$$\widehat{\mathbf{\Psi}} = \frac{\sum_{i=1}^{n-1} k\left(\|\mathbf{y}_i - \bar{\mathbf{y}}\|^2_{\mathbf{S}^{-1}}\right)(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'}{\sum_{i=1}^{n-1} k\left(\|\mathbf{y}_i - \bar{\mathbf{y}}\|^2_{\mathbf{S}^{-1}}\right)}.$$

k est une fonction décroissante de $\mathbb{R}^+$ dans $\mathbb{R}^+$ comme par exemple : $k(u) = e^{-hu}$ où h est un paramètre positif à fixer par l'utilisateur.

Les points suspects ayant été inhibés, il est alors possible de rechercher des groupes en estimant la matrice de covariance intra-groupe inconnue, qui est aussi la covariance de l'erreur, par :

$$\widehat{\mathbf{\Psi}} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k\left(\|\mathbf{y}_i - \mathbf{y}_j\|^2_{\mathbf{S}^{-1}}\right)(\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)'}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k\left(\|\mathbf{y}_i - \mathbf{y}_j\|^2_{\mathbf{S}^{-1}}\right)}.$$

Chapitre VI

Analyse Factorielle des Correspondances

1 Introduction

1.1 Données

On considère dans ce chapitre deux variables qualitatives observées simultanément sur n individus affectés de poids identiques $1/n$. On suppose que la première variable , notée X, possède r modalités notées $x_1, \ldots, x_l, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

La table de contingence, de dimension $r \times c$, est notée $\mathbf{T}$ et son élément générique n_{lh} (effectif conjoint). Elle se présente sous la forme suivante :

	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_{l1}	$\cdots$	n_{lh}	$\cdots$	n_{lc}	n_{l+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

1.2 Notations

Les quantités $\{n_{l+} = \sum_{h=1}^{c} n_{lh}\,; l = 1, \ldots, r\}$ et $\{n_{+h} = \sum_{l=1}^{r} n_{lh}\,; h = 1, \ldots, c\}$ sont les *effectifs marginaux*; et vérifient $\sum_{l=1}^{r} n_{l+} = \sum_{h=1}^{c} n_{+h} = n$. De façon analogue, on définit les notions de *fréquences conjointes* ($f_{lh} = n_{lh}/n$) et de *fréquences marginales*. Ces dernières sont rangées dans les vecteurs :

$$g_r = [f_{1+}, \ldots, f_{r+}]',$$

$$\text{et } g_c = [f_{+1}, \ldots, f_{+c}]'.$$

Elles permettent de définir les matrices :

$$\begin{aligned} \mathbf{D}_r &= \text{diag}(f_{1+}, \ldots, f_{r+}), \\ \text{et } \mathbf{D}_c &= \text{diag}(f_{+1}, \ldots, f_{+c}). \end{aligned}$$

On sera de plus amené à considérer les profils-lignes et les profils-colonnes déduits de $\mathbf{T}$. Le $l^{\text{ième}}$ profil-ligne est

$$\{\frac{n_{l1}}{n_{l+}}, \ldots, \frac{n_{lh}}{n_{l+}}, \ldots, \frac{n_{lc}}{n_{l+}}\}.$$

Il est considéré comme un vecteur de $\mathbb{R}^c$; l'ensemble des r vecteurs ainsi définis sont disposés en colonnes dans la matrice $(c \times r)$

$$\mathbf{A} = \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}.$$

De même, le $h^{\text{ième}}$ profil-colonne est

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

vecteur de $\mathbb{R}^r$, et la matrice $(r \times c)$ des profils-colonnes est

$$\mathbf{B} = \frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}.$$

1.3 Liaison entre deux variables qualitatives

Définition VI.1 *On dit que deux variables X et Y sont* non liées *relativement à T si*

$$\forall (l,h) \in \{1, \ldots, r\} \times \{1, \ldots, c\}, n_{lh} = \frac{n_{l+}n_{+h}}{n}.$$

C'est équivalent à dire que tous les profils lignes sont égaux ou encore que tous les profiles colonnes sont égaux. Cette notion est cohérente avec celle d'indépendance en probabilités.

Soit $\Omega = \{1, \ldots, n\}$ l'ensemble des individus, $(\Omega, \mathcal{P}(\Omega), P)$ l'espace probabilisé où P est l'équiprobabilité ; $\mathcal{M}_X = \{x_1, \ldots, x_r\}$ et $\mathcal{M}_Y = \{y_1, \ldots, y_c\}$ désignent les ensembles de modalités ou valeurs prises par les varables X et Y. On note $\widetilde{X}$ et $\widetilde{Y}$ les variables aléatoires associées aux 2 variables statistiques X et Y :

$$\begin{aligned} \widetilde{X} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), \\ \widetilde{Y} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y)) ; \end{aligned}$$

P_X, P_Y et P_{XY} désignent respectivement les probabilités images définies par $\widetilde{X}, \widetilde{Y}$ et le couple $(\widetilde{X}, \widetilde{Y})$ sur $(\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y))$ et $(\mathcal{M}_X \times \mathcal{M}_Y, \mathcal{P}(\mathcal{M}_X) \times \mathcal{P}(\mathcal{M}_Y))$; ce sont les probabilités empiriques.

Alors, X et Y sont *non liées* si et seulement si $\widetilde{X}$ et $\widetilde{Y}$ sont *indépendantes en probabilité*.

On suppose qu'il existe une liaison entre X et Y que l'on souhaite étudier. La représentation graphique des profils-lignes, ou des profils-colonnes, au moyen de diagrammes en barre parallèles, le calcul de coefficients de liaison (Pearson, Cramer, Tschuprow) donnent une idée assez précise de la variation conjointe des deux variables (cf. chapitre III). Le test du χ^2 permet également de s'assurer de la significativité d'une liaison, il est construit de la manière suivante :

$H0$: $\widetilde{X}$ et $\widetilde{Y}$ sont indépendantes,

$H1$: les variables sont liées.

La statistique du test est :

$$\chi^2 = \sum_{l=1}^{n} \sum_{h=1}^{c} \frac{\left(n_{lh} - \frac{n_{l+} n_{+h}}{n}\right)^2}{\frac{n_{l+} n_{+h}}{n}},$$

et suit asymptotiquement (n grand) une loi du χ^2 à $(r-1)(c-1)$ degrés de liberté.

1.4 Objectifs

Pour aller au delà d'une caractérisation globale, on souhaite définir un modèle susceptible de préciser la liaison entre les variables X et Y et de fournir des paramètres dont la représentation graphique, de type biplot, illustre les "*correspondances*" entre modalités. L'Analyse Factorielle des Correspondances (AFC) s'intéresse ainsi à l'importance relative des écarts $n_{lh} - (n_{l+} n_{+h})/n$ qui s'annulent lors d'une absence de liaison.

Une autre approche, courante dans la littérature francophone, consiste à définir l'A.F.C. comme étant la représentation graphique simultanée issue d'une double A.C.P. :

- A.C.P. des profiles lignes,
- A.C.P. des profiles colonnes,

relativement à la métrique dite du χ^2.

Remarques VI.1 :

i. Toute structure d'ordre éventuelle existant sur les modalités de X ou de Y est ignorée par l'analyse.

ii. Tout individu prend une modalité et une seule de chaque variable.

iii. Chaque modalité doit avoir été observée au moins une fois ; sinon, elle est supprimée.

2 Double A.C.P.

2.1 Métrique du χ^2

Les correspondances entre modalités évoquées au paragraphe précédant se trouvent exprimées en termes de distances au sens d'un certaine métrique. Chaque modalité x_l de X est caractérisée par son profil-ligne représenté par le vecteur a^l de l'espace $\mathbb{R}^c$ muni de la base canonique. De même, chaque modalité y_h de Y est caractérisée par son profil-colonne représenté par le vecteur b^h de l'espace $\mathbb{R}^r$ muni de la base canonique.

Ces espaces sont respectivement munis des métriques, dites du χ^2, de matrices $\mathbf{D}_c^{-1}$ et $\mathbf{D}_r^{-1}$. Ainsi, la distance entre deux modalités x_l et x_i de X s'écrit :

$$\|a^l - a^i\|^2_{\mathbf{D}_c^{-1}} = \sum_{h=1}^{c} \frac{1}{f_{+h}}(a^l_h - a^i_h)^2$$

et de même pour les modalités de Y. La métrique du χ^2 introduit les inverses des fréquences marginales des modalités de Y comme *pondérations* des écarts entre éléments de deux profils relatifs à X (et réciproquement) ; elle attribue donc plus de poids aux écarts correspondants à des modalités de *faible effectif* (rares) pour Y.

2.2 A.C.P. des profils colonnes

On s'intéresse à l'A.C.P. de $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$ qui considère comme "individus" les vecteurs profiles colonnes rangés en lignes dans la matrice $\mathbf{B}'$ et munis du poids de chaque modalité.

Proposition VI.1 *Les éléments de l'A.C.P. de* $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$ *sont fournis par l'analyse spectrale de la matrice carrée* $\mathbf{D}_r^{-1}$*-symétrique positive* $\mathbf{BA}$.

Preuve Elle se construit en remarquant successivement que :

i. le barycentre du nuage des profiles colonnes est le vecteur g_r des fréquence marginales,

ii. la matrice $\mathbf{B}\mathbf{D}_c\mathbf{B}' - g_r\mathbf{D}_c g_r'$ joue le rôle de la matrice de covariances,

iii. la solution de l'A.C.P. est fournie par la D.V.S. de $(\mathbf{B}' - \mathbf{1}g_r', \mathbf{D}_r^{-1}, \mathbf{D}_c)$ qui conduit à rechercher les valeurs et vecteurs propres de la matrice $(\mathbf{SM})$:

$$\mathbf{B}\mathbf{D}_c\mathbf{B}'\mathbf{D}_r^{-1} - g_r\mathbf{D}_c g_r' = \mathbf{BA} - g_r g_r'\mathbf{D}_r^{-1} \quad (\text{ car } \mathbf{D}_c^{-1}\mathbf{A} = \mathbf{B}'\mathbf{D}_r^{-1}),$$

iv. les matrices $\mathbf{BA} - g_r g_r' \mathbf{D}_r^{-1}$ et $\mathbf{BA}$ ont les mêmes vecteurs propres associées aux mêmes valeurs propres à l'exception du vecteur g_r associé à la valeur propre $\lambda_0 = 0$ de $\mathbf{BA} - g_r g_r' \mathbf{D}_r^{-1}$ et à la valeur propre $\lambda_0 = 1$ de $\mathbf{BA}$.

□

On note $\mathbf{V}$ la matrice contenant les vecteurs propres $\mathbf{D}_r^{-1}$-ortnonormés. La représentation des "individus" profiles colonnes amène à une représentation des modalités de la variable Y en utilisant les lignes de la matrice des "composantes principales" ($\mathbf{XMV}$) :

$$\mathbf{C}_c = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{V}.$$

2.3 A.C.P. des profiles lignes

De façon symétrique ou duale, on s'intéresse à l'A.C.P. des "individus" profiles lignes (matrice $\mathbf{A}$') relativement à la métrique du χ^2 lorqu'ils sont munis des poids des classes (matrice $\mathbf{D}_r$) c'est-à-dire à l'A.C.P. de $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$.

Proposition VI.2 *Les éléments de l'A.C.P. de* $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$ *sont fournis par l'analyse spectrale de la matrice carrée* $\mathbf{D}_c^{-1}$*-symétrique positive* $\mathbf{AB}$.

On obtient directement les résultats en échangeant respectivement les matrices $\mathbf{A}$ et $\mathbf{B}$ ainsi que les indices c et r. Notons $\mathbf{U}$ la matrice des vecteurs propres de la matrice $\mathbf{AB}$, les coordonnées représentant les modalités de la variable X sont fournies par la matrice :

$$\mathbf{C}_r = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{U}.$$

Sachant que $\mathbf{U}$ contient les vecteurs propres de $\mathbf{AB}$ et $\mathbf{V}$ ceux de $\mathbf{BA}$, le théorème (A.1) montre qu'il suffit de calculer une seule analyse car les résultats de l'autre s'en déduisent aisément :

$$\begin{aligned} \mathbf{U} &= \mathbf{AV}\mathbf{\Lambda}^{-1/2}, \\ \mathbf{V} &= \mathbf{BU}\mathbf{\Lambda}^{-1/2} \end{aligned}$$

où $\mathbf{\Lambda}$ est la matrice diagonale des valeurs propres (exceptée $\lambda_0 = 0$) communes aux deux A.C.P.

$$\begin{aligned} \mathbf{C}_c &= \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{V} = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{BU}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{ABU}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2}, \\ \mathbf{C}_r &= \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{U} = \mathbf{D}_r^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}. \end{aligned}$$

On en déduit les formules dites de *transition* :

$$\begin{aligned} \mathbf{C}_c &= \mathbf{B}'\mathbf{C}_r\mathbf{\Lambda}^{-1/2}, \\ \mathbf{C}_r &= \mathbf{A}'\mathbf{C}_c\mathbf{\Lambda}^{-1/2}. \end{aligned}$$

La représentation simultanée habituellement construite à partir de ces matrices (option par défaut de SAS) n'est pas justifiée. On va lui donner un sens dans les paragraphes suivants.

3 Modèles pour une table de contingence

On écrit d'abord que chaque fréquence f_{lh} de $\mathbf{T}$ correspond à l'observation d'une probabilité théorique p_{lh} : on modélise donc la table de contingence par cette distribution de probabilités. On précise ensuite le modèle en explicitant l'écriture de p_{lh}. Différents modèles classiques peuvent être considérés.

3.1 Modèle log-linéaire

Il consiste à écrire :

$$\ln(p_{lh}) = \mu + \alpha_l + \beta_h + \gamma_{lh}$$

avec des contraintes le rendant identifiable. Ce modèle très classique ne sera pas développé ici. On pourra se reporter, par exemple, à Bishop et al. (1975).

3.2 Modèle d'association

Il est encore appelé RC-modèle ou modèle de Goodman (1991) :

$$p_{lh} = \gamma\alpha_l\beta_h \exp\left(\sum_{k=1}^{q} \phi_k \mu_{lk} \nu_{hk}\right).$$

Ce modèle, muni des contraintes nécessaires, permet de structurer les interactions et de faire des représentations graphiques des lignes et des colonnes de $\mathbf{T}$ au moyen des μ_{lk} et des ν_{hk}. Les paramètres peuvent être estimés par maximum de vraisemblance ou par moindres carrés.

3.3 Modèle de corrélation

On s'intéresse aux fréquences et on écrit :

$$p_{lh} = p_{l+}p_{+h} + \sum_{k=1}^{q} \sqrt{\lambda_k} u_l^k v_h^k, \qquad \text{(VI.1)}$$

avec $q \leq \inf(r-1, c-1)$, $\lambda_1 \geq \cdots \geq \lambda_q > 0$ et sous les contraintes d'identifiabilité :

$$\begin{aligned} \sum_{l=1}^{r} u_l^k = \sum_{h=1}^{c} v_h^k &= 0, \\ u^{k'}\mathbf{D}_r^{-1}u^j = v^{k'}\mathbf{D}_c^{-1}v^j &= \delta_{kj}. \end{aligned}$$

Remarques VI.2 *:*

i. Le modèle (VI.1) ci-dessus est équivalent au modèle considéré par Goodman (1991) :

$$p_{lh} = p_{l+}p_{+h}\left(1 + \sum_{k=1}^{q} \sqrt{\lambda_k}\alpha_l^k\beta_h^k\right), \qquad \text{(VI.2)}$$

moyennant une homothétie des paramètres.

ii. La quantité $\sum_{k=1}^{q} \sqrt{\lambda_k} u_l^k v_h^k$ *exprime l'écart à l'indépendance pour la cellule considérée.*

iii. Le modèle suppose que cet écart se décompose dans un sous-espace de dimension $q < \min(c-1, r-1)$.

iv. Les estimations des paramètres $p_{l+}, p_{+h}, \lambda_k, u^k, v^k$ *peuvent être réalisées par maximum de vraisemblance*[1] *ou par moindres carrés. Cette dernière solution est ici retenue.*

4 Estimation et AFC

4.1 Critère

Considérons les espaces $\mathbb{R}^c$ et $\mathbb{R}^r$ munis de leurs métriques du χ^2 respectives et notons $\mathbf{P}$ le tableau des probabilités théoriques ; le critère des moindres carrés s'écrit alors :

$$\min_{\mathbf{P}} \left\| \frac{1}{n}\mathbf{T} - \mathbf{P} \right\|^2_{\mathbf{D_r^{-1}D_c^{-1}}} . \tag{VI.3}$$

4.2 Estimation

Proposition VI.3 *L'estimation des paramètres de (VI.1) en résolvant (VI.3) est fournie par la D.V.S. de* $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$ *à l'ordre q. Les probabilités marginales* p_{l+} *et* p_{+h} *sont estimées par* f_{l+} *et* f_{+h} *tandis que les vecteurs* u^k *(resp.* v^k*) sont vecteurs propres de la matrice* $\mathbf{AB}$ *(resp.* $\mathbf{BA}$*) associés aux valeurs propres* λ_k.

On obtient ainsi, d'une autre façon, l'A.F.C. de la table de contingence $\mathbf{T}$.

Preuve

Elle se construit à partir de la D.V.S. de $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$:

$$\frac{1}{n} t_l^h = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k} u_l^k v_h^k,$$

où les vecteurs u^k (resp. v^k) sont vecteurs propres $\mathbf{D}_r^{-1}$-orthonormés (resp. $\mathbf{D}_c^{-1}$-orthonormés) de la matrice

$$\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}\frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1} = \mathbf{BA} \quad (\text{resp. } \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1} = \mathbf{AB}),$$

associés aux valeurs propres λ_k.

1. On suppose alors que les n p_{lh} sont les paramètres de lois de Poisson indépendantes conditionnellement à leur somme qui est fixée et égale à n.

De plus, le vecteur $g_r = u^0$ (resp. $g_c = v^0$) est vecteur propre $\mathbf{D}_r^{-1}$-normé (resp. $\mathbf{D}_c^{-1}$-normé) de la matrice $\mathbf{AB}$ (resp. $\mathbf{BA}$) associé à la valeur propre $\lambda_0 = 1$. Enfin, les matrices $\mathbf{AB}$ et $\mathbf{BA}$ sont stochastiques [2] et donc les valeurs propres vérifient :

$$1 = \lambda_0 > \lambda_1 \geq \cdots \lambda_q > 0.$$

En identifiant les termes, l'approximation de rang $(q+1)$ de la matrice $\mathbf{P}$ s'écrit donc :

$$\widehat{\mathbf{P}}_q = g_r g_c' + \sum_{k=1}^{q} \sqrt{\lambda_k} u^k v^{k'}$$

et les propriétés d'orthonormalité des vecteurs propres assurent que les contraintes du modèle sont vérifiées.

□

5 Représentations graphiques

5.1 Biplot

La décomposition de la matrice $\frac{1}{n}\mathbf{T}$ se transforme encore en :

$$\frac{f_{lh} - f_{l+}f_{+h}}{f_{l+}f_{+h}} = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k} \frac{u_l^k}{f_{l+}} \frac{v_h^k}{f_{+h}}.$$

En se limitant au rang q, on obtient donc, pour chaque cellule (l,h) de la table $\mathbf{T}$, une approximation de son écart relatif à l'indépendance comme produit scalaire des deux vecteurs :

$$\mathbf{D}_r^{-1} u_l \Lambda^{1/4} \text{ et } \mathbf{D}_c^{-1} v_h \Lambda^{1/4}$$

qui sont encore les estimations des vecteurs α_l et β_h du modèle VI.2. Leur représentation (par exemple avec $q = 2$) illustre alors la *correspondance* entre ces deux modalités x_l et y_h : lorsque deux modalités, éloignés de l'origine, sont voisines ou antagonistes, le produit scalaire est de valeur absolue importante ; leur cellule conjointe contribue fortement à la dépendance entre les deux variables.

L'A.F.C. apparait ainsi comme la meilleure reconstitution des fréquences f_{ij} ou encore la meilleure représentation des écarts-types relatifs à l'indépendance.

5.2 Double ACP

Chacune des deux A.C.P. propose une représentation des "individus"-modalités approchant, au mieux, les distances du χ^2 entre les profils-lignes d'une part,

2. Matrice réelle carrée, de termes positifs et dont la somme des termes de chaque ligne (ou chaque colonne) vaut 1.

les profils-colonnes d'autre part. Les coordonnées (composantes principales) sont données cette fois par les vecteurs

$$\mathbf{D}_c^{-1} u_l \Lambda^{1/2} \text{ et } \mathbf{D}_r^{-1} v_h \Lambda^{1/2}.$$

Même si la représentation simultanée n'a plus alors de justification, elle reste couramment employée. En fait, pour les premières valeurs propres, les graphiques obtenus diffèrent peu de ceux du biplot qui sert de "caution" car les interprétations sont identiques.

5.3 Représentations barycentriques

D'autres représentations sont proposées utilisant les vecteurs

$$\mathbf{D}_r^{-1} u_l \text{ et } \mathbf{D}_c^{-1} v_h \Lambda^{1/2},$$

ou encore les vecteurs

$$\mathbf{D}_r^{-1} u_l \Lambda^{1/2} \text{ et } \mathbf{D}_c^{-1} v_h.$$

Celles-ci, souvent illisibles, sont peu utilisées en pratique.

5.4 Aides à l'interprétation

Les qualités de représentation et contributions se déduisent aisément de celles de l'A.C.P. considérée.

6 Exemple

La table de contingence étudiée s'intéresse à la répartition des exploitation agricoles de la région Midi-Pyrénées dans les différents départements. Elle croise la variable qualitative *département* avec la variable *taille de l'exploitation* qui est quantitative découpée en classes. Les données ainsi que les résultats numériques fournis par la procédure `corresp` de SAS/STAT sont listés en annexe. La figure 6 présente le premier plan factoriel utilisant les coordonnées obtenues par défaut, c'est-à-dire celles de la double ACP.

7 Compléments

7.1 Propriétés

- Les *formules de transition* exprimant u^k en fonction de v^k ainsi que les *formules de reconstitution des données* (la table $1/n\mathbf{T}$), qui sont classiques dans la littérature, ne sont que des conséquences de la décomposition en valeurs singulières de $1/n\mathbf{T}$.

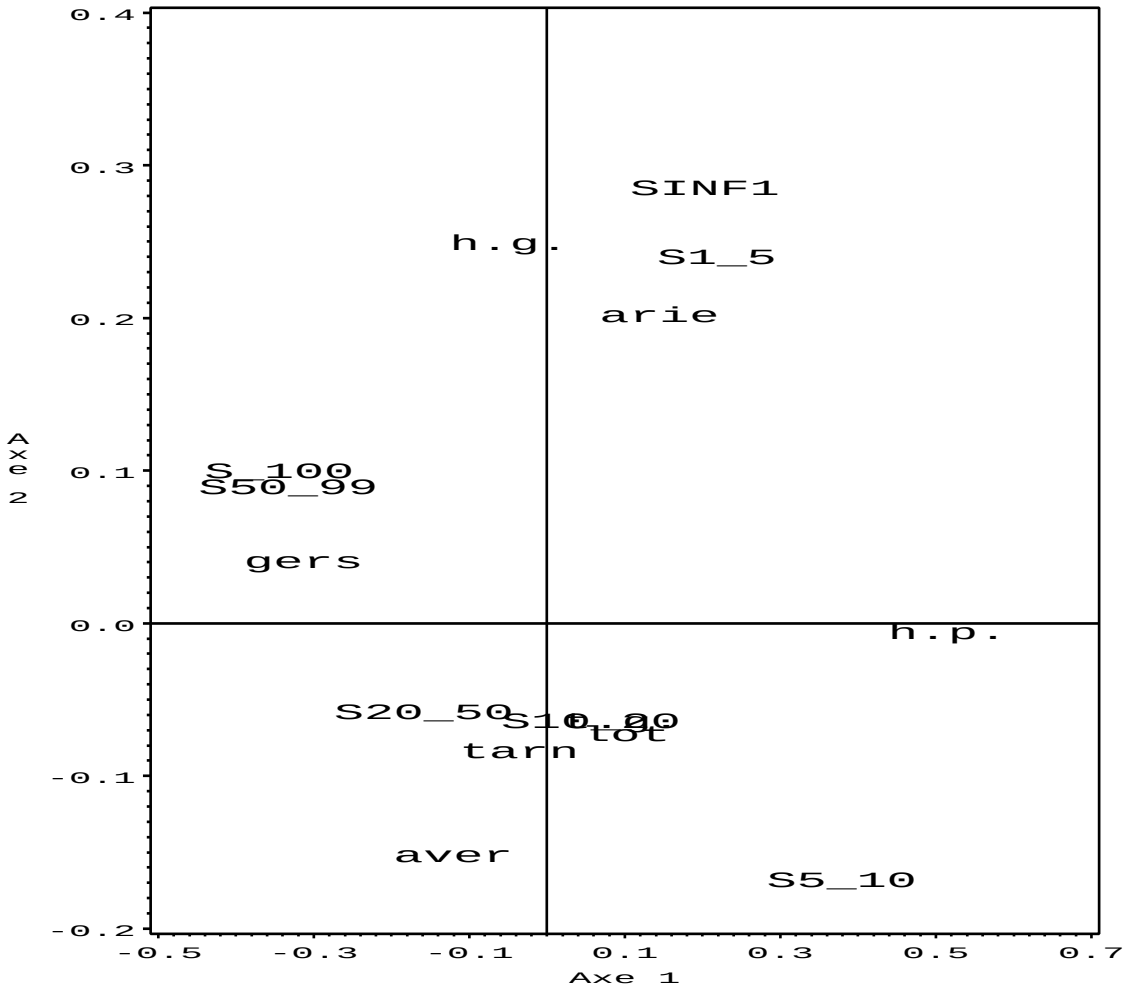


FIG. VI.1 - Répartition des exploitations par département

- Les valeurs propres vérifient :

$$\sum_{k=0}^{\min(r-1,c-1)} \lambda_k = \mathrm{tr}\mathbf{AB} = 1 + \frac{\chi^2}{n} = 1 + \Phi^2.$$

7.2 Invariance

- Les tables de contingence $\mathbf{T}$ et $\alpha\mathbf{T}, \alpha \in \mathbb{R}_+^*$, admettent la même AFC.
- *Invariance distributionnelle* : si deux lignes de $\mathbf{T}$, l et i, ont des effectifs proportionnels, alors les représentations de x_l et x_i sont confondues (les profils sont identiques) et le regroupement de x_l et x_i en une seule modalité laisse inchangées les représentations graphiques (même chose pour les modalités colonnes de Y). Cette propriété est caractéristique de la métrique du χ^2.

7.3 Choix de dimension q

Le choix de dimension pose les mêmes problèmes qu'en ACP. De nombreuses techniques empiriques ont été proposées (part de variance expliquée, éboulis des valeurs propres). Il existe également une approche probabiliste qui peut donner des indications intéressantes sans être théoriquement justifiée.

Posons

$$\widehat{n_{lh}^q} = n f_{l+} f_{+h} + n \sum_{k=1}^{q} \sqrt{\lambda_k} u^k v^{k'},$$

estimation d'ordre q de l'effectif conjoint de la cellule (l,h). Alors, sous certaines conditions (échantillonnage , n grand, modèle multinomial...), on peut montrer que

$$K_q = \sum_l \sum_h \frac{(n_{lh} - \widehat{n_{lh}^q})^2}{\widehat{n_{lh}^q}} \simeq n \sum_{k=q+1}^{\min(r-1,c-1)} \lambda_k$$

suit approximativement une loi du χ^2 à $(r-q-1)(c-q-1)$ degrés de liberté. On peut donc retenir pour valeur de q la plus petite dimension pour laquelle K_q est inférieure à la valeur limite de cette loi. Le choix $(q=0)$ correspond à la situation où les variables sont indépendantes. Les fréquences conjointes sont bien approchées par les produits des fréquences marginales.

Une autre approche, sans hypothèse de type probabiliste, peut être calquée sur celle proposée en ACP et minimisant le risque quadratique moyen sur l'estimation du sous-espace principal de dimension q. Ce risque est, dans ce cas, estimé par bootstrap.

Chapitre VII

Analyse Factorielle Multiple des Correspondances

Cette méthode a pour objectif de proposer une généralisation de l'Analyse des Correspondances en vue de décrire les relations entre p ($p > 2$) variables qualitatives simultanément observées sur n individus.

1 Codages de variables qualitatives

Soit X une variable qualitative à c modalités. On appelle *variable indicatrice* de la $k$ième modalité de x ($k = 1, \ldots, c$), la variable $X_{(k)}$ définie par :

$$X_{(k)}(i) = \begin{cases} 1 & \text{si } X(i) = \mathcal{X}_k \\ 0 & \text{sinon} \end{cases}$$

où i est un individu quelconque et $\mathcal{X}_k$ la $k$ième modalité de X d'effectif n_k.

On note $\mathbf{X}$ $(n \times c)$ la *matrice des indicatrices* des modalités de la variable X et de terme général :

$$x_i^k = X_{(k)}(i) \text{ avec } \sum_{l=1}^{c} x_i^l = 1, \sum_{i=1}^{n} x_i^k = n_k.$$

Considérons p variables qualitatives $X^1, \ldots, X^p$. On note c_j le nombre de modalités de X^j, $c = \sum_{j=1}^{p} c_j$ et $\mathbf{X}_j$ la matrice des indicatrices de X^j. On appelle *tableau disjonctif complet* la matrice $\mathbf{X}(n \times c)$ par blocs obtenue par concaténation des matrices $\mathbf{X}_j$:

$$\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p], \text{ avec } \sum_{l=1}^{c} x_i^l = p, \sum_{i=1}^{n} \sum_{l=1}^{c} x_i^l = np.$$

On appelle *tableau de Burt* la matrice $(c \times c)$:

$$\mathcal{B} = \mathbf{X}'\mathbf{X} = [\mathcal{B}_{jl}]$$

dans laquelle chaque bloc $\mathcal{B}_{jl}$ $(c_j \times c_l)$ défini par

$$\mathcal{B}_{jl} = \mathbf{X}_j' \mathbf{X}_l$$

est la table de contingence obtenue par croisement des variables X^j et X^l.

La matrice $\mathcal{B}$ est symétrique d'effectifs marginaux pn_l^j, d'effectif total np^2 et dont les blocs diagonaux sont des matrices diagonales :

$$\mathcal{B}_{ll} = \text{ diag}(n_1^l, \ldots, n_{c_l}^l).$$

2 A.F.C. du tableau disjonctif complet relatif à deux variables

La généralisation de l'A.F.C. à plusieurs variables repose sur certaines propriétés observées dans le cas élémentaire où $p = 2$. On s'intéresse d'abord aux résultats fournis par une A.F.C. calculée sur le tableau disjonctif complet $\mathbf{X} = [\mathbf{X}_r|\mathbf{X}_c]$ $(n \times (r + c))$ de deux variables considéré comme une table de contingence..

Les matrices usuelles de l'A.F.C. deviennent alors :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X}, \\
\overline{\mathbf{D}_r} &= \frac{1}{n}\mathbf{I}_n, \\
\overline{\mathbf{D}_c} &= \frac{1}{2}\mathbf{\Delta} = \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix}, \\
\overline{\mathbf{A}} &= \frac{1}{2n}\overline{\mathbf{T}}'\overline{\mathbf{D}}_r^{-1} = \frac{1}{2}\mathbf{X}', \\
\overline{\mathbf{B}} &= \frac{1}{2n}\overline{\mathbf{T}\mathbf{D}}_c^{-1} = \frac{1}{n}\mathbf{X}\mathbf{\Delta}^{-1}.
\end{aligned}$$

On considère l'A.F.C. définie comme une double A.C.P. des profils-lignes $\overline{\mathbf{A}}$ et des profils-colonnes $\overline{\mathbf{B}}$.

2.1 A.C.P. des profils lignes

Les profils-lignes sont associés aux n individus observés. Cette A.C.P. conduit donc à une représentation graphique des individus inconnue en A.F.C. classique.

Proposition VII.1 *L'A.C.P. des profils-lignes issue de l'A.F.C. calculée sur le tableau disjonctif complet associé à deux variables X^r et X^c conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}_c}^{-1}$-symétrique positive :*

$$\overline{\mathbf{AB}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix}.$$

Les $(r+c)$ valeurs propres s'écrivent :

$$\nu_k = \frac{1 \pm \sqrt{\lambda_k}}{2},$$

où les λ_k sont les valeurs propres de la matrice $\mathbf{AB}$. *Les vecteurs propres associés* $\overline{\mathbf{D}_c}^{-1}$*-orthonormés se mettent sous la forme :*

$$\overline{\mathbf{U}} = \frac{1}{2}\begin{bmatrix} \mathbf{V} \\ \mathbf{U} \end{bmatrix}.$$

La matrice $\mathbf{U}$ *(resp.* $\mathbf{V}$*) contenant les vecteurs propres* $\mathbf{D}_c^{-1}$*-orthonormés (resp.* $\mathbf{D}_r^{-1}$*-orthonormés) de la matrice* $\mathbf{AB}$ *(resp.* $\mathbf{BA}$*). La matrice des composantes principales s'expriment par :*

$$\overline{\mathbf{C}_r} = \frac{1}{2}\left[\mathbf{X}_r\mathbf{C}_r + \mathbf{X}_c\mathbf{C}_c\right]\mathbf{\Lambda}^{-1/2}.$$

On ne considère que les $r-1$ $(r < c)$ plus grandes valeurs propres différentes de 1 ainsi que les vecteurs propres associés. Ces valeurs propres sont rangées dans la matrice

$$\mathbf{N} = \mathrm{diag}(\nu_k \ ; k = 1, \ldots, r-1).$$

Les autres valeurs propres non nulles sont dues à l'artifice de construction de la matrice à diagonaliser, elles n'ont pas de signification statistique.

2.2 A.C.P. des profils colonnes

Les profils-colonnes sont associés aux $(r+c)$ modalités des variables. Cette A.C.P. conduit donc à une représentation graphique de ces modalités dont on verra qu'elle est très voisine de celle fourni par une A.F.C. classique.

Proposition VII.2 *L'A.C.P. des profils-colonnes issue de l'A.F.C. calculée sur le tableau disjonctif complet associé à deux variables X^r et X^c conduit à l'analyse spectrale de la matrice* $\overline{\mathbf{D}_r}^{-1}$*-symétrique positive :*

$$\overline{\mathbf{BA}} = \frac{1}{2n}\left[\mathbf{X}_r\mathbf{D}_r^{-1}\mathbf{X}_r' + \mathbf{X}_c\mathbf{D}_c^{-1}\mathbf{X}_c'\right].$$

Les $(r+c)$ valeurs propres non nulles s'écrivent :

$$\nu_k = \frac{1 \pm \sqrt{\lambda_k}}{2}$$

où les λ_k sont les valeurs propres de la matrice $\mathbf{BA}$. *Les vecteurs propres* $\overline{\mathbf{D}_r}^{-1}$*-orthonormés associés se mettent sous la forme :*

$$\overline{\mathbf{V}} = \frac{1}{n}\overline{\mathbf{C}_r}\mathbf{N}^{-1/2} \ \textit{avec}\ \mathbf{N}^{1/2} = \frac{1}{2}\left[\mathbf{I}_{r-1} + \mathbf{\Lambda}^{1/2}\right].$$

La matrice des composantes principales s'expriment par:

$$\overline{\mathbf{C}_c} = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{N}^{1/2}.$$

Ainsi, l'A.F.C. du tableau disjonctif complet fournit une représentation graphique des modalités très voisine de celle de l'A.F.C. définie précédemment, une simple homothétie de coefficients $\sqrt{\frac{1+\sqrt{\lambda_k}}{2\lambda_k}}$ permet de passer de l'une à l'autre.

Les coordonnées sont contenues dans les lignes de la matrice $\overline{\mathbf{C}_c}$. Cette approche permet de réaliser également une représentation graphique des individus avec les coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}_r}$. À un facteur près, chaque individu apparaît comme le barycentre des modalités qui lui sont affectées. Dans le cas de n grand, seul un "graphique de densité" : un point par individu, est réalisable.

3 A.F.C. du tableau de Burt relatif à deux variables

On s'intéresse dans cette section aux résultats fournis par une A.F.C. calculée sur le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$ $((r+c)\times(r+c))$ de deux variables considéré comme une table de contingence.

Cette matrice étant symétrique, les profils-lignes et colonnes sont identiques, il suffit de considérer une seule A.C.P.

Les matrices usuelles de l'A.F.C. deviennent dans ce cas:

$$\begin{aligned} \widetilde{\mathbf{T}} &= \mathcal{B} = \begin{bmatrix} n\mathbf{D}_r & \mathbf{T} \\ \mathbf{T}' & n\mathbf{D}_c \end{bmatrix}, \\ \widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{2}\mathbf{\Delta}, \\ \widetilde{\mathbf{A}} &= \tilde{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix}. \end{aligned}$$

On considère l'A.F.C. définie comme une double A.C.P. des profils-lignes $\overline{\mathbf{A}}$ et des profils-colonnes $\overline{\mathbf{B}}$

Proposition VII.3 *L'A.C.P. des profils-lignes (ou colonnes) issue de l'A.F.C. calculée sur le tableau de Burt associé à deux variables X^r et X^c conduit à l'analyse spectrale de la matrice $\widetilde{\mathbf{D}_c}^{-1}$-symétrique positive :*

$$\widetilde{\mathbf{A}}\tilde{\mathbf{B}} = \left[\overline{\mathbf{A}\mathbf{B}}\right]^2.$$

qui admet pour vecteurs propres: $\widetilde{\mathbf{U}} = \overline{\mathbf{U}}$ associés aux valeurs propres $\rho_k = \nu_k^2$.

La matrice des composantes principales s'expriment par:

$$\overline{\mathbf{C}_r} = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{N}.$$

Cette dernière matrice fournit les coordonnées d'une représentation simultanée des modalités des deux variables qui sont, à une homothétie près, identiques à celles de l'A.F.C. de la table de contingence.

Remarques VII.1 *Dans les deux cas d'A.F.C. considérés : du tableau disjonctif ou du tableau de Burt, on trouve, par construction, des valeurs propres non nulles sans signification statistique. Aussi, les critères de qualité s'exprimant comme une "part de variance expliquée" n'ont plus de sens.*

Une A.F.C.M. ne prend en compte que l'information contenue dans le tableau de Burt qui ne considère que les croisements de variables deux à deux. En conséquence, les interactions de niveau plus élevé sont ignorées par cette technique à moins de procéder à des recodages de variables comme l'explique l'exemple ci-après.

4 A.F.C.M.

4.1 Définition

On considère maintenant p variables qualitatives ($p \geq 3$) notées $\{X^j \ ; j = 1, \ldots, p\}$ possédant respectivement c_j modalités avec $c = \sum_{j=1}^{p}$. On suppose que ces variables sont observées sur les mêmes n individus, chacun affecté du poids $1/n$.

Soit $\mathbf{X} = [\mathbf{X}_1|\cdots|\mathbf{X}_p]$ le tableau disjonctif complet et $\mathcal{B} = \mathbf{X}'\mathbf{X}$ le tableau de Burt correspondant.

Définition VII.1 *On appelle analyse factorielle multiple des correspondances des variables $(X^1, \ldots, X^p)$ relativement à l'échantillon, l'A.F.C. effectuée sur les matrices $\mathbf{X}$ ou $\mathcal{B}$.*

On note n_k^j $(1 \leq j \leq p, 1 \leq k \leq c_j)$ l'effectif de la $k$ième modalité de X^j, $\mathbf{D}_j = \text{diag}(n_1^j, \ldots, n_{c_j}^j)/n$ et $\mathbf{\Delta} = \text{diag}(\mathbf{D}_j \ ; j = 1, \ldots, p)$.

4.2 A.F.C. du tableau disjonctif

Les matrices usuelles de l'A.F.C. généralisent le cas précédent $(p = 2)$:

$$\overline{\mathbf{D}_r} \ = \ \frac{1}{n}\mathbf{I}_n,$$

$$\begin{aligned}\overline{\mathbf{D}_c} &= \frac{1}{p}\mathbf{\Delta},\\ \overline{\mathbf{A}} &= \frac{1}{p}\mathbf{X}',\\ \overline{\mathbf{B}} &= \frac{1}{n}\mathbf{X}\mathbf{\Delta}^{-1}.\end{aligned}$$

A.C.P. des profils-lignes

Proposition VII.4 *L'A.C.P. des profils-lignes issue de l'A.F.C. calculée sur le tableau disjonctif complet de p variables conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}_c}^{-1}$-symétrique positive :*

$$\overline{\mathbf{AB}} = \frac{1}{np}\mathcal{B}\mathbf{\Delta}^{-1}.$$

Il y a m $(m < (c-p))$ valeurs propres ν_k, $(0 < \nu_k < 1)$ rangées dans la matrice diagonale $\mathbf{N}$. La matrice des vecteurs propres associés $\overline{\mathbf{D}_c}^{-1}$-orthonormés se décompose en blocs :

$$\overline{\mathbf{U}} = \begin{bmatrix}\mathbf{U}_1\\ \vdots\\ \mathbf{U}_p\end{bmatrix}$$

avec chaque bloc $\mathbf{U}_j$ de dimensions $(c_j \times m)$.

La matrice des composantes principales s'expriment par :

$$\overline{\mathbf{C}_r} = \sum_{j=1}^{p}\mathbf{X}_j\mathbf{D}_j^{-1}\mathbf{U}_j.$$

Comme dans le cas $p = 2$, la matrice des composantes principales fournit une représentation des individus dans laquelle chacun apparaît comme le barycentre des modalités qu'il prend.

Attention *La généralisation $(p > 2)$ restreint les propriétés. Les vecteurs des blocs $\mathbf{U}_j$ ne sont pas les vecteurs propres $\mathbf{D}_j^{-1}$-orthonormés d'une matrice connue.*

A.C.P. des profils-colonnes

Proposition VII.5 *L'A.C.P. des profils-colonnes issue de l'A.F.C. calculée sur le tableau disjonctif complet de p variables conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}_r}^{-1}$-symétrique positive :*

$$\overline{\mathbf{BA}} = \frac{1}{np}\mathbf{X}\mathbf{\Delta}^{-1}\mathbf{X}' = \frac{1}{np}\sum_{j=1}p\mathbf{X}_j\mathbf{D}_j^{-1}\mathbf{X}_j'.$$

La matrice des vecteurs propres $\overline{\mathbf{D}_r}^{-1}$*-orthonormés vérifient :*

$$\overline{\mathbf{V}} = \overline{\mathbf{BU}}\mathbf{N}^{-1/2}.$$

La matrice des composantes principales se décompose en blocs :

$$\overline{\mathbf{C}_c} = p\boldsymbol{\Delta}^{-1}\overline{\mathbf{U}}\mathbf{N}^{1/2} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}.$$

Chaque bloc $\mathbf{C}_j$ contient les coordonnées pour la représentation graphique des modalité de la variable X^j.

4.3 A.F.C. du tableau de Burt

Le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$ $(c \times c)$ étant symétrique, les profils-lignes et colonnes sont identiques, il suffit de considérer une seule A.C.P.

Les matrices usuelles de l'A.F.C. deviennent dans ce cas :

$$\begin{aligned} \tilde{\mathbf{T}} &= \mathcal{B}, \\ \widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{p}\boldsymbol{\Delta}, \\ \widetilde{\mathbf{A}} &= \tilde{\mathbf{B}} = \frac{1}{np}\mathcal{B}\boldsymbol{\Delta}^{-1}. \end{aligned}$$

Proposition VII.6 *L'A.C.P. des profils-lignes (ou colonnes) issue de l'A.F.C. calculée sur le tableau de Burt associé à p variables conduit à l'analyse spectrale de la matrice* $\widetilde{\mathbf{D}_c}^{-1}$*-symétrique positive :*

$$\widetilde{\mathbf{A}}\tilde{\mathbf{B}} = \left[\overline{\mathbf{AB}}\right]^2.$$

qui admet pour vecteurs propres : $\widetilde{\mathbf{U}} = \overline{\mathbf{U}}$ *associés aux valeurs propres*

$$\rho_k = \nu_k^2.$$

La matrice des composantes principales s'expriment par :

$$\overline{\mathbf{C}_r} = .$$

Cette dernière matrice fournit les coordonnées d'une représentation simultanée des modalités de toutes les variables mais pas des individus.

4.4 Variables illustratives

Soit X^0 une variable qualitative supplémentaire à c_0 modalités observées sur les mêmes individus. Soit $\mathbf{T}_{0j}$ la table de contingence $(c_0 \times c_j)$ des variables X^0 et X^j. L'objectif est de représenter les modalités de cette variable supplémentaires dans le graphique d'une A.F.C.M. à laquelle elle n'a pas contribué. On considère les matrices :

$$\begin{aligned}
\mathbf{B}_0 &= [\mathbf{T}_{01}|\ldots|\mathbf{T}_{0p}], \\
\mathbf{D}_0 &= \frac{1}{n}\mathrm{diag}(n_1^0,\ldots,n_{c_0}^0), \\
\mathbf{A}_0 &= \frac{1}{np}\mathbf{D}_0^{-1}\mathbf{B}_0.
\end{aligned}$$

Les coordonnées des modalités de la variable supplémentaires X^0 sont fournies par la matrice :

$$\mathbf{C}_0 = \mathbf{A}_0\overline{\mathbf{D}_r}^{-1}\overline{\mathbf{U}}.$$

4.5 Interprétation

Les représentations graphiques sont interprétées de manière identique à celle de l'A.F.C. de deux variables bien que la représentation simultanée des modalités de toutes les variables ne soit pas, en toute rigueur, bien justifiée. On peux néanmoins considérer, suivant Michael et Greenacre (1988), que la représentation obtenue est similaire au "meilleur" compromis trouvé entre toutes les A.F.C. des variables croisées deux à deux.

Les "principes" suivants sont donc appliqués :

- dans la mesure du possible, essayer de caractériser ou donner un titre à chacun des premiers axes en recherchant les modalités à la plus forte contribution,

- interpréter globalement les proximités et les oppositions entre les modalités des différentes variables, comme en A.F.C., en privilégiant les modalités suffisamment éloignées du centre du graphique. *Attention aux modalités à faible effectif !*

- Les coefficients de qualité globale ou de chaque modalité ne peuvent pas être interprétés, les contributions des modalités à la dispersion totale ou selon les directions des axes sont interprétées comme en A.F.C.,

5 Exemple

5.1 Les données

La littérature anglo-américaine présente souvent des données relatives à plusieurs variables qualitatives sous la forme d'une table de contingence *complète* (5.1). L'exemple ci-dessous est extrait de Bishop et al. (1976). Il décrit les résultats partiels d'une enquête réalisée dans trois centres hospitaliers (Boston, Glamorgan, Tokio) sur des patientes atteintes d'un cancer du sein. On se propose d'étudier la survie de ces patientes trois ans après le diagnostic. En plus de cette information, quatre autres variables sont documentées pour chacune des patientes :

- le centre de diagnostic,
- la tranche d'âge,
- le degré d'inflammation chronique,
- l'apparence relative (bénigne ou maligne).

L'objectif de cette étude est une analyse descriptive de cette table en recherchant à mettre en évidence les facteurs de décès.

5.2 Analyse brute

5.3 Analyse des interactions

Le graphique de l'analyse précédente suggère une influence de l'âge mais aussi une du centre de diagnostic dans les risques de décès avant trois ans. Pour expliciter ces liaisons, les données sont reconsidérées de la façon suivante :

- les variables `centre` et `âge` sont croisées pour construire une variable `c_x_âge` à 9 modalités,
- les variables `inflam` et `appar` sont croisées également pour définir la variable `histol` à 4 modalités,

Une nouvelle analyse est calculée en considérant, comme actives, les deux variables nouvellement créées ainsi que la variable `survie` et, comme illustratives, les variables initiales : `centre, âge, inflam, appar`.

			Histologie			
			Inflammation minime		Grande inflammation	
Centre	*Age*	*Survie*	Maligne	Bénigne	Maligne	Bénigne
Tokyo	< 50	non	9	7	4	3
		oui	26	68	25	9
	$50 - 69$	non	9	9	11	2
		oui	20	46	18	5
	> 70	non	2	3	1	0
		oui	1	6	5	1
Boston	< 50	non	6	7	6	0
		oui	11	24	4	0
	$50 - 69$	non	8	20	3	2
		oui	18	58	10	3
	> 70	non	9	18	3	0
		oui	15	26	1	1
Glamorgan	< 50	non	16	7	3	0
		oui	16	20	8	1
	$50 - 69$	non	14	12	3	0
		oui	27	39	10	4
	> 70	non	3	7	3	0
		oui	12	11	4	1

TAB. VII.1 - Données sous la forme d'une table de contingence complète

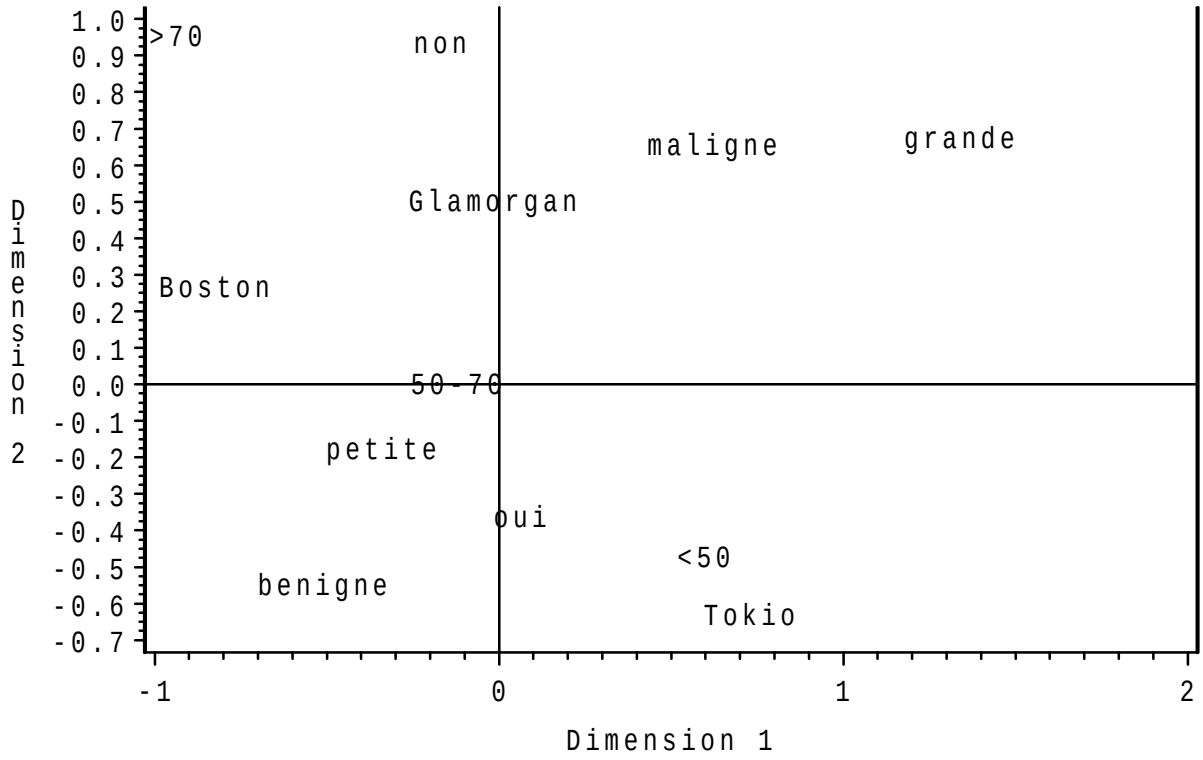


FIG. VII.1 - Cancer du sein : analyse des données brutes.

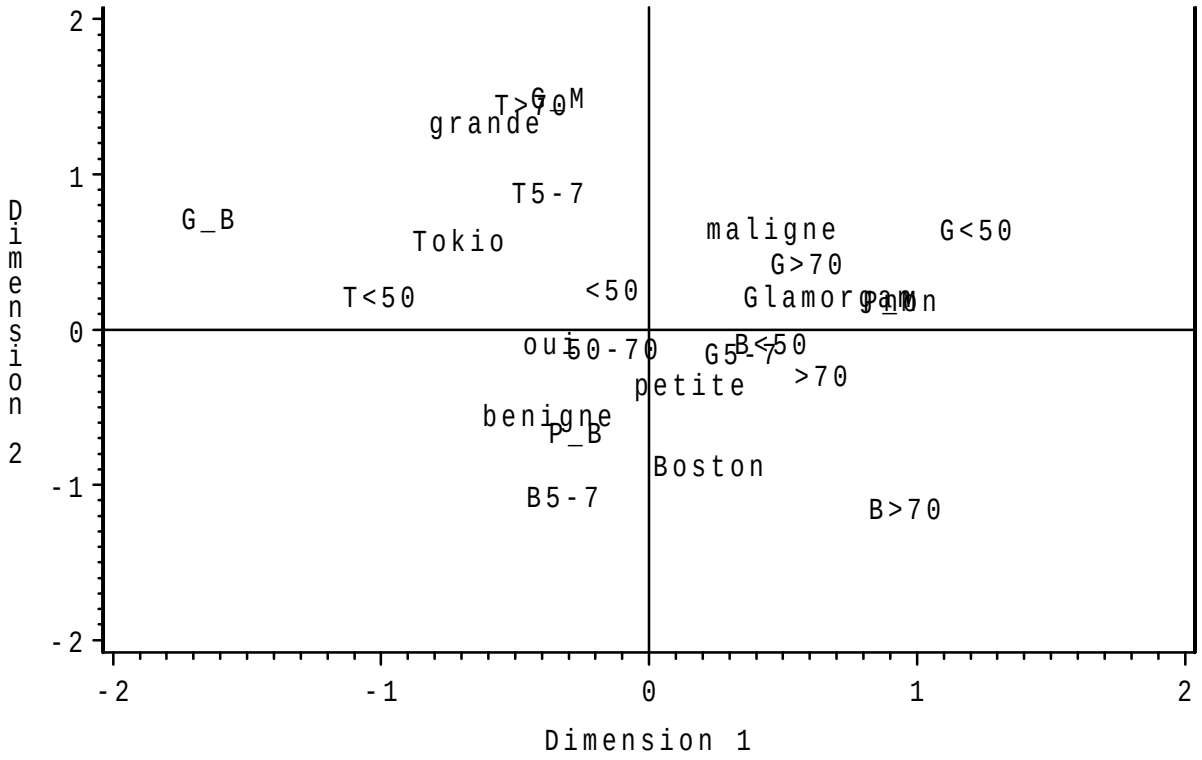


FIG. VII.2 - Cancer du sein : analyse des interactions.

6 Pratique de l'A.F.C.M.

Il n'y a pas de support théorique à une "bonne" utilisation de l'A.F.C.M. dont la mise en œuvre pratique reste très délicate et requiert beaucoup d'expérience pour espérer tirer des interprétations fiables et pertinentes des graphiques obtenus.

C'est pour l'analyse de questionnaires (sondage) que son utilisation est la plus répandue. La question qui se pose alors est la recherche d'une stratégie efficace face à un fichier de données volumineux et comportant beaucoup de variables.

Il est important que le statisticien soit impliqué le plus en amont possible dans ce type d'étude afin qu'il puisse s'assurer que les objectifs poursuivis et les données recueillies soient bien compatibles avec les potentialités des outils statistiques utilisables.

Exemple de chronologie, certains points devant être itérés : :

i. Tester la pertinence du questionnaire sur un échantillon réduit.

ii. Tirer un "bon" échantillon c'est-à-dire représentatif, pour les informations recherchées, de la population visée.

iii. Saisir les questionnaires en vérifiant la cohérence des réponses.

iv. Étude univariée (tri à plat), graphiques élémentaires, vérifications.

v. Regroupement ou élimination de modalités trop rares, élimination de variables non discriminantes ou redondantes, recodage en classes (de préférence de mêmes effectifs) des variables quantitatives.

vi. Dans le cas fréquent de variables trop nombreuses, regroupement de celles-ci par thèmes en répétant à chaque fois l'éventuelle variable "cible" à expliquer.

vii. A.F.C.M. de chaque groupe permettant la sélection des variables les plus pertinentes pour l'objectif de l'étude.

viii. A.F.C.M. de ces quelques variables importantes.

ix. Modélisation (logit, log-linéaire) dans le cas où une variable doit être expliquée.

Chapitre VIII

Bibliographie

Becker, R.A., Chambers, J.M., Wilks, A.R. (1988). *The New S Language, a Programming Environment for Data Analysis and Graphics*, Wadsworth and Brooks/Cole, Pacific Grove, Ca 93950.

Besse, P. (1992). PCA Stability and Choice of Dimensionality. *Statistics & Probability Letters*, 13, 405-410.

Besse, P., Caussinus, H., Ferré, L., Fine, J. (1988). Principal Components Analysis and Optimization of Graphical Displays. *Statistics*, 19, 301-312.

Besse, P., de Falguerolles, A. (1993). Application of resampling Methods to the Choice of Dimension in PCA. in *Computer Intensive Methods in Statistics*, W. Härdle et L. Simar (éditeurs), Physika-Verlag.

Besse, P., Pousse, A. (1992). Extension des Analyses Factorielles, in *Modèles pour l'Analyse des Données Multidimensionnelles*, J.J. Droesbeke et al. (eds.), Economica, Paris.

Bouroche,J.M. et Saporta, G. *L'analyse des données*, Que sais-je, P.U.F., 1980.

Caillez, F. et Pagès, J.M. *Introduction à l'analyse des données*, S.M.A.S.H., 1976.

Caussinus, H. (1986). Models and Uses of Principal Component Analysis, in *Multidimensional Data Analysis*, J. de Leeuw et al. (eds.), DSWO Press, Leiden, 149-170.

Caussinus, H. (1992). Projections révélatrices, in *Modèles pour l'Analyse des Données Multidimensionnelles*, J.J. Droesbeke et al. (eds.), Economica, Paris.

Droesbeke, J.J., Fichet, B., Tassi, P. (éditeurs), *Modèles pour l'analyse des données multidimensionnelles*, Economica, 1992.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Methods*, SIAM, Philadelphie.

Fine, J., Pousse, A. (1991). Asymptotic Study of the Multivariate Functional Model; Application to the Metric Choice in PCA, *Statistics*, 23, 63-83.

Johnson, ., Wichern, . (1988). *xxx*, xxx, xxx.

Jolliffe, I. (1986). *Principal Component Analysis*, Springer-Verlag, New-York.

Kato, T. (1966). *Perturbation Theory for Linear Operator*, Springer-Verlag, New-York.

Mardia K.V., Kent J.T., Bibby J.M. *Multivariate analysis*, Academic Press, 1979.

McDonald, G.C., Schwing, R.C. (1973). Instabilities of Regression Estimates Relating Air Pollution to Mortality. *Technometrics*, 15, 463-481.

Saporta G. *Probabilités, analyse des données et statistique*, Technip, 1990.

SAS (1989), SAS/STAT User's Guide, volume 2, Version 6, fourth edition, Sas Institute Inc, Cary.

Annexe A

Outils algèbriques

Ce chapitre se propose de rassembler des notations et rappels d'algèbre linéaire ainsi que quelques compléments mathématiques du niveau du premier cycle des Universités.

Dans tout ce qui suit, E et F sont deux espaces vectoriels réels munis respectivement des bases canoniques $\mathcal{E} = \{e_j \ ; \ j = 1, \ldots, p\}$ et $\mathcal{F} = \{f_i \ ; \ i = 1, \ldots, n\}$. On note indifféremment soit un vecteur de E ou de F, un endomorphisme de E, ou une application linéaire de E dans F, soit leurs représentations matricielles dans les bases définies ci-dessus.

1 Matrices

1.1 Notations

La matrice d'ordre $(n \times p)$ associée à une application linéaire de E dans F est décrite par un tableau :

$$\mathbf{A} = \begin{bmatrix} a_1^1 & \ldots & a_1^j & \ldots & a_1^p \\ \vdots & & \vdots & & \vdots \\ a_i^1 & \ldots & a_i^j & \ldots & a_i^p \\ \vdots & & \vdots & & \vdots \\ a_n^1 & \ldots & a_n^j & \ldots & a_n^p \end{bmatrix}.$$

On note par la suite :

$$\begin{aligned} a_i^j &= [\mathbf{A}]_i^j \text{ le terme général de la matrice,} \\ a_i &= [a_i^1, \ldots, a_i^p]' \text{ un vecteur-ligne mis en colonne,} \\ a^j &= [a_1^j, \ldots, a_n^j]' \text{ un vecteur-colonne.} \end{aligned}$$

Types de matrices

Une matrice est dite :

- *vecteur*-ligne (colonne) si $n = 1$ ($p = 1$),
- *vecteur*-unité d'ordre p si elle vaut $\mathbf{1}_p = [1, \ldots, 1]'$,
- *scalaire* si $n = 1$ et $p = 1$,
- *carrée* si $n = p$.

Une matrice carrée est dite :

- *identité* ($\mathbf{I}_p$) si $a_i^j = \delta_i^j = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$,
- *diagonale* si $a_i^j = 0$ lorsque $i \neq j$,
- *symétrique* si $a_i^j = a_j^i, \forall (i, j)$,
- *triangulaire* supérieure (inférieure) si

$$a_i^j = 0 \text{ lorsque } i > j \ (i < j).$$

Matrice partitionnée en blocs

Matrices dont les éléments sont eux-mêmes des matrices.
Exemple : $\mathbf{A}(n \times p) = \begin{bmatrix} \mathbf{A}_1^1(r \times s) & \mathbf{A}_1^2(r \times (p-s)) \\ \mathbf{A}_2^1((n-r) \times s) & \mathbf{A}_2^2((n-r) \times (p-s)) \end{bmatrix}$.

1.2 Opérations sur les matrices

Somme

$$[\mathbf{A} + \mathbf{B}]_i^j = a_i^j + b_i^j \text{ pour } \mathbf{A} \text{ et } \mathbf{B} \text{ de même ordre } (n \times p).$$

Multiplication par un scalaire

$$[\alpha \mathbf{A}]_i^j = \alpha a_i^j \text{ pour } \alpha \in \mathbf{R}.$$

Transposition

$$[\mathbf{A}']_i^j = a_j^i, \mathbf{A}' \text{ est d'ordre } (p \times n).$$
$$(\mathbf{A}')' = \mathbf{A} \ ; (\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}' \ ; (\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \ ;$$
$$\begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix}' = \begin{bmatrix} \mathbf{A}_1^{1'} & \mathbf{A}_2^{1'} \\ \mathbf{A}_1^{2'} & \mathbf{A}_2^{2'} \end{bmatrix}.$$

Produit scalaire élémentaire

$$a'b = \sum_{i=1}^{n} a_i b_i \text{ où } a \text{ et } b \text{ sont des vecteurs-colonnes.}$$

Produit

$$[\mathbf{AB}]_i^j = a_i' b^j \text{ avec } \mathbf{A}_{(n\times p)}, \mathbf{B}_{(p\times q)} \text{ et } \mathbf{AB}_{(n\times q)},$$

et pour des matrices par blocs :

$$\begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1^1 & \mathbf{B}_1^2 \\ \mathbf{B}_2^1 & \mathbf{B}_2^2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1^1\mathbf{B}_1^1 + \mathbf{A}_1^2\mathbf{B}_2^1 & \mathbf{A}_1^1\mathbf{B}_1^2 + \mathbf{A}_1^2\mathbf{B}_2^2 \\ \mathbf{A}_2^1\mathbf{B}_1^1 + \mathbf{A}_2^2\mathbf{B}_2^1 & \mathbf{A}_2^1\mathbf{B}_1^2 + \mathbf{A}_2^2\mathbf{B}_2^2 \end{bmatrix}$$

sous réserve de compatibilité des dimensions.

1.3 Propriétés des matrices carrées

La *trace* et le *déterminant* sont des notions intrinsèques, qui ne dépendent pas des bases de représentation choisies, mais uniquement de l'application linéaire sous-jacente.

Trace

Par définition, si $\mathbf{A}$ est une matrice $(p \times p)$,

$$\mathrm{tr}\mathbf{A} = \sum_{j=1}^{p} a_j^j,$$

et il est facile de montrer :

$$\begin{aligned}
\mathrm{tr}\alpha &= \alpha, \\
\mathrm{tr}\alpha\mathbf{A} &= \alpha \mathrm{tr}\mathbf{A}, \\
\mathrm{tr}(\mathbf{A}+\mathbf{B}) &= \mathrm{tr}\mathbf{A} + \mathrm{tr}\mathbf{B}, \\
\mathrm{tr}\mathbf{AB} &= \mathrm{tr}\mathbf{BA}, \\
&\quad \text{reste vrai si } \mathbf{A} \text{ est } (n\times p) \text{ et si } \mathbf{B} \text{ est } (p\times n) \\
\mathrm{tr}\mathbf{CC}' &= \mathrm{tr}\mathbf{C}'\mathbf{C} = \sum_{i=1}^{n}\sum_{j=1}^{p}(c_i^j)^2 \\
&\quad \text{dans ce cas, } \mathbf{C} \text{ est } (n\times p).
\end{aligned}$$

Déterminant

On note $|\mathbf{A}|$ le *déterminant* de la matrice carrée $\mathbf{A}$ $(p \times p)$. Il vérifie :

$$\begin{aligned}
|\mathbf{A}| &= \prod_{j=1}^{p} a_j^j, \text{ si } \mathbf{A} \text{ est triangulaire ou diagonale,} \\
|\alpha\mathbf{A}| &= \alpha^p |\mathbf{A}|, \\
|\mathbf{AB}| &= |\mathbf{A}||\mathbf{B}|, \\
\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{vmatrix} &= |\mathbf{A}||\mathbf{C}|, \\
\begin{vmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{vmatrix} &= |\mathbf{A}_1^1||\mathbf{A}_2^2 - \mathbf{A}_2^1(\mathbf{A}_1^1)^{-1}\mathbf{A}_1^2| \qquad \text{(A.1)} \\
&= |\mathbf{A}_2^2||\mathbf{A}_1^1 - \mathbf{A}_1^2(\mathbf{A}_2^2)^{-1}\mathbf{A}_2^1|, \qquad \text{(A.2)} \\
&\quad \text{sous réserve de la régularité de } \mathbf{A}_1^1 \text{ et } \mathbf{A}_2^2.
\end{aligned}$$

Cette dernière propriété se montre en considérant les matrices :

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} & -\mathbf{A}_1^2(\mathbf{A}_2^2)^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \text{ et } \mathbf{BAB}',$$

puis en comparant les déterminants $|\mathbf{BAB}'|$ et $|\mathbf{A}|$.

Inverse

L'*inverse* de $\mathbf{A}$, lorsqu'elle existe, est la matrice unique notée $\mathbf{A}^{-1}$ telle que :

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} ;$$

elle existe si et seulement si $|\mathbf{A}| \neq 0$.

Quelques propriétés :

$$\begin{aligned}
(\mathbf{A}^{-1})' &= (\mathbf{A}')^{-1}, \\
(\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1}, \\
|\mathbf{A}^{-1}| &= \frac{1}{|\mathbf{A}|}.
\end{aligned}$$

Définitions

Une matrice carrée $\mathbf{A}$ est dite :

symétrique si $\mathbf{A}' = \mathbf{A}$,

singulière si $|\mathbf{A}| = 0$,

régulière si $|\mathbf{A}| \neq 0$,

idempotente si $\mathbf{AA} = \mathbf{A}$,

définie-positive si, $\forall x \in \mathbb{R}^p, x'\mathbf{A}x \geq 0$, et si $x'\mathbf{A}x = 0 \Rightarrow x = 0$,

positive, ou semi-définie-positive, si, $\forall x \in \mathbb{R}^p, x'\mathbf{A}x \geq 0$,

orthogonale si $\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \mathbf{I}$ $(\mathbf{A}' = \mathbf{A}^{-1})$.

2 Espaces euclidiens

E est un espace vectoriel réel de dimension p isomorphe à $\mathbb{R}^p$.

2.1 Sous-espaces

- Un sous-ensemble E_q de E est un *sous-espace vectoriel* (s.e.v.) de E s'il est non vide et stable :

$$\forall (x, y) \in E_q^2, \forall \alpha \in \mathbb{R}, \alpha(x + y) \in E_q.$$

- Le q-uple $\{x_1, \ldots, x_q\}$ de E constitue un système *linéairement indépendant* si et seulement si :

$$\sum_{i=1}^{q} \alpha_i x_i = 0 \Rightarrow \alpha_1 = \cdots = \alpha_q = 0.$$

- Un système linéairement indépendant $\mathcal{E}_q = \{e_1, \ldots, e_q\}$ qui engendre dans E un s.e.v. $E_q = \text{vec}\{e_1, \ldots, e_q\}$ en constitue une *base* et $\dim(E_q) = \text{card}(\mathcal{E}_q) = q$.

2.2 Rang d'une matrice $\mathbf{A}_{(n\times p)}$

Image et noyau

Dans ce sous-paragraphe, $\mathbf{A}$ est la matrice d'une application linéaire de $E = \mathbb{R}^p$ dans $F = \mathbb{R}^n$.

$$\begin{aligned}
\text{Im}(\mathbf{A}) &= \text{vect}\{a^1, \ldots, a^p\} \text{ est le s.e.v. de } F \textit{ image} \text{ de } \mathbf{A} ; \\
\text{Ker}(\mathbf{A}) &= \{x \in E \ ; \ \mathbf{A}x = 0\} \text{ est le s.e.v. de } E \textit{ noyau} \text{ de } \mathbf{A} ; \\
E &= \text{Im}(\mathbf{A}) \oplus \text{Ker}(\mathbf{A}) \text{ si } \mathbf{A} \text{ est carrée associée à un endomorphisme de } E \\
\text{et } p &= \dim(\text{Im}(\mathbf{A})) + \dim(\text{Ker}(\mathbf{A})).
\end{aligned}$$

Rang

$$\begin{aligned}
\text{rang}(\mathbf{A}) &= \dim(\text{Im}(\mathbf{A})),\\
0 \leq \text{rang}(\mathbf{A}) &\leq \min(n,p),\\
\text{rang}(\mathbf{A}) &= \text{rang}(\mathbf{A}'),\\
\text{rang}(\mathbf{A}+\mathbf{B}) &\leq \text{rang}(\mathbf{A}) + \text{rang}(\mathbf{B}),\\
\text{rang}(\mathbf{AB}) &\leq \min(\text{rang}(\mathbf{A}), \text{rang}(\mathbf{B})),\\
\text{rang}(\mathbf{BAC}) &= \text{rang}(\mathbf{A}), \text{ si } \mathbf{B} \text{ et } \mathbf{C} \text{ sont régulières},\\
\text{rang}(\mathbf{A}) &= \text{rang}(\mathbf{AA}') = \text{rang}(\mathbf{A}'\mathbf{A}).
\end{aligned}$$

Enfin, si $\mathbf{B}$ $(p \times q)$ est de rang $q (q < p)$ et $\mathbf{A}$ est carrée $(p \times p)$ de rang p, alors la matrice $\mathbf{B'AB}$ est de rang q.

2.3 Métrique euclidienne

Soit $\mathbf{M}$ une matrice carrée $(p \times p)$, symétrique, définie-positive ; $\mathbf{M}$ définit sur l'espace E :

- un *produit scalaire* : $\langle x, y \rangle_{\mathbf{M}} = x'\mathbf{M}y$,
- une *norme* : $\|x\|_{\mathbf{M}} = \langle x, x \rangle_{\mathbf{M}}^{1/2}$,
- une *distance* : $\mathrm{d}_{\mathbf{M}}(x, y) = \|x - y\|_{\mathbf{M}}$,
- des *angles* : $\cos\theta_{\mathbf{M}}(x, y) = \frac{\langle x,y \rangle_{\mathbf{M}}}{\|x\|_{\mathbf{M}}\|y\|_{\mathbf{M}}}$.

La matrice $\mathbf{M}$ étant donnée, on dit que :

- une matrice $\mathbf{A}$ est $\mathbf{M}$-*symétrique* si $(\mathbf{MA})' = \mathbf{MA}$,
- deux vecteurs x et y sont $\mathbf{M}$-*orthogonaux* si $\langle x, y \rangle_{\mathbf{M}} = 0$,
- un vecteur x est $\mathbf{M}$-*normé* si $\|x\|_{\mathbf{M}} = 1$,
- une base $\mathcal{E}_q = \{e_1, \ldots, e_q\}$ est $\mathbf{M}$-*orthonormée* si

$$\forall (i, j), \langle e_i, e_j \rangle_{\mathbf{M}} = \delta_i^j.$$

2.4 Projection

Soit W un sous-espace de E et $\mathcal{B} = \{b^1, \ldots, b^q\}$ une base de W ; $\mathbf{P}$ $(p \times p)$ est une matrice de projection $\mathbf{M}$-orthogonale sur W si et seulement si :

$$\forall y \in E, \mathbf{P}y \in W \text{ et } \langle \mathbf{P}y, y - \mathbf{P}y \rangle_{\mathbf{M}} = 0.$$

Toute matrice idempotente ($\mathbf{P}^2 = \mathbf{P}$) et $\mathbf{M}$-symétrique ($\mathbf{P}'\mathbf{M} = \mathbf{MP}$) est une matrice de projection $\mathbf{M}$-orthogonale et réciproquement.

Propriétés

- Les valeurs propres de $\mathbf{P}$ sont 0 ou 1 (voir § 3) :

$$\begin{array}{ll} u \in W, & \mathbf{P}u = u, \quad \lambda = 1, \text{ de multiplicité } \dim(W), \\ v \perp W, (\text{on note } v \in W^{\perp}) & \mathbf{P}v = 0, \quad \lambda = 0, \text{ de multiplicité } \dim(W^{\perp}). \end{array}$$

- $\mathrm{tr}\mathbf{P} = \dim(W)$.

- $\mathbf{P} = \mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-1}\mathbf{B}'\mathbf{M}$, où $\mathbf{B} = [b^1, \ldots, b^q]$.

- Dans le cas particulier où les b^j sont $\mathbf{M}$-orthonormés :

$$\mathbf{P} = \mathbf{B}\mathbf{B}'\mathbf{M} = \sum_{i=1}^{q} b^j b^{j\prime}\mathbf{M}.$$

- Dans le cas particulier où $q = 1$ alors :

$$\mathbf{P} = \frac{bb'}{b'\mathbf{M}b}\mathbf{M} = \frac{1}{\|b\|_{\mathbf{M}}}bb'\mathbf{M}.$$

- Si $\mathbf{P}_1, \ldots, \mathbf{P}_q$ sont des matrices de projection $\mathbf{M}$-orthogonales alors la somme $\mathbf{P}_1 + \cdots + \mathbf{P}_q$ est une matrice de projection $\mathbf{M}$-orthogonale si et seulement si : $\mathbf{P}_k\mathbf{P}_j = \delta_k^j\mathbf{P}_j$.

- La matrice $\mathbf{I} - \mathbf{P}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $W^{\perp}$.

3 Eléments propres

Soit $\mathbf{A}$ une matrice carrée $(p \times p)$.

3.1 Définitions

- Par définition, un vecteur v définit une *direction propre* associée à une *valeur propre* λ si l'on a :

$$\mathbf{A}v = \lambda v.$$

- Si λ est une valeur propre de $\mathbf{A}$, le noyau $\mathrm{Ker}(\mathbf{A} - \lambda\mathbf{I})$ est un s.e.v. de E, appelé sous-espace propre, dont la dimension est majoré par l'ordre de multiplicité de λ. Comme cas particulier, $\mathrm{Ker}(\mathbf{A})$ est le sous-espace propre associé, si elle existe, à la valeur propre nulle.

- Les valeurs propres d'une matrice $\mathbf{A}$ sont les racines, avec leur multiplicité, du *polynôme caractéristique* :

$$|\mathbf{A} - \lambda\mathbf{I}| = 0.$$

Théorème A.1 *Soit deux matrices* $\mathbf{A}(n \times p)$ *et* $\mathbf{B}(p \times n)$ *; les valeurs propres non nulles de* $\mathbf{AB}$ *et* $\mathbf{BA}$ *sont identiques avec le même degré de multiplicité. Si* u *est vecteur propre de* $\mathbf{BA}$ *associé à la valeur propre* λ *différente de zéro, alors* $v = \mathbf{A}u$ *est vecteur propre de la matrice* $\mathbf{AB}$ *associé à la même valeur propre.*

Les applications statistiques envisagées dans ce cours ne s'intéressent qu'à des types particuliers de matrices.

Théorème A.2 *Une matrice* $\mathbf{A}$ *réelle symétrique admet p valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base orthonormée de* E *;* $\mathbf{A}$ *se décompose en :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' = \sum_{k=1}^{p} \lambda_k v^k v^{k'}$$

où $\mathbf{V}$ *est une matrice orthogonale* $[v^1, \ldots, v^p]$ *des vecteurs propres orthonormés associés aux valeurs propres* λ_k*, rangées par ordre décroissant dans la matrice diagonale* $\mathbf{\Lambda}$*.*

Théorème A.3 *Une matrice* $\mathbf{A}$ réelle $\mathbf{M}$-symétrique *admet p valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base* $\mathbf{M}$*-orthonormée de* E *;* $\mathbf{A}$ *se décompose en :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'\mathbf{M} = \sum_{k=1}^{p} \lambda_k v^k v^{k'}\mathbf{M}$$

où $\mathbf{V} = [v^1, \ldots, v^p]$ *est une matrice* $\mathbf{M}$*-orthogonale* ($\mathbf{V}'\mathbf{M}\mathbf{V} = \mathbf{I}_p$ *et* $\mathbf{V}\mathbf{V}' = \mathbf{M}^{-1}$) *des vecteurs propres associés aux valeurs propres* λ_k*, rangées par ordre décroissant dans la matrice diagonale* $\mathbf{\Lambda}$*.*

Les décompositions ne sont pas uniques : pour une valeur propre simple (de multiplicité 1) le vecteur propre normé est défini à un signe près, tandis que pour une valeur propre multiple, une infinité de bases $\mathbf{M}$-orthonormées peuvent être extraites du sous-espace propre unique associé.

Le rang de $\mathbf{A}$ est aussi le rang de la matrice $\mathbf{\Lambda}$ associée et donc le nombre (répétées avec leurs multiplicités) de valeurs propres non nulles.

Par définition, si $\mathbf{A}$ est positive, on note la racine carrée de $\mathbf{A}$:

$$\mathbf{A}^{1/2} = \sum_{k=1}^{p} \sqrt{\lambda_k} v^k v^{k'}\mathbf{M} = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{V}'\mathbf{M}.$$

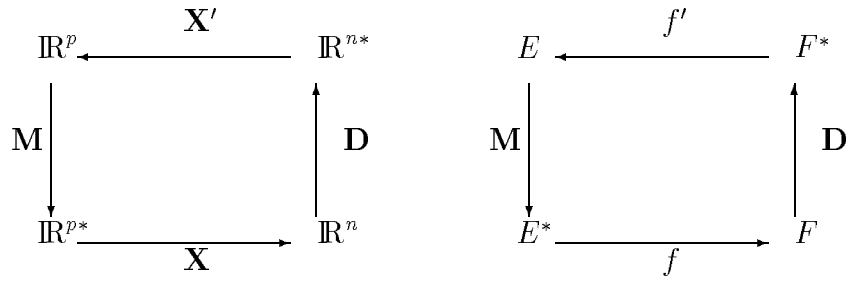


FIG. A.1 - *Schéma de dualité*

3.2 Propriétés

Si $\lambda_k \neq \lambda_j$,	$v^k \perp_{\mathbf{M}} v^j$;
$\mathrm{tr}\mathbf{A} = \sum_{k=1}^{p} \lambda_k$;	$\lvert\mathbf{A}\rvert = \prod_{k=1}^{p} \lambda_k$;
si $\mathbf{A}$ est régulière,	$\forall k, \lambda_k \neq 0$;
si $\mathbf{A}$ est positive,	$\lambda_p \geq 0$;
si $\mathbf{A}$ est définie-positive,	$\lambda_p > 0$;

3.3 Décomposition en Valeurs Singulières (DVS)

Il s'agit, cette fois, de construire la décomposition d'une matrice $\mathbf{X}(n \times p)$ rectangulaire relativement à deux matrices symétriques et positives $\mathbf{D}(n \times n)$ et $\mathbf{M}(p \times p)$.

Théorème A.4 *Une matrice* $\mathbf{X}$ $(n \times p)$ *de rang* r *peut s'écrire :*

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}' = \sum_{k=1}^{r} \sqrt{\lambda_k} u^k v^{k'}; \tag{A.3}$$

$\mathbf{U}$ $(n \times r)$ *contient les vecteurs propres* $\mathbf{D}$*-orthonormés* $(\mathbf{U}'\mathbf{D}\mathbf{U} = \mathbf{I}_r)$ *de la matrice* $\mathbf{D}$*-symétrique positive* $\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}$ *associés aux* r *valeurs propres non nulles* λ_k *rangées par ordre décroissant dans la matrice diagonale* $\mathbf{\Lambda}(r \times r)$ *;* $\mathbf{V}$ $(p \times r)$ *contient les vecteurs propres* $\mathbf{M}$*-orthonormés* $(\mathbf{V}'\mathbf{M}\mathbf{V} = \mathbf{I}_r)$ *de la matrice* $\mathbf{M}$*-symétrique positive* $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}$ *associés aux mêmes valeurs propres. De plus,*

$$\mathbf{U} = \mathbf{X}\mathbf{M}\mathbf{V}\mathbf{\Lambda}^{-1/2} \text{ et } \mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{U}\mathbf{\Lambda}^{-1/2}.$$

4 Dualité

Les éléments précédents se résument dans le diagramme commutatif de la figure (A.1).

5 Optimisation

5.1 Norme d'une matrice

L'espace vectoriel E de dimension p (resp. F de dimension n) est muni de sa base canonique et d'une métrique de matrice $\mathbf{M}$ (resp. $\mathbf{D}$). Soit $\mathbf{X}$ une matrice $(n \times p)$. L'ensemble $\mathcal{M}_{n,p}$ des matrices $(n \times p)$ est un espace vectoriel de dimension np; on le munit du *produit scalaire* :

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{M},\mathbf{D}} = \mathrm{tr}\mathbf{X}\mathbf{M}\mathbf{Y}'\mathbf{D}. \tag{A.4}$$

Dans le cas particulier où $\mathbf{M} = \mathbf{I}_p$ et $\mathbf{D} = \mathbf{I}_n$, et en notant $\mathrm{vec}(\mathbf{X}) = \left[x^{1\prime}, \ldots, x^{p\prime}\right]'$ la matrice "vectorisée", ce produit scalaire devient :

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{I_p},\mathbf{I_n}} = \mathrm{tr}\mathbf{X}\mathbf{Y}' = \sum_{i=1}^{n}\sum_{j=1}^{p} x_i^j y_i^j = \mathrm{vec}(\mathbf{X})'\mathrm{vec}(\mathbf{Y}).$$

La *norme* associée à ce produit scalaire (A.4) est appelée norme trace :

$$\begin{aligned} \|\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} &= \mathrm{tr}\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}, \\ \|\mathbf{X}\|^2_{\mathbf{I_p},\mathbf{I_n}} &= \mathrm{tr}\mathbf{X}\mathbf{X}' = \mathrm{SSQ}(\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_i^j)^2 \\ &\qquad \text{(SSQ signifie "sum of squares").} \end{aligned}$$

La *distance* associée à cette norme devient, dans le cas où $\mathbf{D}$ est une matrice diagonale ($\mathbf{D} = \mathrm{diag}(w_1, \ldots, w_n)$), le critère usuel des *moindres carrés* :

$$\mathrm{d}^2(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|^2_{\mathbf{M},\mathbf{D}} = \sum_{i=1}^{n} w_i \|x_i - y_i\|^2_{\mathbf{M}}.$$

5.2 Approximation d'une matrice

Les matrices $\mathbf{X}$, $\mathbf{M}$ et $\mathbf{D}$ sont définies comme ci-dessus ; $\mathbf{X}$ est supposée de rang r. On cherche la matrice $\mathbf{Z}_q$, de rang q inférieur à r, qui soit la plus proche possible de $\mathbf{X}$.

Théorème A.5 *La solution du problème :*

$$\min_{\mathbf{Z}} \left\{ \|\mathbf{X} - \mathbf{Z}\|^2_{\mathbf{M},\mathbf{D}} \,;\, \mathbf{Z} \in \mathcal{M}_{n,p},\ \mathit{rang}(\mathbf{Z}) = q < r \right\} \tag{A.5}$$

est donnée par la somme des q premiers termes de la décomposition en valeurs singulières (A.3) de $\mathbf{X}$:

$$\mathbf{Z}_q = \sum_{k=1}^{q} \sqrt{\lambda_k} u^k v^{k\prime} = \mathbf{U}_q \Lambda_q^{1/2} \mathbf{V}_q'.$$

Le minimum atteint est :

$$\|\mathbf{X} - \mathbf{Z}_q\|^2_{\mathbf{M},\mathbf{D}} = \sum_{k=q+1}^{r} \lambda_k.$$

Les matrices $\mathbf{U}_q, \mathbf{\Lambda}_q$ et $\mathbf{V}_q$ contiennent les q premiers vecteurs et valeurs propres donnés par la DVS de $\mathbf{X}$; $\mathbf{Z}_q$ est appelée approximation de rang q de $\mathbf{X}$.

Ce théorème peut se re-formuler d'une manière équivalente. On note $\widehat{\mathbf{P}_q}$ (resp. $\widehat{\mathbf{Q}_q}$) la projection $\mathbf{M}$-orthogonale sur $E_q = \mathrm{Im}(\mathbf{V}_q)$ (resp. $\mathbf{D}$-orthogonale sur $F_q = \mathrm{Im}(\mathbf{U}_q)$) :

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \sum_{k=1}^{q} v^k v^{k\prime}\mathbf{M} = \mathbf{V}_q\mathbf{V}_q'\mathbf{M} \\
\widehat{\mathbf{Q}_q} &= \sum_{k=1}^{q} u^k u^{k\prime}\mathbf{D} = \mathbf{U}_q\mathbf{U}_q'\mathbf{D}, \\
\mathbf{Z}_q &= \widehat{\mathbf{Q}_q}\mathbf{X} = \mathbf{X}\widehat{\mathbf{P}_q}'.
\end{aligned}$$

Proposition A.6 *Avec les notations précédentes :*

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \arg\max_{\mathbf{P}_q} \left\{ \left\|\mathbf{X}\mathbf{P}_q'\right\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\qquad \left. \mathbf{P}_q \text{ projection } \mathbf{M}\text{-orthogonale de rang } q < r \right\}, \\
\widehat{\mathbf{Q}_q} &= \arg\max_{\mathbf{Q}_q} \left\{ \|\mathbf{Q}_q\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\qquad \left. \mathbf{Q}_q \text{ projection } \mathbf{D}\text{-orthogonale de rang } q < r \right\}.
\end{aligned}$$

Annexe B

Sorties numériques

1 A.C.P. des températures

A.c.p. des donnees de temp

Statistiques elementaires

_ TYPE_	_ NAME_	JANV	FEVR	MARS	AVRI	MAI	JUIN	JUIL	AOUT	SEPT	OCT	NOV	DEC
MEAN		3.93	4.79	8.13	10.88	14.35	17.73	19.75	19.45	16.9	12.27	7.87	4.78
N		32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.0	32.00	32.00	32.00

Matrice des covariances ou des correlations

NAME	JANV	FEVR	MARS	AVRI	MAI	JUIN	JUIL	AOUT	SEPT	OCT	NOV	DEC
JANV	5.29	4.98	3.73	2.72	2.04	2.33	2.59	2.93	3.59	4.46	4.83	5.33
FEVR	4.98	4.78	3.70	2.81	2.23	2.59	2.91	3.19	3.73	4.42	4.66	5.02
MARS	3.73	3.70	3.06	2.45	2.07	2.43	2.76	2.91	3.20	3.55	3.61	3.76
AVRI	2.72	2.81	2.45	2.18	2.01	2.39	2.74	2.77	2.82	2.89	2.79	2.77
MAI	2.04	2.23	2.07	2.01	2.01	2.40	2.80	2.74	2.63	2.49	2.28	2.10
JUIN	2.33	2.59	2.43	2.39	2.40	2.93	3.41	3.32	3.14	2.93	2.65	2.41
JUIL	2.59	2.91	2.76	2.74	2.80	3.41	4.04	3.91	3.66	3.37	3.00	2.67
AOUT	2.93	3.19	2.91	2.77	2.74	3.32	3.91	3.86	3.73	3.56	3.26	3.01
SEPT	3.59	3.73	3.20	2.82	2.63	3.14	3.66	3.73	3.81	3.91	3.75	3.67
OCT	4.46	4.42	3.55	2.89	2.49	2.93	3.37	3.56	3.91	4.35	4.39	4.53
NOV	4.83	4.66	3.61	2.79	2.28	2.65	3.00	3.26	3.75	4.39	4.62	4.92
DEC	5.33	5.02	3.76	2.77	2.10	2.41	2.67	3.01	3.67	4.53	4.92	5.43

Valeurs propres, variances expliquees

K	LAMBDA	PCTVAR	CUMPCT
1	40.37	0.87	0.87
2	5.60	0.12	0.99
3	0.18	0.00	0.99
4	0.12	0.00	1.00
5	0.04	0.00	1.00
6	0.02	0.00	1.00
7	0.02	0.00	1.00
8	0.02	0.00	1.00
9	0.01	0.00	1.00
10	0.01	0.00	1.00
11	0.00	0.00	1.00
12	0.00	0.00	1.00

Vecteurs propres = coordonnees des variables du biplot

NAME	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
JANV	0.33	-0.39	0.10	0.12	-0.37	0.40	0.03	0.01	0.57	0.28	0.05	0.10
FEVR	0.33	-0.26	0.11	-0.14	-0.26	0.36	-0.12	-0.03	-0.75	-0.03	-0.13	0.02
MARS	0.27	-0.05	0.59	-0.46	-0.06	-0.50	-0.12	0.11	0.07	-0.01	0.19	0.18
AVRI	0.22	0.13	0.47	0.07	0.26	0.17	0.25	0.03	0.13	-0.23	-0.47	-0.51
MAI	0.20	0.27	0.22	0.24	0.50	0.33	0.14	0.25	-0.13	0.21	0.35	0.40
JUIN	0.23	0.36	0.13	0.34	-0.06	0.00	-0.56	-0.57	0.08	-0.15	0.01	0.10
JUIL	0.27	0.45	-0.16	0.17	-0.46	-0.08	-0.06	0.56	-0.02	-0.13	0.20	-0.27
AOUT	0.28	0.36	-0.19	-0.09	-0.20	-0.17	0.46	-0.17	-0.01	0.16	-0.47	0.43
SEPT	0.30	0.18	-0.24	-0.36	0.16	0.03	0.08	-0.34	-0.02	0.46	0.32	-0.48
OCT	0.33	-0.05	-0.37	-0.37	0.23	0.20	-0.01	0.01	0.21	-0.66	0.11	0.16
NOV	0.33	-0.21	-0.29	0.09	0.36	-0.25	-0.46	0.33	0.03	0.28	-0.41	-0.01
DEC	0.34	-0.38	-0.08	0.50	0.06	-0.42	0.37	-0.18	-0.14	-0.18	0.26	-0.09

Coordonnees des variables

NAME	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
JANV	2.10	-0.92	0.04	0.04	-0.08	0.06	0.00	0.00	0.05	0.02	0.00	0.00
FEVR	2.10	-0.60	0.04	-0.05	-0.06	0.06	-0.02	0.00	-0.07	0.00	-0.01	0.00
MARS	1.72	-0.12	0.25	-0.16	-0.01	-0.08	-0.02	0.01	0.01	0.00	0.01	0.01
AVRI	1.43	0.32	0.20	0.02	0.06	0.03	0.03	0.00	0.01	-0.02	-0.03	-0.02
MAI	1.25	0.64	0.09	0.08	0.11	0.05	0.02	0.03	-0.01	0.02	0.02	0.01
JUIN	1.48	0.85	0.06	0.12	-0.01	0.00	-0.08	-0.07	0.01	-0.01	0.00	0.00
JUIL	1.69	1.07	-0.07	0.06	-0.10	-0.01	-0.01	0.07	0.00	-0.01	0.01	-0.01
AOUT	1.77	0.85	-0.08	-0.03	-0.04	-0.03	0.06	-0.02	0.00	0.01	-0.03	0.01
SEPT	1.90	0.42	-0.10	-0.12	0.03	0.00	0.01	-0.04	0.00	0.03	0.02	-0.02
OCT	2.07	-0.13	-0.16	-0.13	0.05	0.03	0.00	0.00	0.02	-0.05	0.01	0.01
NOV	2.08	-0.50	-0.12	0.03	0.08	-0.04	-0.06	0.04	0.00	0.02	-0.02	0.00
DEC	2.14	-0.90	-0.03	0.17	0.01	-0.07	0.05	-0.02	-0.01	-0.01	0.02	0.00

Correlations variables x facteurs

NAME	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
JANV	0.91	-0.40	0.02	0.02	-0.03	0.03	0.00	0.00	0.02	0.01	0.00	0.00
FEVR	0.96	-0.28	0.02	-0.02	-0.03	0.03	-0.01	0.00	-0.03	0.00	0.00	0.00
MARS	0.98	-0.07	0.14	-0.09	-0.01	-0.05	-0.01	0.01	0.00	0.00	0.01	0.00
AVRI	0.97	0.22	0.13	0.02	0.04	0.02	0.02	0.00	0.01	-0.01	-0.02	-0.01
MAI	0.88	0.45	0.07	0.06	0.08	0.04	0.01	0.02	-0.01	0.01	0.01	0.01
JUIN	0.86	0.50	0.03	0.07	-0.01	0.00	-0.04	-0.04	0.00	-0.01	0.00	0.00
JUIL	0.84	0.53	-0.03	0.03	-0.05	-0.01	0.00	0.03	0.00	0.00	0.01	0.00
AOUT	0.90	0.43	-0.04	-0.02	-0.02	-0.01	0.03	-0.01	0.00	0.01	-0.01	0.01
SEPT	0.97	0.22	-0.05	-0.06	0.02	0.00	0.01	-0.02	0.00	0.02	0.01	-0.01
OCT	0.99	-0.06	-0.07	-0.06	0.02	0.02	0.00	0.00	0.01	-0.02	0.00	0.00
NOV	0.97	-0.23	-0.06	0.01	0.04	-0.02	-0.03	0.02	0.00	0.01	-0.01	0.00
DEC	0.92	-0.39	-0.01	0.07	0.01	-0.03	0.02	-0.01	-0.01	-0.01	0.01	0.00

Coordonnees des individus contributions et cosinus carres

VILLE	PRIN1	PRIN2	PRIN3	CONTG	CONT1	CONT2	CONT3	COSCA1	COSCA2	COSCA3
ajac	10.65	-1.11	-0.99	0.08	8.79	0.68	17.60	0.98	0.01	0.01
ange	-1.41	-1.61	0.26	0.00	0.15	1.45	1.21	0.42	0.55	0.01
ango	0.87	-0.94	0.61	0.00	0.06	0.49	6.66	0.33	0.38	0.16
besa	-6.14	1.66	-0.15	0.03	2.92	1.55	0.41	0.93	0.07	0.00
biar	6.89	-3.76	0.37	0.04	3.68	7.91	2.50	0.76	0.23	0.00
bord	5.44	0.50	1.16	0.02	2.29	0.14	24.07	0.94	0.01	0.04
bres	-2.42	-7.84	-0.26	0.05	0.45	34.36	1.16	0.09	0.91	0.00
cler	-2.87	0.81	0.08	0.01	0.64	0.37	0.11	0.91	0.07	0.00
dijo	-4.71	2.72	0.38	0.02	1.72	4.13	2.55	0.75	0.25	0.00
embr	-8.09	2.52	-0.65	0.05	5.07	3.55	7.56	0.89	0.09	0.01
gren	-2.99	2.82	0.14	0.01	0.69	4.44	0.37	0.50	0.45	0.00
lill	-6.74	-1.97	-0.59	0.03	3.52	2.16	6.28	0.91	0.08	0.01
limo	-3.93	-0.95	0.49	0.01	1.20	0.51	4.25	0.92	0.05	0.01
lyon	-1.72	3.14	0.16	0.01	0.23	5.51	0.44	0.23	0.76	0.00
mars	8.43	2.80	-0.36	0.05	5.51	4.37	2.33	0.90	0.10	0.00
mont	7.34	1.95	-0.15	0.04	4.17	2.13	0.38	0.93	0.07	0.00
nanc	-8.02	1.38	-0.23	0.04	4.98	1.07	0.96	0.97	0.03	0.00
nant	0.05	-2.05	0.43	0.00	0.00	2.35	3.28	0.00	0.94	0.04
nice	10.89	0.12	-0.25	0.08	9.18	0.01	1.07	1.00	0.00	0.00
nime	8.21	3.00	-0.03	0.05	5.21	5.03	0.01	0.88	0.12	0.00
orle	-4.18	-0.47	-0.11	0.01	1.36	0.12	0.21	0.99	0.01	0.00
pari	-2.00	-0.05	0.48	0.00	0.31	0.00	4.13	0.89	0.00	0.05
perp	12.12	1.35	0.17	0.10	11.37	1.01	0.52	0.99	0.01	0.00
reim	-5.88	0.05	-0.32	0.02	2.68	0.00	1.82	0.99	0.00	0.00
renn	-1.73	-3.29	0.28	0.01	0.23	6.03	1.44	0.21	0.77	0.01
roue	-4.59	-2.46	-0.12	0.02	1.63	3.37	0.28	0.77	0.22	0.00
stqu	-6.45	-1.15	-0.39	0.03	3.22	0.74	2.74	0.96	0.03	0.00
stra	-7.48	2.93	0.07	0.04	4.34	4.81	0.09	0.86	0.13	0.00
toul	12.62	-1.33	-0.43	0.11	12.33	0.99	3.26	0.99	0.01	0.00
tlse	3.25	0.84	-0.05	0.01	0.82	0.40	0.04	0.91	0.06	0.00
tour	-1.78	-0.33	0.25	0.00	0.24	0.06	1.09	0.93	0.03	0.02
vich	-3.63	0.70	-0.26	0.01	1.02	0.27	1.19	0.95	0.04	0.00

2 A.C.P. des données de criminalité

A.c.p. des donnees de crime

Statistiques elementaires

TYPE	_NAME_	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MEAN		7.44	25.73	124.09	211.30	1291.9	2671.29	377.53
STD		3.83	10.65	87.46	99.25	428.1	718.61	191.45
N		50.00	50.00	50.00	50.00	50.0	50.00	50.00

Matrice des covariances ou des correlations

NAME	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MURDER	1.00	0.60	0.48	0.65	0.39	0.10	0.07
RAPE	0.60	1.00	0.59	0.74	0.71	0.61	0.35
ROBBERY	0.48	0.59	1.00	0.56	0.64	0.45	0.59
ASSAULT	0.65	0.74	0.56	1.00	0.62	0.40	0.28
BURGLARY	0.39	0.71	0.64	0.62	1.00	0.79	0.56
LARCENY	0.10	0.61	0.45	0.40	0.79	1.00	0.44
AUTO	0.07	0.35	0.59	0.28	0.56	0.44	1.00

Valeurs propres, variances expliquees

K	LAMBDA	PCTVAR	CUMPCT
1	4.11	0.59	0.59
2	1.24	0.18	0.76
3	0.73	0.10	0.87
4	0.32	0.05	0.91
5	0.26	0.04	0.95
6	0.22	0.03	0.98
7	0.12	0.02	1.00

Vecteurs propres = coordonnees des variables du biplot

NAME	V1	V2	V3	V4	V5	V6	V7
MURDER	0.30	-0.63	0.18	-0.23	0.54	0.26	0.27
RAPE	0.43	-0.17	-0.24	0.06	0.19	-0.77	-0.30
ROBBERY	0.40	0.04	0.50	-0.56	-0.52	-0.11	0.00
ASSAULT	0.40	-0.34	-0.07	0.63	-0.51	0.17	0.19
BURGLARY	0.44	0.20	-0.21	-0.06	0.10	0.54	-0.65
LARCENY	0.36	0.40	-0.54	-0.23	0.03	0.04	0.60
AUTO	0.30	0.50	0.57	0.42	0.37	-0.06	0.15

Coordonnees des variables

NAME	V1	V2	V3	V4	V5	V6	V7
MURDER	0.61	-0.70	0.15	-0.13	0.27	0.12	0.09
RAPE	0.88	-0.19	-0.21	0.03	0.10	-0.36	-0.10
ROBBERY	0.81	0.05	0.42	-0.31	-0.26	-0.05	0.00
ASSAULT	0.80	-0.38	-0.06	0.35	-0.26	0.08	0.07
BURGLARY	0.89	0.23	-0.18	-0.03	0.05	0.25	-0.23
LARCENY	0.72	0.45	-0.46	-0.13	0.02	0.02	0.21
AUTO	0.60	0.56	0.48	0.24	0.19	-0.03	0.05

Correlations variables x facteurs

NAME	V1	V2	V3	V4	V5	V6	V7
MURDER	0.61	-0.70	0.15	-0.13	0.27	0.12	0.09
RAPE	0.88	-0.19	-0.21	0.03	0.10	-0.36	-0.10
ROBBERY	0.81	0.05	0.42	-0.31	-0.26	-0.05	0.00
ASSAULT	0.80	-0.38	-0.06	0.35	-0.26	0.08	0.07
BURGLARY	0.89	0.23	-0.18	-0.03	0.05	0.25	-0.23
LARCENY	0.72	0.45	-0.46	-0.13	0.02	0.02	0.21
AUTO	0.60	0.56	0.48	0.24	0.19	-0.03	0.05

Coordonnees des individus contributions et cosinus carres

STATEN	PRIN1	PRIN2	PRIN3	CONTG	CONT1	CONT2	CONT3	COSCA1	COSCA2	COSCA3
Alabama	-0.05	-2.12	0.51	0.02	0.00	7.24	0.71	0.00	0.85	0.05
Alaska	2.45	0.17	-0.07	0.03	2.91	0.05	0.01	0.49	0.00	0.00
Arizona	3.04	0.85	-1.77	0.04	4.51	1.18	8.63	0.64	0.05	0.22
Arkansas	-1.07	-1.36	-0.02	0.01	0.55	2.98	0.00	0.35	0.57	0.00
California	4.33	0.14	0.28	0.05	9.10	0.03	0.21	0.98	0.00	0.00
Colorado	2.53	0.93	-1.16	0.03	3.12	1.38	3.73	0.73	0.10	0.15
Connecticut	-0.55	1.52	0.79	0.01	0.15	3.71	1.73	0.09	0.69	0.19
Delaware	0.97	1.31	-0.53	0.01	0.46	2.77	0.78	0.28	0.51	0.08
Florida	3.14	-0.61	-1.23	0.04	4.80	0.60	4.15	0.77	0.03	0.12
Georgia	0.50	-1.39	0.25	0.01	0.12	3.14	0.17	0.10	0.81	0.03
Hawaii	0.83	1.84	-0.79	0.02	0.34	5.48	1.72	0.09	0.44	0.08
Idaho	-1.45	-0.01	-0.64	0.01	1.02	0.00	1.13	0.82	0.00	0.16
Illinois	0.52	0.10	1.13	0.01	0.13	0.01	3.53	0.11	0.00	0.52
Indiana	-0.50	0.00	0.25	0.00	0.12	0.00	0.17	0.40	0.00	0.10
Iowa	-2.61	0.83	-0.52	0.02	3.31	1.12	0.75	0.83	0.09	0.03
Kansas	-0.64	-0.03	-0.50	0.00	0.20	0.00	0.69	0.49	0.00	0.30
Kentucky	-1.74	-1.16	0.66	0.02	1.48	2.17	1.22	0.57	0.25	0.08
Louisiana	1.13	-2.10	0.37	0.02	0.62	7.15	0.38	0.20	0.68	0.02
Maine	-1.84	0.58	-0.54	0.01	1.65	0.55	0.80	0.70	0.07	0.06
Maryland	2.20	-0.20	0.39	0.02	2.36	0.06	0.41	0.67	0.01	0.02
Massachusetts	0.99	2.66	2.57	0.05	0.47	11.40	18.17	0.05	0.39	0.36
Michigan	2.30	0.16	0.54	0.02	2.56	0.04	0.81	0.85	0.00	0.05

Minnesota	-1.57	1.07	-0.15	0.01	1.20	1.84	0.06	0.64	0.30	0.01
Mississippi	-1.52	-2.57	0.71	0.03	1.13	10.69	1.39	0.23	0.65	0.05
Missouri	0.56	-0.56	0.57	0.00	0.15	0.51	0.89	0.29	0.29	0.29
Montana	-1.68	0.27	-0.37	0.01	1.38	0.12	0.38	0.77	0.02	0.04
Nebraska	-2.17	0.23	-0.11	0.01	2.29	0.08	0.03	0.93	0.01	0.00
Nevada	5.32	-0.26	-0.31	0.08	13.76	0.11	0.26	0.95	0.00	0.00
New Hampshire	-2.49	0.83	-0.21	0.02	3.02	1.12	0.12	0.87	0.10	0.01
New Jersey	0.22	0.97	0.61	0.00	0.02	1.53	1.03	0.03	0.62	0.24
New Mexico	1.23	-0.96	-1.08	0.01	0.73	1.49	3.24	0.36	0.22	0.28
New York	3.49	0.44	2.76	0.06	5.91	0.31	21.06	0.54	0.01	0.34
North Carolina	-0.71	-1.69	-0.09	0.01	0.24	4.60	0.02	0.10	0.59	0.00
North Dakota	-4.00	0.39	-0.09	0.05	7.79	0.25	0.02	0.97	0.01	0.00
Ohio	0.24	0.09	0.46	0.00	0.03	0.01	0.59	0.08	0.01	0.29
Oklahoma	-0.32	-0.63	-0.12	0.00	0.05	0.64	0.04	0.11	0.40	0.02
Oregon	1.46	0.59	-1.26	0.01	1.04	0.57	4.36	0.45	0.07	0.33
Pennsylvania	-1.74	-0.20	1.02	0.01	1.47	0.06	2.86	0.68	0.01	0.23
Rhode Island	-0.20	2.17	0.97	0.02	0.02	7.59	2.57	0.01	0.57	0.11
South Carolina	1.62	-2.18	-0.56	0.03	1.27	7.70	0.86	0.24	0.44	0.03
South Dakota	-3.20	-0.26	-0.14	0.03	4.99	0.11	0.05	0.91	0.01	0.00
Tennessee	-0.14	-1.15	0.66	0.01	0.01	2.12	1.20	0.01	0.56	0.18
Texas	1.41	-0.69	-0.08	0.01	0.97	0.76	0.02	0.53	0.13	0.00
Utah	-1.06	0.95	-0.65	0.01	0.55	1.45	1.15	0.45	0.36	0.17
Vermont	-2.09	0.95	-0.52	0.02	2.11	1.47	0.73	0.67	0.14	0.04
Virginia	-0.93	-0.70	-0.21	0.00	0.42	0.79	0.12	0.49	0.28	0.02
Washington	0.94	0.75	-1.32	0.01	0.43	0.90	4.78	0.22	0.14	0.44
West Virginia	-3.18	-0.82	0.54	0.03	4.91	1.09	0.81	0.90	0.06	0.03
Wisconsin	-2.53	0.79	-0.43	0.02	3.11	1.00	0.51	0.84	0.08	0.02
Wyoming	-1.44	0.06	-0.58	0.01	1.01	0.01	0.94	0.69	0.00	0.11

3 AFC des exploitations

The Correspondence Analysis Procedure

Contingency Table

	SINF1	S1__5	S5_10	S10_20
arie	620	830	760	1270
aver	420	1210	2130	4610
h.g.	830	2400	980	2900
gers	540	1530	680	3090
lot	460	1190	1630	2510
h.p.	590	1880	2190	2570
tarn	650	1230	1740	2920
t.g.	500	1410	1600	2980
Sum	4610	11680	11710	22850

Contingency Table

	S20_50	S50_99	S_100	Sum
arie	1530	600	120	5730
aver	6370	2030	550	17320
h.g.	4120	1680	450	13360
gers	6650	2170	390	15050
lot	3340	870	190	10190
h.p.	2200	180	10	9620
tarn	5250	1250	230	13270
t.g.	3880	810	170	11350
Sum	33340	9590	2110	95890

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	13 26 39 52 65
0.22033	0.04855	4655.05	65.96%	**********
0.13224	0.01749	1676.89	23.76%	*********
0.06056	0.00367	351.67	4.98%	**
0.05875	0.00345	330.99	4.69%	**
0.01870	0.00035	33.55	0.48%	
0.00977	0.00010	9.15	0.13%	
	0.07360	7057.3 (Degrees of Freedom = 42)		

Row Coordinates

	Dim1	Dim2
arie	0.144675	0.203728
aver	-.120423	-.149990
h.g.	-.047974	0.250562
gers	-.312741	0.042192
lot	0.101230	-.070955
h.p.	0.515438	-.004171
tarn	-.035629	-.081818
t.g.	0.095787	-.061953

Summary Statistics for the Row Points

	Quality	Mass	Inertia
arie	0.620524	0.059756	0.081695
aver	0.810221	0.180624	0.112071
h.g.	0.927156	0.139326	0.132887
gers	0.946888	0.156951	0.224286
lot	0.944660	0.106268	0.023358
h.p.	0.985382	0.100323	0.367547

tarn	0.469056	0.138388	0.031924
t.g.	0.797840	0.118365	0.026232

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
arie	0.025764	0.141825
aver	0.053957	0.232364
h.g.	0.006605	0.500187
gers	0.316214	0.015977
lot	0.022432	0.030594
h.p.	0.549038	0.000100
tarn	0.003619	0.052974
t.g.	0.022371	0.025978

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
arie	0	2	2
aver	0	2	2
h.g.	0	2	2
gers	1	0	1
lot	0	0	2
h.p.	1	0	1
tarn	0	0	2
t.g.	0	0	2

Squared Cosines for the Row Points

	Dim1	Dim2
arie	0.208022	0.412502
aver	0.317568	0.492653
h.g.	0.032787	0.894369
gers	0.929962	0.016926
lot	0.633447	0.311213
h.p.	0.985318	0.000065
tarn	0.074770	0.394286
t.g.	0.562528	0.235313

Column Coordinates

	Dim1	Dim2
SINF1	0.202644	0.286745
S1__5	0.218172	0.241648
S5_10	0.379388	-.165757
S10_20	0.056263	-.062236

S20_50	-.158791	-.056005
S50_99	-.332706	0.090935
S_100	-.344058	0.101379

Summary Statistics for the Column Points

	Quality	Mass	Inertia
SINF1	0.730675	0.048076	0.110219
S1__5	0.917894	0.121806	0.191112
S5_10	0.987035	0.122119	0.288154
S10_20	0.670044	0.238294	0.034013
S20_50	0.889324	0.347690	0.150604
S50_99	0.958834	0.100010	0.168596
S_100	0.671270	0.022004	0.057302

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
SINF1	0.040667	0.226041
S1__5	0.119431	0.406729
S5_10	0.362076	0.191866
S10_20	0.015539	0.052779
S20_50	0.180589	0.062362
S50_99	0.228042	0.047291
S_100	0.053656	0.012932

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
SINF1	0	2	2
S1__5	2	2	2
S5_10	1	1	1
S10_20	0	0	2
S20_50	1	0	1
S50_99	1	0	1
S_100	0	0	1

Squared Cosines for the Column Points

	Dim1	Dim2
SINF1	0.243374	0.487301
S1__5	0.412206	0.505688
S5_10	0.828823	0.158212
S10_20	0.301338	0.368707
S20_50	0.790935	0.098389
S50_99	0.892184	0.066650

```
S_100       0.617644      0.053626
```

4 A.F.D. des insectes

```
                    Canonical Discriminant Analysis

                74 Observations        73 DF Total
                 6 Variables           71 DF Within Classes
                 3 Classes              2 DF Between Classes


                         Class Level Information

                Y            Frequency        Weight     Proportion

                a                   21       21.0000       0.283784
                b                   22       22.0000       0.297297
                c                   31       31.0000       0.418919
                          Canonical Discriminant Analysis

                Between-Class Correlation Coefficients  /  Prob > |R|

Variable          X1          X2          X3          X4          X5          X6

X1             1.00000    -0.43892    -0.82713    -0.50655     0.97139    -0.79651
                0.0         0.7107      0.3799      0.6618      0.1526      0.4134

X2            -0.43892     1.00000     0.86803     0.99705    -0.21299     0.89288
                0.7107      0.0         0.3308      0.0489      0.8634      0.2974

X3            -0.82713     0.86803     1.00000     0.90355    -0.67001     0.99862
                0.3799      0.3308      0.0         0.2819      0.5326      0.0334

X4            -0.50655     0.99705     0.90355     1.00000    -0.28730     0.92479
                0.6618      0.0489      0.2819      0.0         0.8145      0.2485

X5             0.97139    -0.21299    -0.67001    -0.28730     1.00000    -0.63014
                0.1526      0.8634      0.5326      0.8145      0.0         0.5660

X6            -0.79651     0.89288     0.99862     0.92479    -0.63014     1.00000
                0.4134      0.2974      0.0334      0.2485      0.5660      0.0


                Total-Sample Correlation Coefficients  /  Prob > |R|

Variable          X1          X2          X3          X4          X5          X6

X1             1.00000     0.02635    -0.09573    -0.33496     0.78123    -0.57171
                0.0         0.8237      0.4171      0.0035      0.0001      0.0001
```

X2	0.02635	1.00000	0.67322	0.56151	-0.12218	0.48825
	0.8237	0.0	0.0001	0.0001	0.2997	0.0001
X3	-0.09573	0.67322	1.00000	0.59290	-0.31288	0.51611
	0.4171	0.0001	0.0	0.0001	0.0066	0.0001
X4	-0.33496	0.56151	0.59290	1.00000	-0.25094	0.78457
	0.0035	0.0001	0.0001	0.0	0.0310	0.0001
X5	0.78123	-0.12218	-0.31288	-0.25094	1.00000	-0.47861
	0.0001	0.2997	0.0066	0.0310	0.0	0.0001
X6	-0.57171	0.48825	0.51611	0.78457	-0.47861	1.00000
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0

Total-Sample

Variable	N	Sum	Mean	Variance	Std Dev
X1	74	13117	177.25676	865.09756	29.41254
X2	74	9173	123.95946	71.92984	8.48115
X3	74	3726	50.35135	7.57349	2.75200
X4	74	9976	134.81081	107.14180	10.35093
X5	74	961.00000	12.98649	4.58886	2.14216
X6	74	7058	95.37838	204.62199	14.30461

Y = a

Variable	N	Sum	Mean	Variance	Std Dev
X1	21	3845	183.09524	147.49048	12.14457
X2	21	2722	129.61905	51.24762	7.15874
X3	21	1076	51.23810	4.99048	2.23394
X4	21	3070	146.19048	31.66190	5.62689
X5	21	296.00000	14.09524	0.79048	0.88909
X6	21	2202	104.85714	38.22857	6.18293

Y = b

Variable	N	Sum	Mean	Variance	Std Dev
X1	22	3041	138.22727	87.32684	9.34488
X2	22	2752	125.09091	73.03896	8.54628
X3	22	1135	51.59091	8.06277	2.83950
X4	22	3042	138.27273	17.16017	4.14248
X5	22	222.00000	10.09091	0.94372	0.97145
X6	22	2345	106.59091	34.25325	5.85263

Y = c

Variable	N	Sum	Mean	Variance	Std Dev
X1	31	6231	201.00000	222.13333	14.90414
X2	31	3699	119.32258	44.15914	6.64523
X3	31	1515	48.87097	5.51613	2.34864
X4	31	3864	124.64516	21.36989	4.62276
X5	31	443.00000	14.29032	1.21290	1.10132
X6	31	2511	81.00000	79.73333	8.92935

Multivariate Statistics and F Approximations

S=2 M=1.5 N=32

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.01090038	94.3591	12	132	0.0001
Pillai's Trace	1.74204806	75.4128	12	134	0.0001
Hotelling-Lawley Trace	21.66449535	117.3493	12	130	0.0001
Roy's Greatest Root	17.77934399	198.5360	6	67	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

Canonical Discriminant Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.973011	0.970907	0.006232	0.946750
2	0.891795	0.887636	0.023959	0.795298

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	17.7793	13.8942	0.8207	0.8207
2	3.8852	.	0.1793	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F

1	0.01090038	94.3591	12	132	0.0001
2	0.20470195	52.0610	5	67	0.0001

Total Canonical Structure

	CAN1	CAN2
X1	0.898868	0.260064
X2	-0.343287	0.432595
X3	-0.448482	0.171774
X4	-0.649544	0.701667
X5	0.799566	0.475423
X6	-0.817965	0.366944

Between Canonical Structure

	CAN1	CAN2
X1	0.966593	0.256315
X2	-0.654566	0.756005
X3	-0.943551	0.331226
X4	-0.710623	0.703573
X5	0.878074	0.478525
X6	-0.924874	0.380273

Pooled Within Canonical Structure

	CAN1	CAN2
X1	0.487182	0.276360
X2	-0.092112	0.227585
X3	-0.116725	0.087655
X4	-0.327862	0.694407
X5	0.397939	0.463921
X6	-0.370550	0.325921

Canonical Discriminant Analysis

Total-Sample Standardized Canonical Coefficients

	CAN1	CAN2
X1	2.726839899	0.400878440
X2	-0.502351246	0.352478483
X3	-0.377793991	-0.799140019
X4	-0.673969346	2.091980100
X5	0.616056010	1.119569529
X6	-0.672930736	0.163908193

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2
X1	1.177217024	0.173065138
X2	-0.438064356	0.307371109
X3	-0.339647523	-0.718449563
X4	-0.312428278	0.969767755
X5	0.289634104	0.526357200
X6	-0.347576765	0.084660540

Raw Canonical Coefficients

	CAN1	CAN2
X1	0.0927101102	0.0136295073
X2	-.0592315278	0.0415602413
X3	-.1372799083	-.2903854248
X4	-.0651119461	0.2021054762
X5	0.2875861621	0.5226354401
X6	-.0470429135	0.0114584139

Class Means on Canonical Variables

Y	CAN1	CAN2
a	-0.783671161	3.045269156
b	-5.441221850	-1.530060092
c	4.392386293	-0.977075169

5 AFCM de l'enquête sur les cancers du sein

The Correspondence Analysis Procedure

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	4 8 12 16 20
0.54763	0.29990	1245.95	21.42%	**********
0.50983	0.25992	1079.87	18.57%	*********
0.45565	0.20762	862.56	14.83%	*******
0.44391	0.19705	818.66	14.08%	*******
0.42211	0.17818	740.26	12.73%	******
0.37332	0.13936	579.00	9.95%	****

```
0.34345   0.11796    490.06   8.43% ***********
-------   -------
1.40000   5816.36 (Degrees of Freedom = 121)
```

Column Coordinates

	Dim1	Dim2
Boston	-0.82461	0.27372
Glamorgan	-0.01366	0.50788
Tokio	0.72981	-0.62584
50-70	-0.12131	0.00814
<50	0.59686	-0.46534
>70	-0.93641	0.95818
non	-0.16813	0.94042
oui	0.06373	-0.35648
grande	1.34057	0.68005
petite	-0.33844	-0.17169
benigne	-0.51012	-0.54324
maligne	0.62282	0.66326

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
Boston	0.150168	0.019091
Glamorgan	0.000036	0.057412
Tokio	0.134827	0.114397
50-70	0.004508	0.000023
<50	0.087069	0.061065
>70	0.101799	0.122981
non	0.005182	0.187047
oui	0.001964	0.070902
grande	0.241577	0.071729
petite	0.060988	0.018109
benigne	0.095402	0.124832
maligne	0.116479	0.152411

The Correspondence Analysis Procedure
Inertia and Chi-Square Decomposition

```
  Singular  Principal Chi-
  Values    Inertias  Squares Percents    2    4    6    8   10
----+----+----+----+----+---
  0.65465   0.42857   1000.91  10.71% ***************************
  0.62473   0.39028    911.50   9.76% ************************
  0.60451   0.36543    853.44   9.14% ***********************
  0.58621   0.34364    802.56   8.59% *********************
  0.57735   0.33333    778.49   8.33% *********************
  0.57735   0.33333    778.49   8.33% *********************
  0.57735   0.33333    778.49   8.33% *********************
```

```
   0.57735   0.33333    778.49   8.33% ********************
   0.56360   0.31764    741.84   7.94% ********************
   0.54340   0.29529    689.63   7.38% ******************
   0.52287   0.27339    638.49   6.83% ****************
   0.50242   0.25243    589.53   6.31% ***************
   -------   -------
   4.00000   9341.84 (Degrees of Freedom = 196)


        Column Coordinates

             Dim1          Dim2

   non       0.97484       0.19771
   oui      -0.36953      -0.07494
   B5-7     -0.32167      -1.06740
   B<50      0.45642      -0.07923
   B>70      0.95688      -1.14027
   G5-7      0.34401      -0.14015
   G<50      1.22129       0.65705
   G>70      0.58883       0.43943
   T5-7     -0.37662       0.89522
   T<50     -1.00580       0.22199
   T>70     -0.44001       1.45888
   G_B      -1.64346       0.72400
   G_M      -0.33947       1.50399
   P_B      -0.27254      -0.64722
   P_M       0.89979       0.20030

   Supplementary Column Coordinates
               Dim1          Dim2

Boston          0.22562      -0.86189
Glamorgan       0.67127       0.22349
Tokio          -0.70838       0.58161
50-70          -0.13374      -0.10847
<50            -0.13819       0.26992
>70             0.64387      -0.28199
grande         -0.61043       1.34192
petite          0.15411      -0.33878
benigne        -0.37699      -0.54275
maligne         0.46028       0.66265
```

Table des matières