UNIVERSITE
PAUL
SABATIER
TOULOUSE III

PUBLICATIONS DU LABORATOIRE
DE
STATISTIQUE ET PROBABILITÉS

Statistique Descriptive Multidimensionnelle

ALAIN BACCINI & PHILIPPE BESSE

Version juillet 1999

Laboratoire de Statistique et Probabilités — UMR CNRS C5583
Université Paul Sabatier — 31062 – Toulouse cedex 4.

Sommaire

Chapitre 1

Introduction

Le corpus des méthodes, développées dans cette partie et la suivante, apparaissent dans différents environnements ou problématiques et souvent donc sous des appellations différentes. Il est instructif de resituer ces méthodes et leurs appellations dans un cadre chronologique afin de les rapprocher des tâches exploratoires habituelles d'un statisticien dans une entreprise.

1 Chronologie

Les bases théoriques de ces méthodes sont anciennes et sont principalement issues de psychomètres américains : 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 dans le courant des années 60. Sous l'appellation *"Multivariate Analysis"* elles poursuivent des objectifs sensiblement différents à ceux qui apparaîtront en France. Un individu ou unité statistique n'y est 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.

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, leur objectif *exploratoire* et une absence quasi systématique d'hypothèses de nature *probabiliste* au profit de la géométrie euclidienne. Elles insistent sur les représentations graphiques en particulier de celles des individus qui sont considérés au même titre que les variables.

Depuis la fin des années 1970, de nombreux travaux ont permis de rapprocher ou concilier les deux points de vue 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 ...

2 Méthodes

Les méthodes de *Statistique Multidimensionnelle* concernées sont généralement les 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.
  vi. Analyse en Facteurs ("Factor Analysis"), ou analyse en facteurs communs et spécifiques.
- Méthodes de classification :
  i. Classifications Hiérarchiques,
  ii. Classifications non Hiérarchiques,
  iii. Segmentation.

Les références les plus utiles pour ce cours sont : Bouroche & Saporta (1980), Jobson (1991), Droesbeke, Fichet & Tassi (1992), Everitt & Dunn (1991), Mardia, Kent & Bibby (1979), Saporta (1990), Lebart, Morineau & Piron (1995).

Le développement des moyens informatiques de stockage (bases de données) et de calcul permet maintenant le traitement et l'analyse d'ensembles de données très volumineux. Plus récemment, le perfectionnement des interfaces offrent aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples des outils logiciels. Cette évolution, ainsi que la popularisation de nouvelles méthodes algorithmiques (réseaux de neurones) et outils graphiques, conduit au développement et à la commercialisation de logiciels intégrant un sous-ensemble des méthodes cités sous la terminologie de *Data mining*.

Chapitre 2

Analyse en Composantes Principales

L'analyse en Composantes Principales (A.C.P.) est introduite ici comme l'estimation au sein d'un modèle, afin de préciser la signification statistique des résultats obtenus.

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 B.

1 Introduction

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

$$\begin{aligned} \forall i = 1, \ldots, n \ : w_i &> 0 \text{ et } \sum_{i=1}^{n} w_i = 1 \ ; \\ \forall i = 1, \ldots, n \ : y_i^j &= Y^j(i), \text{ mesure de } Y^j \text{ sur le } i^{ème} \text{ individu.} \end{aligned}$$

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

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

1.1 Représentation vectorielle de données quantitatives

- À 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*.

- À 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 :

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

  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 :	$\overline{y^j}$	$=$	$\left\langle y^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = y^{j'}\mathbf{D}\mathbf{1}_n$.
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	$=$	$(x^{j'}\mathbf{D}x^j)^{1/2} = \left\|x^j\right\|_{\mathbf{D}}$.
Covariance de Y^j et Y^k :	$x^{j'}\mathbf{D}x^k$	$=$	$\left\langle x^j, x^k \right\rangle_{\mathbf{D}}$.
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 :	$\frac{\left\langle x^j, x^k \right\rangle_{\mathbf{D}}}{\|x^j\|_{\mathbf{D}}\|x^k\|_{\mathbf{D}}}$	$=$	$\cos\theta_{\mathbf{D}}(x^j, x^k)$.

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 2.1. — *Soit $Y^1, \ldots, Y^p$, p variables quantitatives observées sur n individus de poids w_i. On appelle Analyse en Composantes Principales (A.C.P.) du triplet $(\mathbf{Y}, \mathbf{M}, \mathbf{D})$ les tableaux et graphiques obtenus à partir de 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 hypothèses 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 Modèle

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 $\mathbf{D} = \mathrm{diag}(w_1, \ldots, w_n)$.

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 A.C.P., 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 $\sigma^2\Gamma$, 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 :

$$\begin{array}{l} \{y_i \ ; i = 1, \ldots, n\},\ n \text{ vecteurs aléatoires indépendants de } E, \\ y_i = z_i + \varepsilon_i \ , i = 1, \ldots, n \text{ avec } \left\{ \begin{array}{l} E(\varepsilon_i) = 0,\ \mathrm{var}(\varepsilon_i) = \sigma^2\Gamma, \\ \sigma > 0 \text{ inconnu, } \Gamma \text{ régulière et connue,} \end{array} \right. \\ \exists A_q,\ \text{sous-espace affine de dimension } q \text{ de } E \text{ tel que } \forall i, z_i \in A_q \ (q < p). \end{array} \tag{2.1}$$

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*).

2.1 Estimation

PROPOSITION 2.2. — *L'estimation des paramètres de (2.1) 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* $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ *:*

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

Preuve

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

$$\min_{E_q, z_i} \left\{ \sum_{i=1}^{n} w_i \left\| y_i - z_i \right\|_{\mathbf{M}}^2 \ ; \ \dim(E_q) = q, z_i - \bar{z} \in E_q \right\}. \tag{2.2}$$

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 $(z_i - \bar{z})'$.

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

le problème (2.2) conduit alors à prendre $\widehat{\bar{z}} = \overline{y}$ et devient équivalent à résoudre :

$$\min_{\mathbf{Z}} \left\{ \left\| \mathbf{X} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p}, \ \operatorname{rang}(\mathbf{Z}) = q \right\}. \tag{2.3}$$

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 :

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

Remarque. —:

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

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

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 :

$$\widetilde{\mathbf{X}} = \mathbf{X}\Sigma^{-1/2} \text{ où } \Sigma = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2), \text{ avec } \sigma_j^2 = \text{Var } (Y^j) = \text{Var } (X^j) \text{ ;}$$

$\widetilde{\mathbf{S}}$ est alors la matrice $\mathbf{R} = \Sigma^{-1/2}\mathbf{S}\Sigma^{-1/2}$ des *corrélations*.

Remarque. — Sous l'hypothèse que la distribution de l'erreur est gaussienne, une estimation par maximum de vraisemblance conduit à la même solution.

2.2 Définition équivalente

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"

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

PROPOSITION 2.3. — *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.*

- 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* ;

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

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

$$\mathbf{C} = \mathbf{XF} = \mathbf{XMV} = \mathbf{U}\Lambda^{1/2}$$

est la matrice des composantes principales.

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

3 Représentations graphiques

3.1 Les individus

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

Projection

Chaque individu i représenté par x_i est approché par sa projection **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 :

$$\left\langle x_i, v^k \right\rangle_{\mathbf{M}} = x_i' \mathbf{M} v^k = e_i' \mathbf{X} \mathbf{M} v^k = c_i^k.$$

PROPOSITION 2.4. — *Les coordonnées de la projection* **M***-orthogonale de* x_i *sur* $\widehat{E_q}$ *sont les* q *premiers élément de la* $i^{ème}$ *ligne de la matrice* **C** *des composantes principales.*

Mesures de "qualité"

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

$$r_q = \frac{\mathrm{tr}\mathbf{S}\mathbf{M}\widehat{\mathbf{P}_q}}{\mathrm{tr}\mathbf{S}\mathbf{M}} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

Remarque. — La dispersion d'un nuage de points unidimensionnel par rapport à sa moyenne se mesure par la variance. Dans le cas multidimensionnel, la dispersion du nuage $\mathcal{N}$ par rapport à son barycentre g se mesure par l'*inertie*, généralisation de la variance :

$$I_g(\mathcal{N}) = \sum_{i=1}^{n} w_i \left\| y_i - g \right\|_{\mathbf{M}}^2 = \sum_{i=1}^{n} w_i \left\| x_i \right\|_{\mathbf{M}}^2 = \left\| \mathbf{X} \right\|_{\mathbf{M},\mathbf{D}}^2 = \mathrm{tr}\ (\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}) = \mathrm{tr}\ (\mathbf{S}\mathbf{M}).$$

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

$$[\cos\theta(x_i, \widehat{z_i}^q)]^2 = \frac{\left\| \widehat{\mathbf{P}_q} x_i \right\|_{\mathbf{M}}^2}{\left\| x_i \right\|_{\mathbf{M}}^2} = \frac{\sum_{k=1}^{q} (c_i^k)^2}{\sum_{k=1}^{p} (c_i^k)^2}.$$

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.

Contributions

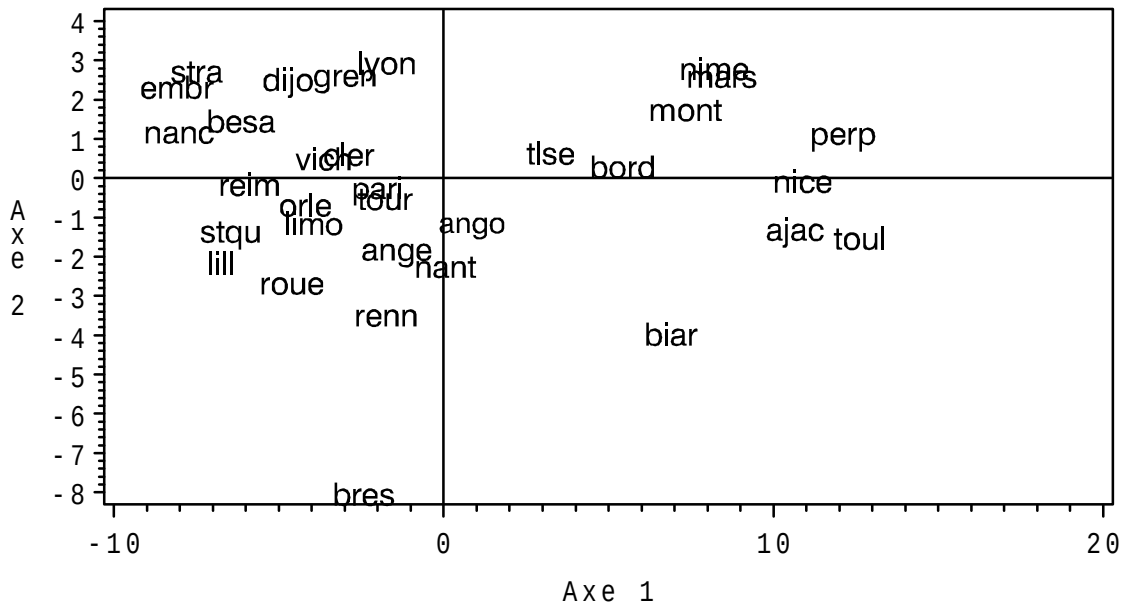
Les contributions de chaque individu à l'inertie de leur nuage

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

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.



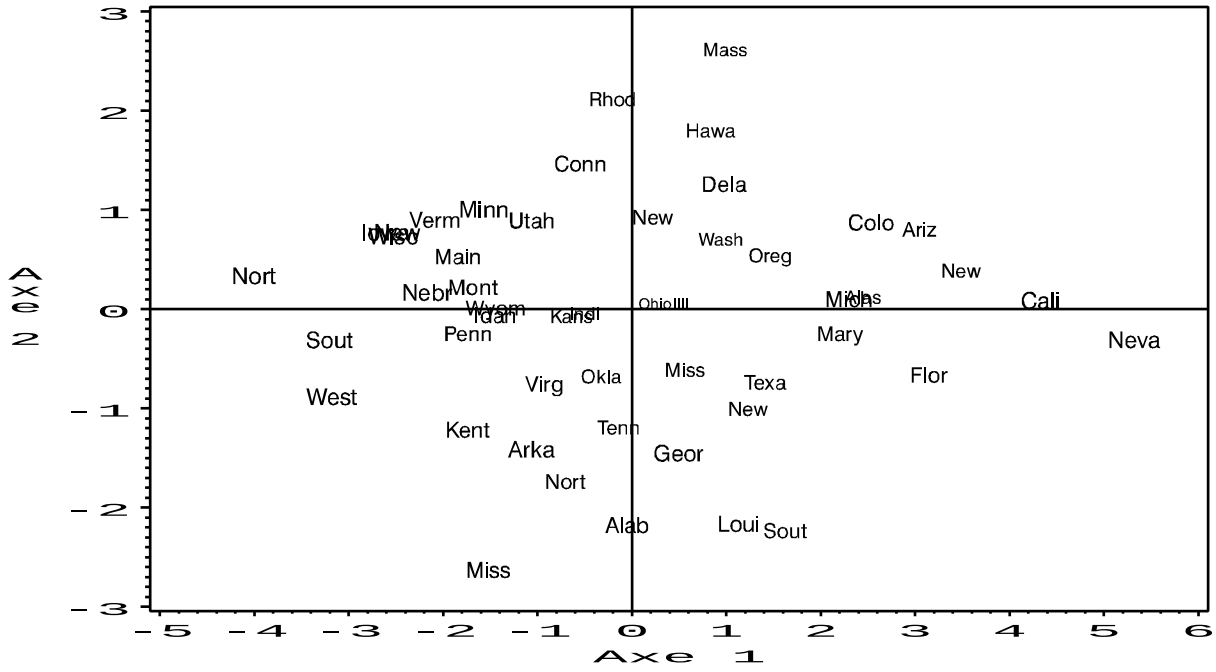


FIG. 2.2: Criminalité: premier plan des individus.

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 :

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

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

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

3.2 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 :

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

PROPOSITION 2.5. — *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 $\mathbf{V}\Lambda^{1/2}$.*

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 :

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

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 $\Sigma^{-1/2}\mathbf{V}\Lambda^{1/2}$.

Cercle des corrélations

Dans le cas de variables réduites $\tilde{x}^j = \sigma_j^{-1}x^j$, $\left\|\tilde{x}^j\right\|_{\mathbf{D}} = 1$, 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 :

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

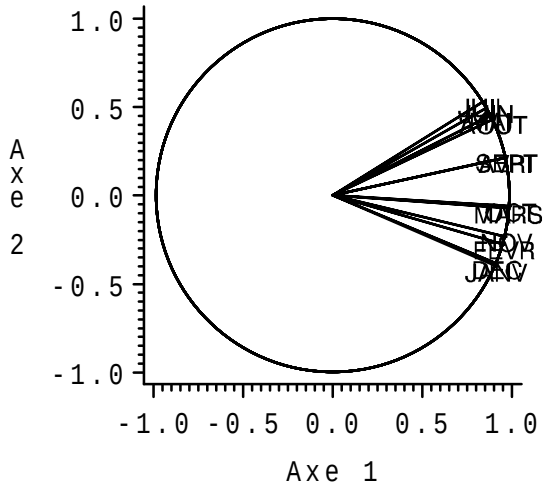


FIG. 2.3: Températures : premier plan des variables.

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.

3.3 Représentation simultanée ou "biplot"

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

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

s'exprime comme produit scalaire usuel des vecteurs

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

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 $\mathbf{V}\Lambda^{1/2}$; elle permet d'interpréter les angles entre vecteurs variables (corrélations) et les produits scalaires comme précédemment.

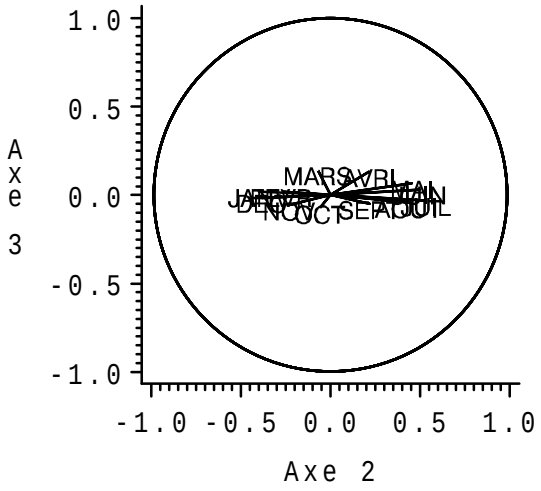


FIG. 2.4: Températures : deuxième plan des variables.

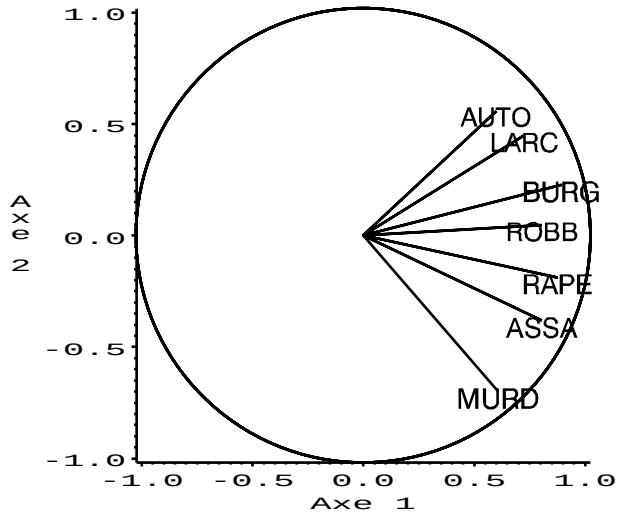


FIG. 2.5: Criminalité : premier plan des variables.

Remarque. — .

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.

4 Choix de dimension

La qualité des estimations auxquelles conduit l'A.C.P. dépend, de façon évidente, 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 pour q ont été proposés dans la littérature. Nous présentons ici les plus courants.

4.1 Part d'inertie

La "qualité globale" des représentations est mesurée par la *part d'inertie 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 d'inertie expliquée r_q soit supérieure à une valeur seuil fixée a priori par l'utilisateur. C'est souvent le seul critère employé.

4.2 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.

4.3 Éboulis des valeurs propres

C'est le graphique (figures 4.3 et 4.3) 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 $\widehat{E_q}$.

4.4 Boîtes-à-moustaches des variables principales

Un graphique (figure 4.4 et 4.4) 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.

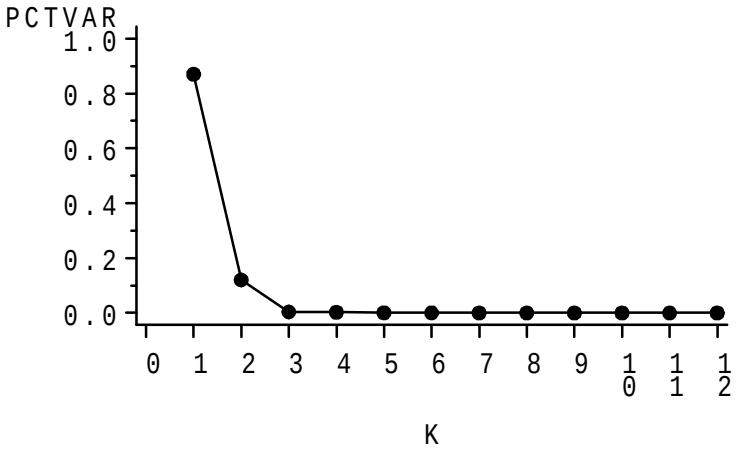


FIG. 2.6: Températures : éboulis des valeurs propres.

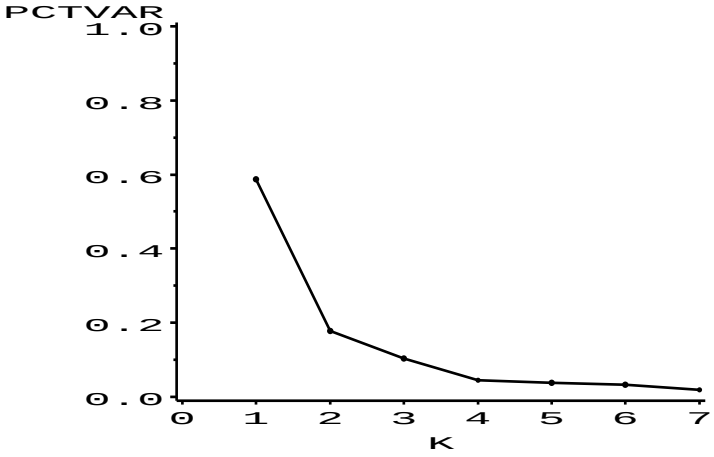


FIG. 2.7: Criminalité : éboulis des valeurs propres.

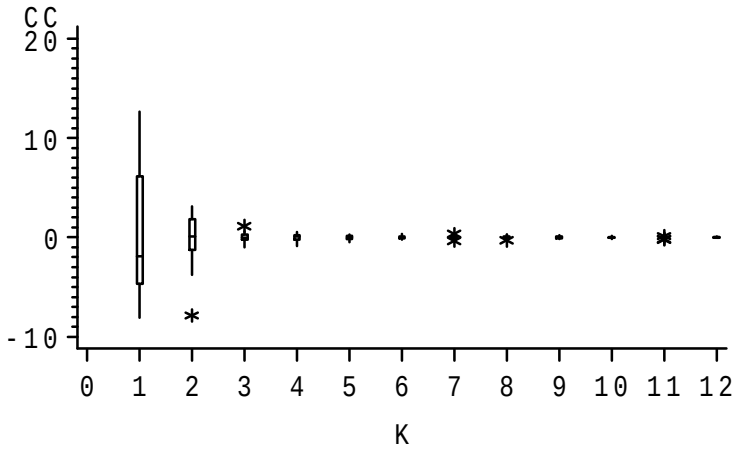


FIG. 2.8: Températures : composantes en boîtes.

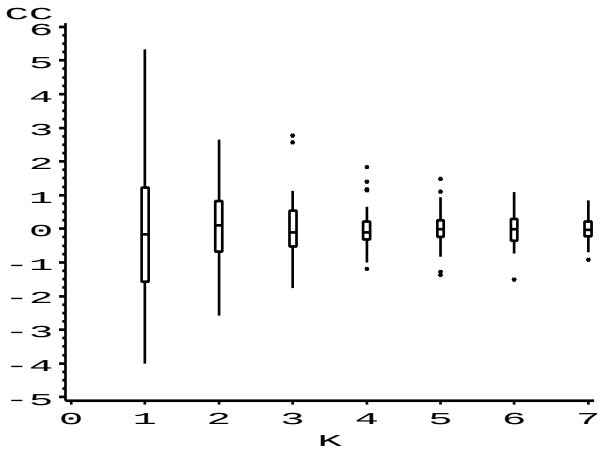


FIG. 2.9: Criminalité : composantes en boîtes.

5 Pratique de l'A.C.P.

5.1 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.

5.2 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 : $\mathbf{M} = \text{diag}(a_1^2, \ldots, a_p^2)$.

5.3 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 2.6. — *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 $(\mathbf{Y}\mathbf{M}^{1/2}, \mathbf{I}_p, \mathbf{D})$

5.4 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 ! *Remarque.* — 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 3

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 associée à ces observations, de dimension $r \times c$, est notée $\mathbf{T}$; son élément générique est 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* vérifiant $\sum_{l=1}^{r} n_{l+} = \sum_{h=1}^{c} n_{+h} = n$. 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 :

$$\begin{aligned} g_r &= [f_{1+}, \ldots, f_{r+}]', \\ \text{et } g_c &= [f_{+1}, \ldots, f_{+c}]'. \end{aligned}$$

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 également 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 $\mathbb{R}^c$ et les 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 3.1. — *On dit que deux variables X et Y sont* non liées *relativement à T si et seulement si :*

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

Il est équivalent de dire que tous les profils–lignes sont égaux, ou encore que tous les profils–colonnes sont égaux (voir volume 1, chapitre 3, paragraphe 3).

Cette notion est cohérente avec celle d'indépendance en probabilités. En effet, soit $\Omega = \{1,\ldots,n\}$ l'ensemble des individus observés et $(\Omega, \mathcal{P}(\Omega), P)$ l'espace probabilisé associé 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 variables X et Y. On note $\widetilde{X}$ et $\widetilde{Y}$ les variables aléatoires associées aux 2 variables statistiques X et Y :

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

$P_{\widetilde{X}}$, $P_{\widetilde{Y}}$ 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* (la vérification est immédiate).

On suppose maintenant 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 barres parallèles, ainsi que le calcul de coefficients de liaison (Cramer ou Tschuprow) donnent une première idée de la variation conjointe des deux variables (voir volume 1, chapitre 3, paragraphe 3). Le test du χ^2 permet de plus de s'assurer du caractère significatif de cette liaison. Il est construit de la manière suivante :

l'hypothèse nulle est $H0$: $\widetilde{X}$ et $\widetilde{Y}$ sont indépendantes en probabilités ;

l'hypothèse alternative est $H1$: les variables $\widetilde{X}$ et $\widetilde{Y}$ ne sont pas indépendantes.

La statistique de test est alors

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

elle suit asymptotiquement (pour les grandes valeurs de n), et si l'hypothèse $H0$ est vraie, une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté. On rejette donc $H0$ (et l'on conclut au caractère significatif de la liaison) si χ^2 dépasse une valeur particulière (valeur ayant une probabilité faible et fixée a priori – en général 0,05 – d' être dépassée par une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté).

1.4 Objectifs

Pour préciser la liaison existant entre les variables X et Y, on souhaite définir un modèle statistique susceptible de fournir des paramètres dont la représentation graphique (de type biplot) illustrera les "*correspondances*" entre les modalités de ces 2 variables. Cette approche sera développée au paragraphe 3.

Une autre approche, très courante dans la littérature francophone, consiste à définir l'Analyse Factorielle des Correspondances (A.F.C.) comme étant le résultat d'une double Analyse en Composantes Principales

- l'A.C.P. des profils–lignes,
- l'A.C.P. des profils–colonnes,

relativement à la métrique dite du χ^2. Cette approche est présentée au paragraphe 2.
Remarque. — :

i. Toute structure d'ordre existant éventuellement sur les modalités de X ou de Y est ignorée par l'A.F.C.

ii. Tout individu présente 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 Métriques du χ^2

Les correspondances entre modalités évoquées au paragraphe précédant se trouvent exprimées en termes de distances au sens d'une certaine métrique. Ainsi, 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 (les coordonnées de a^l sont les éléments de la l–ième colonne de $\mathbf{A}$). 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 ici à l'A.C.P. du triplet $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$. Dans cette A.C.P., les "individus" sont les modalités de Y, caractérisées par les profils–colonnes de $\mathbf{T}$, pondérées par les fréquences marginales correspondantes et rangées en lignes dans la matrice $\mathbf{B}'$.

PROPOSITION 3.2. — *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 et semi–définie positive* $\mathbf{BA}$.

Preuve Elle se construit en remarquant successivement que :

i. le barycentre du nuage des profils–colonnes est le vecteur g_r des fréquence marginales de X,

ii. la matrice $\mathbf{B}\mathbf{D}_c\mathbf{B}' - g_r\mathbf{D}_c g_r'$ joue le rôle de la matrice des variances–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 $(\mathbf{SM})$

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

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 $\mathbf{U}$ la matrice contenant les vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés de $\mathbf{BA}$. La représentation des "individus" de l'A.C.P. réalisée fournit une représentation des modalités de la variable Y. Elle se fait au moyen des lignes de la matrice des "composantes principales" $(\mathbf{XMV})$:

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

2.3 A.C.P. des profils–lignes

De façon symétrique (ou duale), on s'intéresse à l'A.C.P. des "individus" modalités de X ou profils–lignes (la matrice des données est $\mathbf{A}'$), pondérés par les fréquences marginales des lignes de $\mathbf{T}$ (la matrice diagonale des poids est $\mathbf{D}_r$) et utilisant la métrique du χ^2. Il s'agit donc de l'A.C.P. de $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$.

PROPOSITION 3.3. — *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 et semi–définie positive* $\mathbf{AB}$.

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

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

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

$$\begin{aligned} \mathbf{V} &= \mathbf{A}\mathbf{U}\Lambda^{-1/2}, \\ \mathbf{U} &= \mathbf{B}\mathbf{V}\Lambda^{-1/2}; \end{aligned}$$

Λ est la matrice diagonale des valeurs propres (exceptée $\lambda_0 = 0$) communes aux deux A.C.P.

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

On en déduit les formules dites de *transition*:

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

La représentation simultanée habituellement construite à partir de ces matrices (option par défaut de SAS) n'est pas a priori justifiée. On lui donnera 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 Le 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 Le 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 paramètres μ_{lk} et ν_{hk}. Ces paramètres peuvent être estimés par maximum de vraisemblance ou par moindres carrés.

3.3 Le modèle de corrélation

On écrit ici :

$$p_{lh} = p_{l+}p_{+h} + \sum_{k=1}^{q}\sqrt{\lambda_k}u_l^k v_h^k, \tag{3.1}$$

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

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

Remarque. — :

i. Le modèle (3.1) ci-dessus est équivalent au modèle considéré par Goodman (1991) :

$$p_{lh} = p_{l+} p_{+h} \left(1 + \sum_{k=1}^{q} \sqrt{\lambda_k} \xi_l^k \eta_h^k \right), \tag{3.2}$$

moyennant une homothétie sur les 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. Dans le contexte de la statistique descriptive, qui est celui de ce cours, il est naturel de retenir cette dernière solution.

3.4 Estimation Moindres Carrés dans le modèle de corrélation

Critère

Considérons les espaces $\mathbb{R}^c$ et $\mathbb{R}^r$ munis de leur base canonique et de leur métrique du χ^2 respectives et notons $\mathbf{P}$ le tableau des probabilités théoriques définies selon le modèle (3.1). Le critère des moindres carrés s'écrit alors :

$$\min_{\mathbf{P}} \left\| \frac{1}{n}\mathbf{T} - \mathbf{P} \right\|^2_{\mathbf{D_r^{-1}} \mathbf{D_c^{-1}}}. \tag{3.3}$$

Estimation

Proposition 3.4. — *L'estimation des paramètres de (3.1) en résolvant (3.3) 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* $\mathbf{BA}$ *(resp.* $\mathbf{AB}$*) 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 $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$:

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

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.

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

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

associés aux valeurs propres λ_k.

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 $\mathbf{BA}$ (resp. $\mathbf{AB}$) 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 :

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

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.

□

4 Représentations graphiques

4.1 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

$$\frac{u_l^k}{f_{l+}}\lambda_k^{1/4} \text{ et } \frac{v_h^k}{f_{+h}}\lambda_k^{1/4},$$

termes génériques respectifs des matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\Lambda^{1/4} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\Lambda^{1/4},$$

qui sont encore les estimations des vecteurs ξ_l et η_h du modèle 3.2. Leur représentation (par exemple avec $q = 2$) illustre alors la *correspondance* entre les deux modalités x_l et y_h : lorsque deux modalités, éloignées de l'origine, sont voisines (resp. opposées), leur produit scalaire est de valeur absolue importante ; leur cellule conjointe contribue alors fortement et de manière positive (resp. négative) à la dépendance entre les deux variables.

L'A.F.C. apparaît ainsi comme la meilleure reconstitution des fréquences f_{lh}, ou encore la meilleure représentation des écarts relatifs à l'indépendance. La représentation simultanée des modalités de X et de Y se trouve ainsi pleinement justifiée.

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

4.2 Double A.C.P.

Chacune des deux A.C.P. réalisée permet une représentation des "individus" (modalités) approchant, au mieux, les distances du χ^2 entre les profils–lignes d'une part, les profils–colonnes d'autre part. Les coordonnées sont fournies cette fois par les matrices (de composantes principales)

$$\mathbf{C}_r = \mathbf{D}_r^{-1}\mathbf{U}\Lambda^{1/2} \text{ et } \mathbf{C}_c = \mathbf{D}_c^{-1}\mathbf{V}\Lambda^{1/2}.$$

Même si la représentation simultanée n'a plus alors de justification, elle reste couramment employée. En fait, les graphiques obtenus diffèrent très peu de ceux du biplot ; ce dernier sert donc de "caution" puisque les interprétations des graphiques sont identiques. On notera que cette représentation issue de la double A.C.P. est celle réalisée par la plupart des logiciels statistiques (c'est en particulier le cas de SAS).

4.3 Représentations barycentriques

D'autres représentations simultanées, appelées barycentriques, sont proposées en utilisant les matrices

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

ou encore les matrices

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

Si l'on considère alors, par exemple, la formule de transition

$$\mathbf{C}_r = \mathbf{A}'\mathbf{C}_c\Lambda^{-1/2} \Longleftrightarrow \mathbf{C}_r\Lambda^{1/2} = \mathbf{A}'\mathbf{C}_c \Longleftrightarrow \mathbf{D}_r^{-1}\mathbf{U}\Lambda = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V}\Lambda^{1/2},$$

on voit que dans la seconde des représentations ci–dessus, chaque modalité x_l de X est représentée par un vecteur qui est barycentre de l'ensemble des vecteurs associés aux modalités de Y, chacun d'eux ayant pour poids l'élément correspondant du l–ième profil–ligne. Là encore, la représentation simultanée s'en trouve parfaitement justifiée. Malheureusement, dans la pratique, les représentations barycentriques sont souvent illisibles ; elles sont, de ce fait, très peu utilisées.

4.4 Autre représentation

La pratique de l'A.F.C. montre que l'interprétation des graphiques est toujours la même, quelle que soit la représentation simultanée choisie parmi les 3 ci–dessus.

On peut ainsi envisager d'utiliser, pour une représentation simultanée des modalités de X et de Y, les coordonnées fournies respectivement par les lignes des matrices

$$\mathbf{D}_r^{-1}\mathbf{U} \quad \text{et} \quad \mathbf{D}_c^{-1}\mathbf{V}.$$

L'interprétation du graphique sera toujours la même et les matrices ci–dessus, outre leur simplicité, présentent l'avantage de conduire a une représentation graphique qui reste invariante lorsque l'on utilise la technique d'Analyse Factorielle des Correspondances Multiples (voir chapitre suivant) sur les données considérées ici.

4.5 Aides à l'interprétation

Les qualités de représentation dans la dimension choisie et les contributions des modalités de X ou de Y se déduisent aisément de celles de l'A.C.P. Ces quantités sont utilisées à la fois pour choisir la dimension de l'A.F.C. et pour interpréter ses résultats dans la dimension choisie.

Mesure de la qualité globale

Pour une dimension donnée q ($1 \leq q \leq d = \inf(r-1, c-1)$), la qualité globale des représentations graphiques en dimension q se mesure par le rapport entre la somme des q premières valeurs propres de l'A.F.C. et leur somme complète de 1 à d.

Compte–tenue de la propriété $\sum_{k=1}^{d} \lambda_k = \Phi^2$ (voir en 6.1), la qualité de la représentation dans la k–ième dimension s'écrit

$$\frac{n\lambda_k}{\chi^2}.$$

On parle encore de part du khi–deux expliquée par la k–ième dimension (voir les sorties du logiciel SAS).

Mesure de la qualité de chaque modalité

Pour chaque modalité de X (resp. de Y), la qualité de sa représentation en dimension q se mesure par le cosinus carré de l'angle entre le vecteur représentant cette modalité dans $\mathbb{R}^c$ (resp. dans $\mathbb{R}^r$) et sa projection $\mathbf{D}_c^{-1}$–orthogonale (resp. $\mathbf{D}_r^{-1}$–orthogonale) dans le sous–espace principal de dimension q.

Ces cosinus carrés s'obtiennent en faisant le rapport des sommes appropriées des carrés des coordonnées extraites des lignes de $\mathbf{C}_r$ (resp. de $\mathbf{C}_c$).

Contributions à l'inertie totale

L'inertie totale (en dimension d) du nuage des profils–lignes (resp. des profils–colonnes) est égale à la somme des d valeurs propres. La part due au i–ième profil–ligne (resp. au j–ième profil–colonne) valant $f_{l+} \sum_{k=1}^{d} (c_{rl}^k)^2$ (resp. $f_{+h} \sum_{k=1}^{d} (c_{ch}^k)^2$), les contributions à l'inertie totale s'en déduisent immédiatement.

Contributions à l'inertie selon chaque axe

Il s'agit de quantités analogues à celles ci–dessus, dans lesquelles il n'y a pas de sommation sur l'indice k. Ces quantités sont utilisées dans la pratique pour sélectionner les modalités les plus importantes, c'est–à–dire celles qui contribuent le plus à la définition de la liaison entre les 2 variables X et Y.

Remarque

En général, on n'interprète pas les axes d'une A.F.C. (en particulier parce qu'il n'y a pas de variable quantitative intervenant dans l'analyse). L'interprétation s'appuie surtout sur la position relative des différentes modalités repérées comme les plus importantes.

5 Exemple

La table de contingence étudiée s'intéresse à la répartition des exploitations agricoles de la région Midi–Pyrénées dans les différents départements en fonction de leur taille. Elle croise la variable qualitative *département*, à 8 modalités, avec la variable *taille de l'exploitation*, quantitative découpée en 6 classes. Les données, ainsi que les résultats numériques obtenus avec la procédure `corresp` de SAS/STAT, sont fournis en annexe.

La figure 5 présente le premier plan factoriel utilisant les coordonnées obtenues par défaut, c'est–à–dire celles de la double ACP.

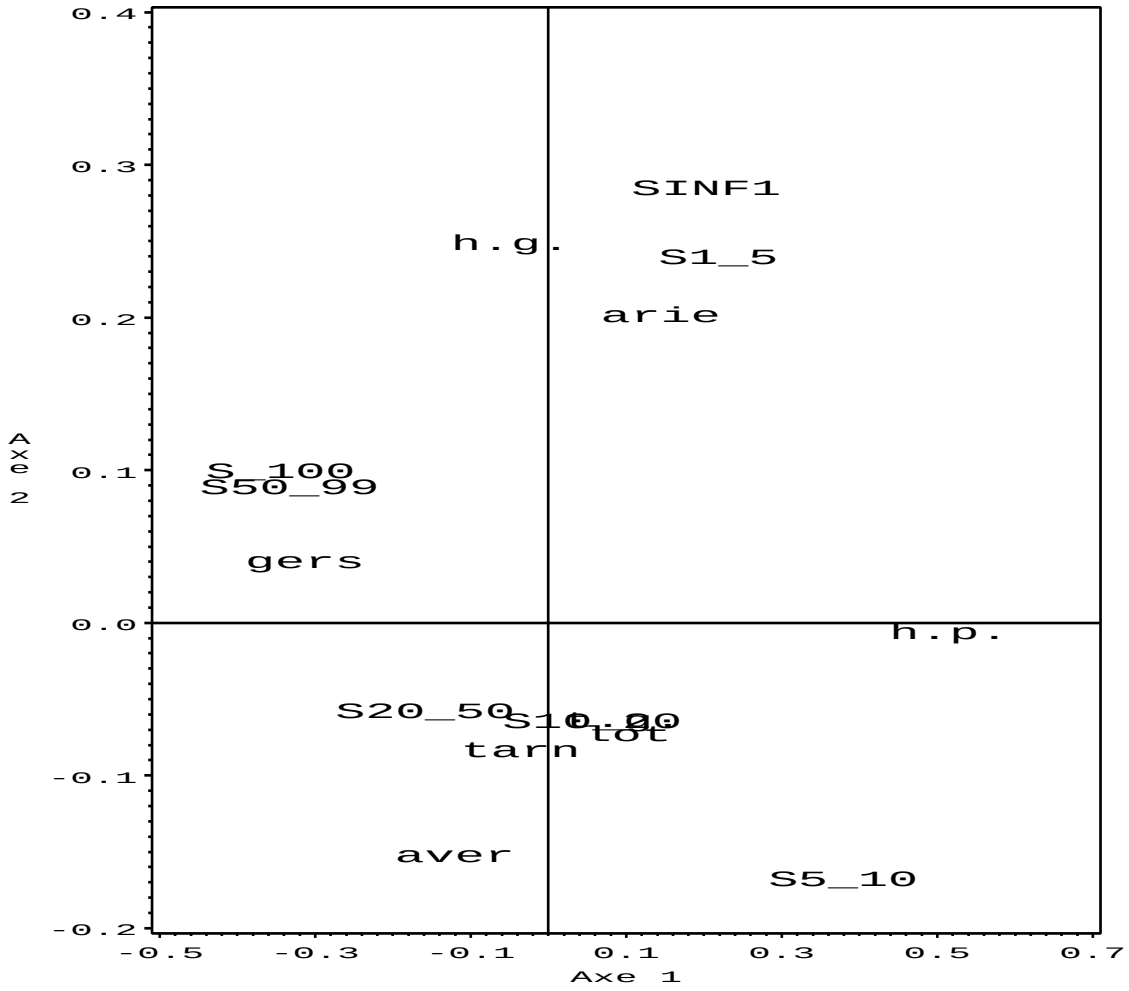


FIG. 3.1: Répartition des exploitations agricoles par taille et par département

6 Compléments

6.1 Propriétés

- *Formule de reconstitution des données.* On appelle ainsi l'approximation d'ordre q (c'est–à–dire fournie par l'A.F.C. en dimension q) de la table des fréquences initiales $(\frac{1}{n}\mathbf{T})$:

$$f_{lh} \simeq f_{l+} f_{+h} \sum_{k=1}^{q} \sqrt{\lambda_k} u_l^k v_h^k.$$

- Les valeurs propres vérifient :

$$\sum_{k=1}^{d} \lambda_k = \Phi^2.$$

En effet, on vérifie facilement :

$$\mathrm{tr}\mathbf{AB} = \sum_{k=0}^{d} \lambda_k = 1 + \frac{\chi^2}{n} = 1 + \Phi^2;$$

d'où le résultat.

6.2 Invariance

- Les tables de contingence $\mathbf{T}$ et $\alpha\mathbf{T}, \alpha \in \mathbb{R}_+^*$, admettent la même A.F.C. (évident).
- *Propriété d'équivalence 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 (leurs profils sont identiques) et le regroupement de x_l et x_i en une seule modalité (en additionnant les effectifs) laisse inchangées les représentations graphiques (même chose pour les colonnes de $\mathbf{T}$). Cette propriété est une conséquence de la métrique du χ^2.

6.3 Choix de la dimension q

Le choix de la dimension pose les mêmes problèmes qu'en A.C.P. De nombreuses techniques empiriques ont été proposées (essentiellement : part d'inertie expliquée, éboulis des valeurs propres). Il existe également une approche probabiliste qui peut donner des indications intéressantes. Nous la détaillons ci–dessous.

Posons

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

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

$$K_q = \sum_{l=1}^{r} \sum_{h=1}^{c} \frac{(n_{lh} - \widehat{n_{lh}^q})^2}{\widehat{n_{lh}^q}} \simeq n \sum_{k=q+1}^{d} \lambda_k$$

suit approximativement une loi de 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 proche de l'indépendance en probabilités ; les fréquences conjointes sont alors bien approchées par les produits des fréquences marginales.

Chapitre 4

Analyse Factorielle des Correspondances Multiples

Cette méthode est une généralisation de l'Analyse Factorielle des Correspondances, permettant de décrire les relations entre p ($p > 2$) variables qualitatives simultanément observées sur n individus.

1 Codages de variables qualitatives

1.1 Tableau disjonctif complet

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 $\mathcal{X}_k$ est la k–ième modalité de X. On notera n_k l'effectif de $\mathcal{X}_k$.

On appelle *matrice des indicatrices* des modalités de X, et l'on notera $\mathbf{X}$, la matrice $n \times c$ de terme général :

$$x_i^k = X_{(k)}(i).$$

On vérifie :

$$\sum_{k=1}^{c} x_i^k = 1, \forall i \quad \text{et} \quad \sum_{i=1}^{n} x_i^k = n_k.$$

Considérons maintenant 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 alors *tableau disjonctif complet* la matrice $\mathbf{X}, n \times c$, obtenue par concaténation des matrices $\mathbf{X}_j$:

$$\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p].$$

$\mathbf{X}$ vérifie :

$$\sum_{k=1}^{c} x_i^k = p, \forall i \quad \text{et} \quad \sum_{i=1}^{n} \sum_{k=1}^{c} x_i^k = np.$$

D'autre part, la somme des éléments d'une colonne de $\mathbf{X}$ est égale à l'effectif marginal de la modalité de la variable X^j correspondant à cette colonne.

1.2 Tableau de Burt

On observe toujours p variables qualitatives sur un ensemble de n individus. On appelle *tableau de Burt* la matrice $\mathcal{B}$, $c \times c$, définie par :

$$\mathcal{B} = \mathbf{X}'\mathbf{X}.$$

On peut écrire $\mathcal{B} = [\mathcal{B}_{jl}]$ $(j = 1, \ldots, p\,;\, l = 1, \ldots, p)$; chaque bloc $\mathcal{B}_{jl}$, de dimension $c_j \times c_l$, est défini par :

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

Si $j \neq l$, $\mathcal{B}_{jl}$ est la table de contingence obtenue par croisement des variables X^j en lignes et X^l en colonnes. Si $j = l$, le bloc diagonal $\mathcal{B}_{jj}$ est lui–même une matrice diagonale vérifiant :

$$\mathcal{B}_{jj} = \text{ diag } (n_1^j, \ldots, n_{c_j}^j).$$

La matrice $\mathcal{B}$ est symétrique, d'effectifs marginaux $n_l^j p$ et d'effectif total np^2.

1.3 La démarche suivie dans ce chapitre

La généralisation de l'A.F.C. à plusieurs variables qualitatives repose sur certaines propriétés observées dans le cas élémentaire où $p = 2$. On s'intéresse tout d'abord aux résultats fournis par l'A.F.C. usuelle réalisée sur le tableau disjonctif complet $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ relatif à 2 variables qualitatives X^1 et X^2 ; $\mathbf{X}$ est alors considéré comme une table de contingence (paragraphe 2). Ensuite, on suit la même démarche avec l'A.F.C. réalisée sur le tableau de Burt $\mathcal{B}$ relatif à X^1 et X^2 (paragraphe 3). Enfin, en utilisant les propriétés obtenues dans les deux premiers cas, on généralise cette double approche à un nombre quelconque p de variables qualitatives ; on définit ainsi l'Analyse Factorielle des Correspondances Multiples (paragraphe 4).

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

2.1 Données

On note toujours X^1 et X^2 les 2 variables qualitatives considérées et r et c leurs nombres respectifs de modalités.

Les matrices intervenant dans l'A.F.C. usuelle sont reprises ici avec les mêmes notations, mais surlignées. On obtient ainsi :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]\ ; \\
\overline{\mathbf{D}_r} &= \frac{1}{n}\mathbf{I}_n\ ; \\
\overline{\mathbf{D}_c} &= \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\Delta\ ; \\
\overline{\mathbf{A}} &= \frac{1}{2n}\overline{\mathbf{T}}'\overline{\mathbf{D}}_r^{-1} = \frac{1}{2}\mathbf{X}'\ ; \\
\overline{\mathbf{B}} &= \frac{1}{2n}\overline{\mathbf{T}\mathbf{D}}_c^{-1} = \frac{1}{n}\mathbf{X}\Delta^{-1}.
\end{aligned}$$

On considère ici l'A.F.C. comme une double A.C.P. : celle des profils–lignes $\overline{\mathbf{A}}$, puis celle des profils–colonnes $\overline{\mathbf{B}}$.

2.2 A.C.P. des profils–lignes

Les profils–lignes, provenant de $\overline{\mathbf{T}}$, sont associés aux n individus observés. Leur A.C.P. conduit ainsi à une représentation graphique des individus, inconnue en A.F.C. classique.

PROPOSITION 4.1. —
L'A.C.P. des profils–lignes issue de l'A.F.C. réalisée sur le tableau disjonctif complet associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}_c}^{-1}$*–symétrique et 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 de* $\overline{\mathbf{AB}}$ *s'écrivent*

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

où les λ_k *sont les valeurs propres de la matrice* $\mathbf{AB}$ *(donc celles de l'A.F.C. classique de* X^1 *et* X^2*).*

Les vecteurs propres $\overline{\mathbf{D}_c}^{-1}$*–orthonormés associés se mettent sous la forme*

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

la matrice $\mathbf{U}$ *(resp.* $\mathbf{V}$*) contient les vecteurs propres* $\mathbf{D}_r^{-1}$*–orthonormés (resp.* $\mathbf{D}_c^{-1}$*-orthonormés) de la matrice* $\mathbf{BA}$ *(resp.* $\mathbf{AB}$*); autrement dit, les matrices* U *et* V *sont les matrices de vecteurs propres obtenues en faisant l'A.F.C. classique de la table de contingence croisant* X^1 *et* X^2*.*

La matrice des composantes principales s'écrit

$$\overline{\mathbf{C}_r} = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r + \mathbf{X}_2\mathbf{C}_c\right]\Lambda^{-1/2},$$

où $\mathbf{C}_r$ *et* $\mathbf{C}_c$ *sont encore les matrices de composantes principales de l'A.F.C. classique.*

Dans la pratique, on ne considère que les $d = \inf(r-1, c-1)$ plus grandes valeurs propres différentes de 1, ainsi que les vecteurs propres associés. Les valeurs propres sont rangées dans la matrice

$$\mathbf{M} = \text{ diag } (\mu_1, \ldots, \mu_d) = \frac{1}{2}\left[\mathbf{I}_d + \Lambda^{1/2}\right].$$

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

On notera que la matrice $\overline{\mathbf{C}_r}$, $n \times d$, fournit les coordonnées permettant la représentation graphique des individus sur les axes factoriels.

2.3 A.C.P. des profils–colonnes

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

PROPOSITION 4.2. — *L'A.C.P. des profils–colonnes issue de l'A.F.C. réalisée sur le tableau disjonctif complet associé à 2 variables conduit à l'analyse spectrale de la matrice* $\overline{\mathbf{D}_r}^{-1}$*–symétrique et positive :*

$$\overline{\mathbf{BA}} = \frac{1}{2n}\left[\mathbf{X}_1\mathbf{D}_r^{-1}\mathbf{X}_1' + \mathbf{X}_2\mathbf{D}_c^{-1}\mathbf{X}_2'\right].$$

Les $r+c$ valeurs propres non nulles de $\overline{\mathbf{BA}}$ sont les μ_k.
Les vecteurs propres $\overline{\mathbf{D}_r}^{-1}$–orthonormés associés se mettent sous la forme :

$$\overline{\mathbf{U}} = \frac{1}{n}\overline{\mathbf{C}_r}\mathbf{M}^{-1/2}.$$

La matrice des composantes principales s'écrit :

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

Ainsi, l'A.F.C. du tableau disjonctif complet permet, grâce aux coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}_c}$, une représentation simultanée des modalités des 2 variables. Cette représentation est très voisine de celle obtenue par l'A.F.C. classique, définie au chapitre précédent. Une simple homothétie sur chaque axe factoriel, de rapport $\sqrt{\frac{1+\sqrt{\lambda_k}}{2\lambda_k}}$, permet de passer de l'une à l'autre.

De plus, cette approche permet aussi de réaliser 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 2 modalités qu'il a présentées. Dans le cas où n est grand, le graphique des individus a néanmoins peu d'intérêt ; seule sa forme générale peut en avoir un.
Remarque. — Si, dans l'A.F.C. classique, on choisit d'utiliser, pour la représentation simultanée des modalités de X^1 et de X^2, les lignes des matrices

$$\mathbf{C}_r^* = \mathbf{D}_r^{-1}\mathbf{U} = \mathbf{C}_r\Lambda^{-\mathbf{1/2}} \text{ et } \mathbf{C}_c^* = \mathbf{D}_c^{-1}\mathbf{V} = \mathbf{C}_c\Lambda^{-\mathbf{1/2}}$$

(voir chapitre précédent, sous–section 4.4), alors on obtient par A.F.C. du tableau disjonctif complet la matrice

$$\overline{\mathbf{C}_c^*} = \overline{\mathbf{C}_c}\mathbf{M}^{-1/2} = \begin{bmatrix} \mathbf{C}_r^* \\ \mathbf{C}_c^* \end{bmatrix};$$

il y a invariance de la représentation des modalités lorsqu'on passe d'une méthode à l'autre. Pour les individus, on obtient

$$\overline{\mathbf{C}_r^*} = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r^* + \mathbf{X}_2\mathbf{C}_c^*\right]\mathbf{M}^{-1/2}$$

(le commentaire est alors le même qu'avec $\overline{\mathbf{C}_r}$).

3 A.F.C. du tableau de Burt relatif à 2 variables

Dans cette section, on s'intéresse aux résultats fournis par l'A.F.C. réalisée sur le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, $(r+c)\times(r+c)$, relatif aux 2 variables X^1 et X^2 ; $\mathcal{B}$ est encore considéré comme une table de contingence. La matrice $\mathcal{B}$ étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; il suffit donc de considérer une seule A.C.P.

Les notations des matrices usuelles de l'A.F.C. sont maintenant réutilisées surmontées d'un tilde. On obtient ainsi :

$$\begin{aligned}
\widetilde{\mathbf{T}} &= \mathcal{B} = \begin{bmatrix} n\mathbf{D}_r & \mathbf{T} \\ \mathbf{T}' & n\mathbf{D}_c \end{bmatrix}; \\
\widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\Delta = \overline{\mathbf{D}_c}\,; \\
\widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix} = \overline{\mathbf{AB}}.
\end{aligned}$$

On considère encore l'A.F.C. comme l'A.C.P. des profils–lignes $\widetilde{\mathbf{A}}$ (ou des profils–colonnes $\widetilde{\mathbf{B}}$).

PROPOSITION 4.3. — *L'A.C.P. des profils–lignes (ou des profils–colonnes) issue de l'A.F.C. réalisée sur le tableau de Burt associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice* $\widetilde{\mathbf{D}_c}^{-1}$*–symétrique et positive :*

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

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$*–orthonormés*

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

Les valeurs propres associées vérifient : $\nu_k = \mu_k^2$.
La matrice des composantes principales s'écrit:

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

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant une représentation simultanée des modalités des deux variables. À une homothétie près, cette représentation est identique à celle de l'A.F.C. classique, réalisée sur la table de contingence $\mathbf{T}$ (mais le rapport d'homothétie, sur chaque axe, n'est plus le même qu'avec $\overline{\mathbf{C}_c}$).
Remarque. —

- En reprenant les notations de la remarque 2.3, on obtient ici :

$$\widetilde{\mathbf{C}_r^*}\ (=\widetilde{\mathbf{C}_c^*}) = \widetilde{\mathbf{C}_r}\mathbf{M}^{-1} = \overline{\mathbf{C}_c^*} = \begin{bmatrix}\mathbf{C}_r^*\\ \mathbf{C}_c^*\end{bmatrix}.$$

  Ainsi, si l'on utilise ce mode de représentation graphique, les trois approches de l'A.F.C. que nous avons présentées conduisent à la même représentation simultanée des modalités des 2 variables : il y a donc invariance de cette représentation.

- Dans les deux cas d'A.F.C. considérés dans ce chapitre (sur tableau disjonctif complet et sur tableau de Burt) on trouve, par construction, des valeurs propres non nulles sans signification statistique. En conséquence, les critères de qualité s'exprimant comme une "part d'inertie expliquée" n'ont plus de signification.

- L'A.F.C. sur tableau de Burt ne prend en compte que l'information contenue dans $\mathcal{B}$ qui ne considère que les croisements de variables prises deux à deux. En conséquence, les interactions de niveau plus élevé sont ignorées par cette approche, à moins de procéder à des recodages de variables comme l'explique l'exemple présenté dans la section 5.

4 Analyse Factorielle des Correspondances Multiples

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 $c = \sum_{j=1}^{p} c_j$. 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 des observations ($\mathbf{X}$ est $n \times c$) et $\mathcal{B} = \mathbf{X}'\mathbf{X}$ le tableau de Burt correspondant ($\mathcal{B}$ est carré d'ordre c, symétrique).

DÉFINITION 4.4. — *On appelle Analyse Factorielle des Correspondances Multiples (A.F.C.M.) des variables $(X^1, \ldots, X^p)$ relativement à l'échantillon considéré, l'A.F.C. réalisée soit sur la matrice $\mathbf{X}$ soit sur la matrice $\mathcal{B}$.*

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, $\mathbf{D}_j = \dfrac{1}{n}$ diag $(n_1^j, \ldots, n_{c_j}^j)$ et $\Delta =$ diag $(\mathbf{D}_1 \ldots \mathbf{D}_p)$ (Δ est carrée d'ordre c et diagonale).

4.2 A.F.C. du tableau disjonctif complet X

Comme dans le cas $p = 2$, on reprend les notations de l'A.F.C. classique en les surlignant. On obtient ainsi :

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

A.C.P. des profils–lignes

PROPOSITION 4.5. — *L'A.C.P. des profils–lignes issue de l'A.F.C. réalisée sur le tableau disjonctif complet de p variables qualitatives conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}_c}^{-1}$–symétrique et positive :*

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

Il y a m $(m \leq c - p)$ valeurs propres notées μ_k, $(0 < \mu_k < 1)$ rangées dans la matrice diagonale $\mathbf{M}$.

La matrice des vecteurs propres $\overline{\mathbf{D}_c}^{-1}$–orthonormés associés se décompose en blocs de la façon suivante :

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

chaque bloc $\mathbf{V}_j$ est de dimension $c_j \times m$.

La matrice des composantes principales s'écrit :

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

Comme dans le cas $p = 2$, la matrice des composantes principales permet de réaliser une représentation graphique des individus dans laquelle chacun apparaît, à un facteur près, comme le barycentre des p modalités qu'il a présentées.

Remarque. — La généralisation au cas $p > 2$ restreint les propriétés. Ainsi, les vecteurs des blocs $\mathbf{V}_j$ ne sont pas les vecteurs propres $\mathbf{D}_j^{-1}$–orthonormés d'une matrice connue.

A.C.P. des profils–colonnes

PROPOSITION 4.6. — *L'A.C.P. des profils-colonnes issue de l'A.F.C. réalisée sur le tableau disjonctif complet de p variables conduit à l'analyse spectrale de la matrice* $\overline{\mathbf{D}_r}^{-1}$*–symétrique et positive :*

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

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

$$\overline{\mathbf{U}} = \overline{\mathbf{BV}}\mathbf{M}^{-1/2}.$$

La matrice des composantes principales s'écrit:

$$\overline{\mathbf{C}_c} = p\Delta^{-1}\overline{\mathbf{V}}\mathbf{M}^{1/2} \ ;$$

elle se décompose en blocs sous la forme :

$$\overline{\mathbf{C}_c} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}.$$

Chaque bloc $\mathbf{C}_j$, de dimension $c_j \times m$, fournit en lignes les coordonnées des modalités de la variable X^j permettant la représentation graphique simultanée.

4.3 A.F.C. du tableau de Burt $\mathcal{B}$

Le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, carré d'ordre c, étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; on ne considère donc ici qu'une seule A.C.P.

En utilisant encore le tilde dans ce cas, les matrices usuelles de l'A.F.C. deviennent :

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

PROPOSITION 4.7. — *L'A.C.P. des profils–lignes (ou des profils–colonnes) issue de l'A.F.C. réalisée sur le tableau de Burt associé à p variables qualitatives conduit à l'analyse spectrale de la matrice* $\widetilde{\mathbf{D}_c}^{-1}$*–symétrique et positive :*

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

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$*–orthonormés* $\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}}$.
Les valeurs propres associées vérifient $\nu_k = \mu_k^2$.
La matrice des composantes principales s'écrit:

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \overline{\mathbf{C}_c}\mathbf{M}^{1/2} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}\mathbf{M}^{1/2}.$$

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant la représentation simultanée des modalités de toutes les variables (on ne peut pas faire de représentation des individus si l'on fait l'A.F.C. du tableau de Burt).

4.4 Variables illustratives

Soit X^0 une variable qualitative, à c_0 modalités, observée sur les mêmes n individus que les X^j et n'étant pas intervenue dans l'A.F.C.M. Soit $\mathbf{T}_{0j}$ la table de contingence $c_0 \times c_j$ croisant les variables X^0 en lignes et X^j en colonnes. L'objectif est maintenant de représenter les modalités de cette variable supplémentaire X^0 dans le graphique de l'A.F.C.M. réalisée sur $X^1, \ldots, X^p$. Pour cela, on considère les matrices :

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

Les coordonnées des modalités de la variable supplémentaires X^0 sur les axes factoriels sont alors fournies dans les lignes de la matrice

$$\mathbf{C}_0 = \mathbf{A}_0 \widetilde{\mathbf{D}_c}^{-1} \widetilde{\mathbf{V}} = p\mathbf{A}_0\Delta^{-1}\overline{\mathbf{V}}.$$

4.5 Interprétation

Les représentations graphiques sont interprétées de manière analogue à ce qui est fait dans l'A.F.C. de deux variables, bien que la représentation simultanée des modalités de toutes les variables ne soit pas, en toute rigueur, réellement justifiée.

Les "principes" suivants sont donc appliqués :

- on interprète 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 rapports de valeurs propres ne sont pas interprétables comme indicateurs de qualité globale ; on peut néanmoins regarder la décroissance des premières valeurs propres pour choisir la dimension ;
- les coefficients de qualité de chaque modalité ne peuvent pas être interprétés ; seules les contributions des modalités à l'inertie selon les axes sont interprétées, selon le même principe qu'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). C'est le cas de l'exemple ci–dessous qui décrit les résultats partiels d'une enquête réalisée dans trois centres hospitaliers (Boston, Glamorgan, Tokyo) 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 connues pour chacune des patientes :

- le centre de diagnostic,

			Histologie			
			Inflammation minime		Grande inflammation	
Centre	*Âge*	*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. 4.1: Données sous la forme d'une table de contingence complète

- 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

On se reportera à la figure 5.1. La variable survie, qui joue en quelques sortes le rôle de variable à expliquer, est très proche de l'axe 2 et semble liée à chacune des autres variables.

5.3 Analyse des interactions

Pour essayer de mettre en évidence d'éventuelles interactions entre variables, 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 également croisées pour définir la variable `histol`, à 4 modalités.

Une nouvelle analyse est alors réalisée en considérant comme actives les deux variables nouvellement créées, ainsi que la variable `survie`, et comme illustratives les variables initiales : `centre, âge, inflam, appar`. Les résultats sont donnés dans la figure 5.3.

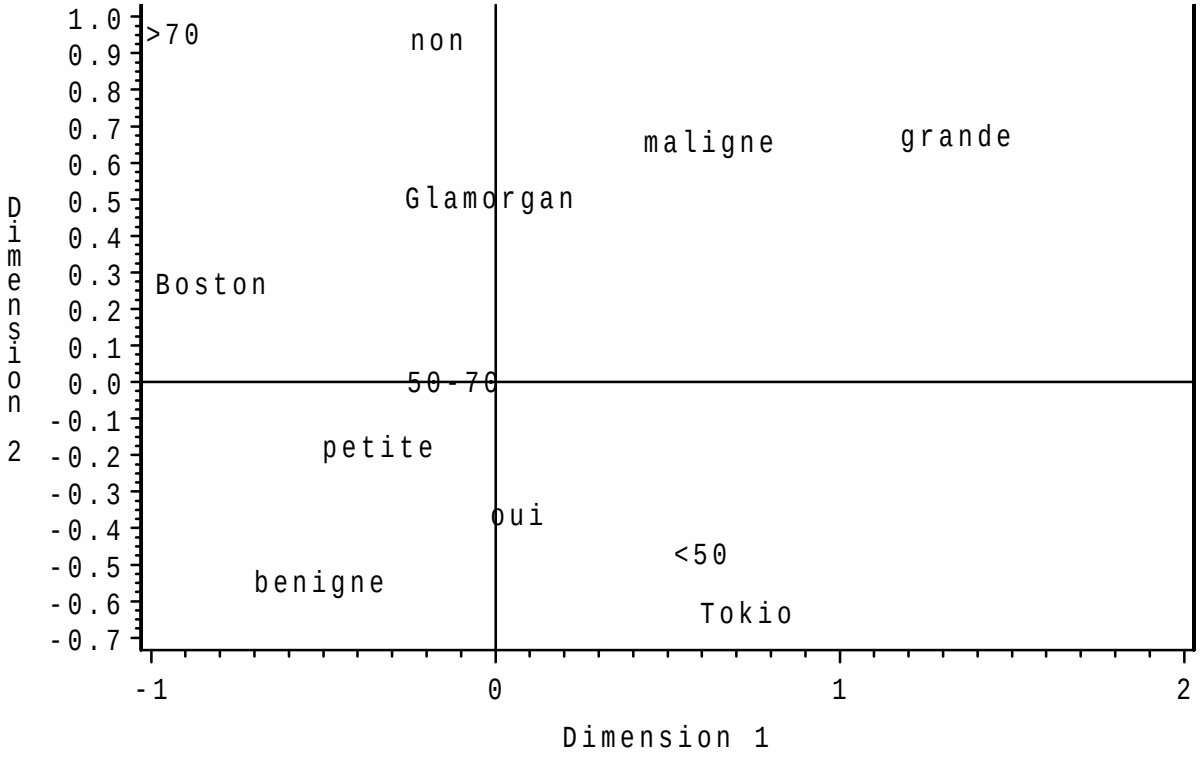


FIG. 4.1: Cancer du sein : analyse des données brutes.

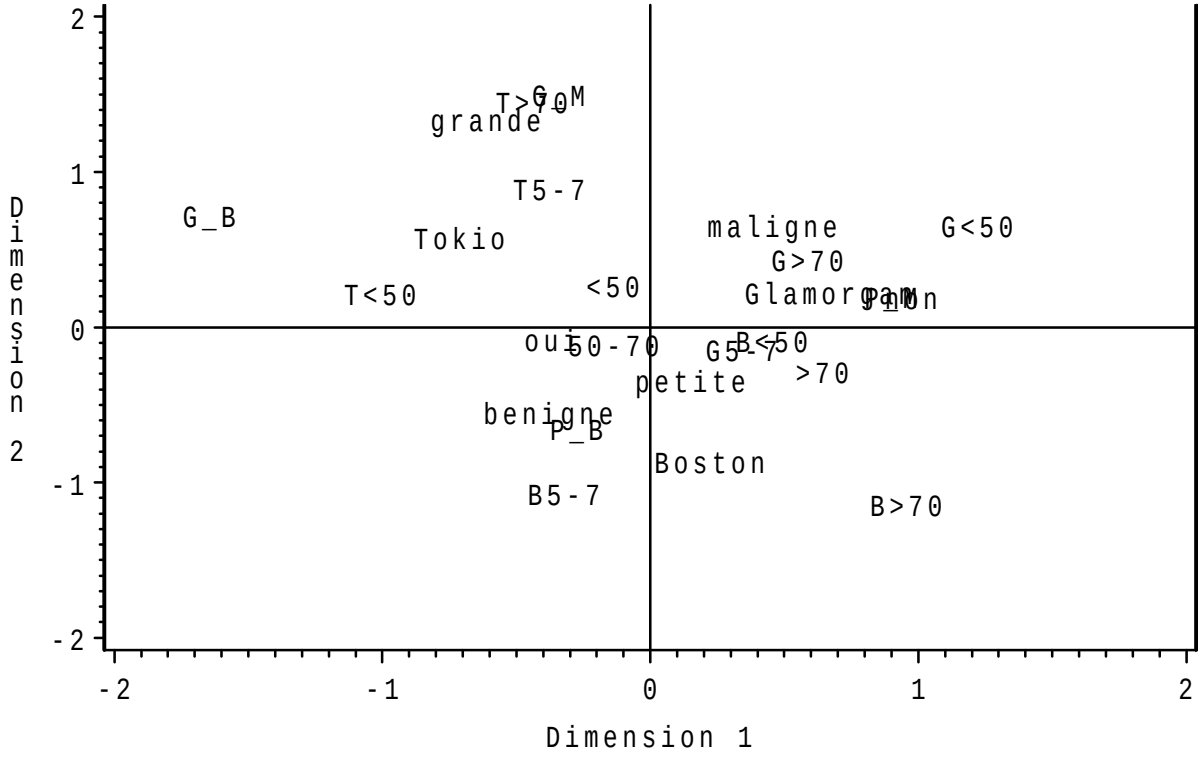


FIG. 4.2: Cancer du sein : analyse des interactions.

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 (sondages) 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 sont compatibles avec les potentialités des outils statistiques utilisables.

Exemple de chronologie, certains points pouvant être itérés :

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

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

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

iv. Faire l'étude univariée (tris à plat), des graphiques élémentaires, les vérifications nécessaires.

v. Regrouper ou éliminer les modalités trop rares, éliminer les variables non discriminantes ou redondantes, recoder en classes (de préférence d'effectifs voisins) les variables quantitatives.

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

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

viii. Faire l'A.F.C.M. de ces quelques variables importantes.

ix. Réaliser une modélisation des données (logit, log–linéaire ...), dans le cas où une variable doit être expliquée.

Chapitre 5

Analyse Factorielle Discriminante

1 Introduction à l'A.F.D.

1.1 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

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

1.2 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). C'est l'objet du chapitre suivant.

Remarque. — 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*.

1.3 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 5.1. — *La matrice des covariances se décompose en*

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

2 Définition

2.1 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 :

$$\begin{array}{l} \{y_i \ ; i = 1, \ldots, n\}, \ n \text{ vecteurs indépendants de } E, \\ \forall k, \forall i \in \Omega_k, y_i = z_k + \varepsilon_i \text{ avec } \left\{ \begin{array}{l} E(\varepsilon_i) = 0, \ \mathrm{var}(\varepsilon_i) = \Gamma, \\ \Gamma \text{ régulière et inconnue,} \end{array} \right. \\ \exists A_q, \text{ sous-espace affine de de dimension } q \text{ de } E \text{ tel que} \\ \forall k, z_k \in A_q, \ (q < \min(p, m-1)). \end{array} \tag{5.1}$$

Remarque. — 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.

2.2 Estimation

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

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

Comme on a

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

on est conduit à résoudre :

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

La covariance $\sigma^2\Gamma$ du modèle (5.1) 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 Γ. 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 :

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

.

Proposition 5.2. — *L'estimation des paramètres E_q et z_k du modèle 5.1 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 $(\mathbf{Y}|\mathbf{T}, \mathbf{D})$.*

3 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..

3.1 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

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

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

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

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*.

3.2 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.

3.3 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 $\mathbf{V}\Lambda^{1/2}$.

3.4 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 : $\Sigma_e^{-1}\mathbf{V}\Lambda^{1/2}$. La matrice Σ_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 Variantes de l'A.F.D.

4.1 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 :	Λ^*	=	$\frac{n-m}{m-1}\Lambda$,
– 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 :	$\mathbf{V}^*\Lambda^{*1/2}$	=	$\sqrt{\frac{n}{m-1}}\mathbf{V}\Lambda^{1/2}$,
– corrélations variables-facteurs :	$\Sigma_e^{*-1}\mathbf{V}^*\Lambda^{*1/2}$	=	$\Sigma_e^{-1}\mathbf{V}\Lambda^{1/2}$.

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.

4.2 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 5.3. — *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 : $\Lambda(\mathbf{I}+\Lambda)^{-1}$,
- vecteurs propres : $\mathbf{V}(\mathbf{I}+\Lambda)^{1/2}$,
- représentation des barycentres : $\overline{\mathbf{C}}(\mathbf{I}+\Lambda)^{-1/2}$,
- représentation des variables : $\mathbf{V}\Lambda^{1/2}$,
- corrélations variables-facteurs : $\Sigma_e^{-1}\mathbf{V}\Lambda^{1/2}$.

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.

5 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 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.

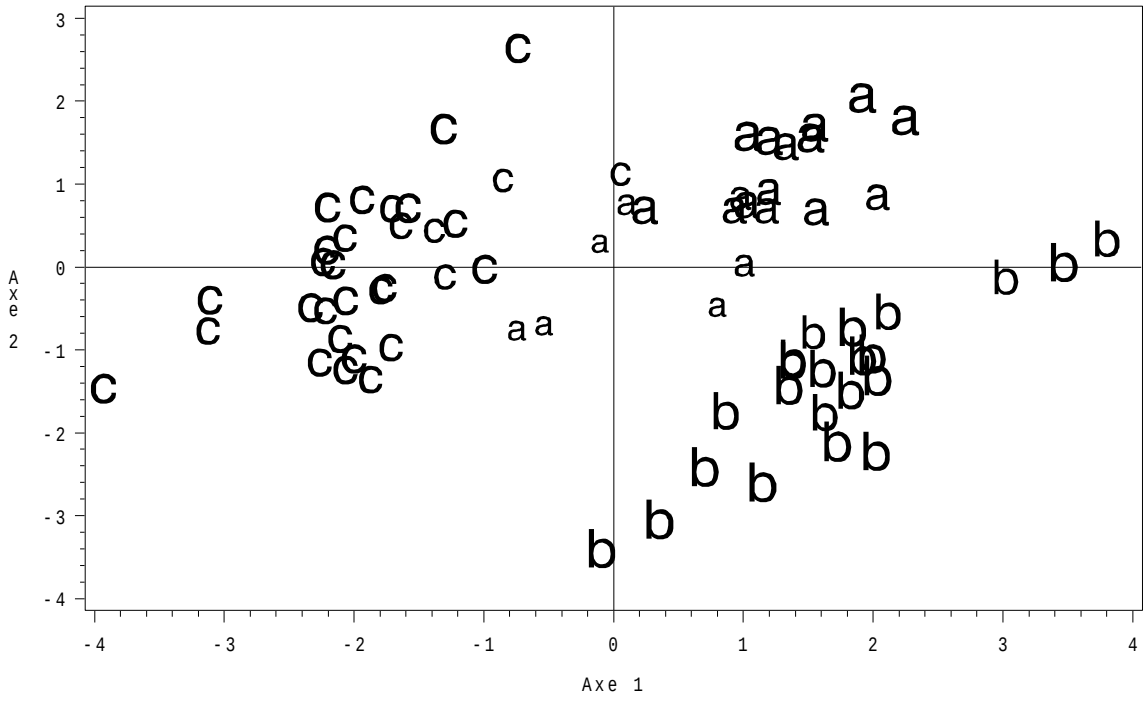


FIG. 5.1: Insectes : premier plan factoriel de l'A.C.P.

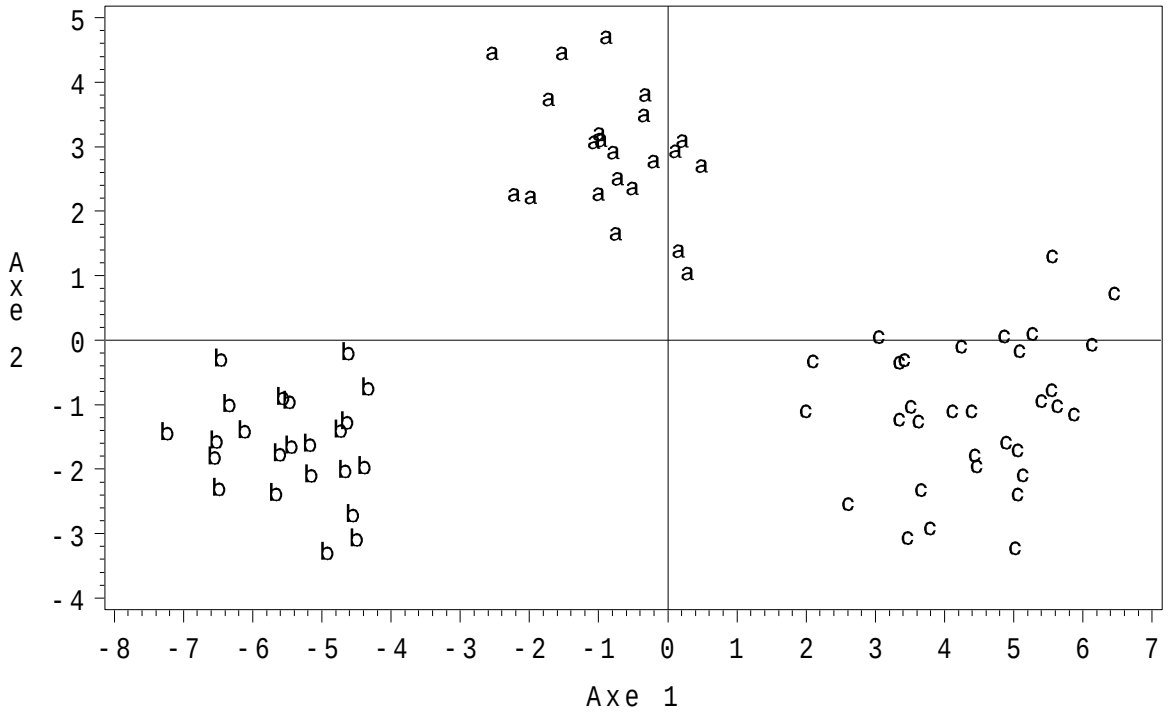


FIG. 5.2: Insectes : premier plan factoriel de l'A.F.D.

Chapitre 6

Analyse Discriminante Décisionnelle

1 Introduction

Dans le chapitre précédent on disposait d'un échantillon Ω de taille n sur lequel étaient observées à la fois les variables explicatives $Y^j, j = 1, \ldots, p$ et la variable T. En général, cet échantillon est appelé *échantillon d'apprentissage*.

On suppose maintenant que l'on dispose d'un nouvel individu (ou de plusieurs, c'est la même chose) sur lequel on a observé les Y^j mais pas T. On doit alors *décider* de la modalité de T_l (ou de la classe correspondante) de ce nouvel individu. On parle aussi de problème d'*affectation*.

Pour cela, on va définir et étudier dans ce chapitre des *règles de décision* (ou d'affectation) et donner ensuite les moyens de les évaluer.

On ne s'intéressera qu'à un seul individu pour décrire ces règles ; on notera $y = (y^1, \ldots, y^p)$ les observations des variables explicatives sur cet individu et $x = (x^1, \ldots, x^p)$ les valeurs centrées :

$$x^j = y^j - \overline{y}^j, \quad \text{avec} \quad \overline{y}^j = \sum_{i=1}^{n} w_i y_i^j.$$

On note que l'observation de Y^j sur le nouvel individu n'intervient pas dans le calcul de $\overline{y}^j$.

2 Règle de décision issue de l'AFD

2.1 Cas général : m quelconque

DÉFINITION 6.1. — *On affectera l'individu x à la modalité de T minimisant :*

$$d^2_{\mathbf{S}_r^{-1}}(x, g_l), l = 1, \ldots, m.$$

Cette distance se décompose en

$$d^2_{\mathbf{S}_r^{-1}}(x, g_l) = \|x - g_l\|^2_{\mathbf{S}_r^{-1}} = (x - g_l)' \mathbf{S}_r^{-1} (x - g_l)$$

et le problème revient donc à maximiser

$$g_l' \mathbf{S}_r^{-1} x - \frac{1}{2} g_l' \mathbf{S}_r^{-1} g_l.$$

Il s'agit bien d'une règle linéaire en x car elle peut s'écrire : $\mathbf{A}_l x + b_l$.

2.2 Cas particulier : $m = 2$

Dans ce cas, la dimension r de l'AFD vaut 1. Il n'y a qu'une seule valeur propre non nulle λ_1, un seul vecteur discriminant v^1 et un seul axe discriminant Δ_1. Les 2 barycentres g_1 et g_2 sont sur Δ_1, de sorte que v^1 est colinéaire à $g_1 - g_2$.

L'application de la règle de décision permet d'affecter x à T_1 si :

$$g_1' \mathbf{S}_r^{-1} x - \frac{1}{2} g_1' \mathbf{S}_r^{-1} g_1 > g_2' \mathbf{S}_r^{-1} x - \frac{1}{2} g_2' \mathbf{S}_r^{-1} g_2$$

c'est-à-dire encore si

$$(g_1 - g_2)' \mathbf{S}_r^{-1} x > (g_1 - g_2)' \mathbf{S}_r^{-1} \frac{g_1 + g_2}{2}.$$

On retrouve ainsi la formulation due à Fisher (1936) dans les premiers travaux sur l'analyse discriminante. Si la norme de la projection de x sur Δ_1 est plus grande que $\frac{g_1+g_2}{2}$ (mid-point) on choisit T_1 sinon on choisit T_2.

2.3 Remarque

La règle de décision liée à l'AFD est simple mais elle est limitée et insuffisante notamment si les variances des classes ne sont pas identiques. De plus, elle ne tient pas compte de l'échantillonnage pour x : tous les groupes n'ont pas nécessairement la même probabilité d'occurence.

3 Règle de décision bayésienne

Dans la suite de ce chapitre, on travaille sur les données brutes et l'observation à affecter sera y et non plus x.

3.1 Introduction

Dans cette optique, on considère que l'échantillon d'apprentissage Ω est issue d'une population plus vaste $\mathcal{T}$ sur laquelle est définie une partition $\{T_1, \ldots, T_l, \ldots, T_m\}$. Les probabilités ou proportions théoriques de ces classes sont notées $\pi_1, \ldots, \pi_m$; dans cette optique bayésienne, il s'agit de probabilités *a priori*.

Ayant observé y, il faut l'affecter à la classe la plus probable, c'est-à-dire celle qui maximise la probabilité $P[T_l \mid y]$: probabilité conditionnelle *a posteriori*.

3.2 Explicitation de la règle

Par le théorème de Bayes, on a :

$$P[T_l \mid y] = \frac{P[T_l \text{et} y]}{P[y]} = \frac{P[T_l].P[y \mid T_l]}{P[y]}$$

avec le principe des probabilités totales : $P[y] = \sum_{l=1}^{m} P[T_l].P[y \mid T_l]$.

Comme $P[y]$ ne dépend pas de l, la règle consistera à choisir T_l maximisant

$$P[T_l].P[y \mid T_l] = \pi_l.P[y \mid T_l];$$

$P[y \mid T_l]$ est la probabilité d'observer y au sein de la classe T_l. Pour une loi discrète, il s'agit d'une probabilité du type $P[y = y_k^l \mid T_l]$ et d'une densité $f(y \mid T_l)$ pour une loi continue. Dans tous les cas nous utiliserons la notation $f_l(y)$.

La règle de décision s'écrit finalement sous la forme :

$$\max_{l=1,\dots,m} \pi_l f_l(y).$$

3.3 Détermination des probabilités

Les probabilités *a priori* π_l peuvent effectivement être connues *a priori* : proportions de divers groupes dans une population, de diverses maladies ... ; sinon elles sont estimées sur l'échantillon d'apprentissage :

$$\widehat{\pi_l} = w_l = \frac{n_l}{n} \quad \text{(si tous les individus ont le même poids).}$$

Les différentes méthodes d'estimation des densités conditionnelles $f_l(y)$ conduisent aux méthodes classiques de discrimination bayésienne objets de la section suivante.

3.4 Cas particuliers

- Dans le cas où les probabilités *a priori* sont égales, c'est par exemple le cas du choix de probabilités non informatives, la règle de décision bayésienne revient alors à maximiser $f_l(y)$ qui est la vraisemblance, au sein de T_l, de l'observation y. La règle consiste alors à choisir la classe pour laquelle cette vraisemblance est maximum.

- Dans le cas où $m = 2$, on affecte y à T_1 si :

$$\frac{f_1(y)}{f_2(y)} > \frac{\pi_2}{\pi_1}$$

  faisant ainsi apparaître un rapport de vraisemblance. D'autre part, l'introduction de coûts de mauvais classement différents selon les classes amène à modifier la valeur limite π_2/π_1.

4 Règle bayésienne avec modèle normal

On suppose dans cette section que, conditionnellement à T_l, $y = (y_1, \dots, y_p)$ est l'observation d'un vecteur aléatoire gaussien $\mathcal{N}(\mu_l, \Sigma_l)$; μ_l est un vecteur de $\mathbb{R}^p$ et Σ_l une matrice $(p \times p)$ symétrique et définie-positive. La densité de la loi, au sein de la classe l, s'écrit donc :

$$f_l(y) = \frac{1}{\sqrt{2\pi}(\det(\Sigma_l))^{1/2}} \exp\left[-\frac{1}{2}(y-\mu_l)'\Sigma_l^{-1}(y-\mu_l)\right].$$

L'affectation de y à une classe se fait en maximisant $\pi_l . f_l(y)$ par rapport à l soit encore la quantité :

$$\ln(\pi_l) - \frac{1}{2}\ln(\det(\Sigma_l)) - \frac{1}{2}(y-\mu_l)'\Sigma_l^{-1}(y-\mu_l).$$

4.1 Hétéroscédasticité

Dans le cas général, il n'y a pas d'hypothèse supplémentaire sur la loi de y et donc les matrices Σ_l sont fonction de l. Le critère d'affectation est alors *quadratique* en y. Les probabilités π_l sont supposées connues mais il est nécessaire d'estimer les moyennes μ_l ainsi que les covariances Σ_l en maximisant, compte tenu de l'hypothèse de normalité, la vraisemblance. Ceci conduit à estimer la moyenne

$$\widehat{\mu_l} = \overline{y_l}$$

par la moyenne empirique de y dans la classe l pour l'échantillon d'apprentissage et Σ_l par la matrice de covariance empirique $\mathbf{S}^*_{Rl}$:

$$\mathbf{S}^*_{Rl} = \frac{1}{n_l - 1} \sum_{i \in \Omega_l} (y_i - \overline{y_l})(y_i - \overline{y_l})'$$

pour ce même échantillon.

4.2 Homoscédasticité

On suppose dans ce cas que les lois de chaque classe partagent la même structure de covariance $\Sigma_l = \Sigma$. Supprimant les termes indépendants de l, le critère à maximiser devient

$$\ln(\pi_l) - \frac{1}{2}\mu_l' \Sigma_l^{-1} \mu_l + \mu_l' \Sigma_l^{-1} y$$

qui est cette fois *linéaire* en y. Les moyennes μ_l sont estimées comme précédemment tandis que Σ est estimée par la matrice de covariance intra empirique :

$$\mathbf{S}^*_R = \frac{1}{n - m} \sum_{l=1}^{m} \sum_{i \in \Omega_l} (y_i - \overline{y_l})(y_i - \overline{y_l})'.$$

Si, de plus, les probabilités π_l sont égales, après estimation le critère s'écrit :

$$\overline{y_l}' \mathbf{S}_R^{*-1} y - \frac{1}{2} \overline{y_l}' \mathbf{S}_R^{*-1} \overline{y_l}.$$

On retrouve alors le critère de la section 2 issu de l'AFD.

4.3 Commentaire

Les hypothèses : normalité, éventuellement l'homoscédasticité, doivent être vérifiées par la connaissance *a priori* du phénomène ou par une étude préalable de l'échantillon d'apprentissage. L'hypothèse d'homoscédasticité, lorqu'elle est vérifiée, permet de réduire très sensiblement le nombre de paramètres à estimer et d'aboutir à des estimateurs plus fiables car de variance moins élevée. Dans le cas contraire, l'échantillon d'apprentissage doit être de taille importante.

5 Règle bayésienne avec estimation non paramétrique

5.1 Introduction

En Statistique, on parle d'estimation non paramétrique ou fonctionnelle lorsque le nombre de paramètres à estimer est infini. L'objet statistique à estimer est alors une fonction par exemple de régression $y = f(x)$ ou encore une densité de probabilité. Dans ce cas, au lieu de supposer qu'on a affaire à une densité de type connu (normale) dont on estime les paramètres, on cherche une estimation $\widehat{f}$ de la fonction de densité f. Pour tout x de $\mathbb{R}$, $f(x)$ est donc estimée par $\widehat{f}(x)$.

Cette approche très souple a l'avantage de ne pas nécessiter d'hypothèse particulière sur la loi (seulement la régularité de f pour de bonnes propriétés de convergence), en revanche elle n'est applicable qu'avec des échantillons de grande taille d'autant plus que le nombre de dimensions p est grand (curse of dimensionality).

Dans le cadre de l'analyse discriminante, ces méthodes permettent d'estimer directement les densités $f_l(y)$. On considère ici deux approches : la méthode du noyau et celle des k plus proches voisins.

5.2 Méthode du noyau

Estimation de densité

Soit $y_1, \ldots, y_n$ n observations équipondérées d'une v.a.r. continue Y de densité f inconnue. Soit $K(y)$ (le *noyau*) une densité de probabilité unidimensionnelle (sans rapport avec f) et h un réel strictement positif. On appelle estimation de f par la méthode du noyau la fonction

$$\widehat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right).$$

Il est immédiat de vérifier que

$$\forall y \in \mathbb{R}, \widehat{f}(y) \geq 0 \quad \text{et} \quad \int_{-\infty}^{+\infty} \widehat{f}(y)dy = 1;$$

h est appelé *largeur de fenêtre* ou paramètre de *lissage*; plus h est grand, plus l'estimation $\widehat{f}$ de f est régulière. Le noyau K est choisi centré en 0, unimodal et symétrique. Les cas les plus usuels sont la densité gaussienne, celle uniforme sur $[-1, 1]$ ou triangulaire: $K(x) = [1 - |x|]\mathbf{1}_{[-1,1]}(x)$. La forme du noyau n'est pas très déterminante sur la qualité de l"estimation contrairement à la valeur de h.

Application à l'analyse discriminante

La méthode du noyau est utilisée pour calculer une estimation non paramétrique de chaque densité $f_l(y)$ qui sont alors des fonctions à valeurs dans $\mathbb{R}^p$. Le noyau K^* dont donc être choisi multidimensionnel et

$$\widehat{f}_l(y) = \frac{1}{n_l h^p} \sum_{i \in \Omega_l} K^*\left(\frac{y - y_i}{h}\right).$$

Un noyau multidimensionnel peut être défini à partir de la densité usuelle de lois: multinormale $\mathcal{N}_p(0, \Sigma_p)$ ou uniforme sur la sphère unité ou encore par produit de noyaux unidimensionnels:

$$K^*(y) = \prod_{j=1}^{p} K(y_j).$$

5.3 k plus proches voisins

Cette méthode d'affectation d'un vecteur y consiste à enchaîner les étapes suivantes:

i. choix d'un entier k : $1 \geq k \geq n$;

ii. calculer les distances $d_{\mathbf{M}}(y, y_i)$, $\quad i = 1, \ldots, n$ où $\mathbf{M}$ est la métrique de Mahalanobis c'est-à-dire la matrice inverse de la matrice de variance (ou de variance intra);

iii. retenir les k observations $y_{(1)}, \ldots, y_{(k)}$ pour lesquelles ces distances sont les plus petites

iv. compter les nombres de fois $k_1, \ldots, k_m$ que ces k observations apparaissent dans chacune des classes,

v. estimer les densités par

$$\widehat{f}_l(y) = \frac{k_l}{kV_k(y)};$$

où $V_k(y)$ est le volume de l'ellipsoïde $\{z | (z - y)'\mathbf{M}(z - y) = d_{\mathbf{M}}(y, y_{(k)})\}$.

Pour $k = 1$, y est affecté à la classe du plus proche élément.

6 Évaluation des règles de décision

On décrit les trois méthodes d'évaluation les plus courantes d'une règle de décision en analyse discriminante. Cette étape est indispensable afin de s'assurer de la fiabilité des résultats.

6.1 Resubstitution

Cette méthode consiste à appliquer la règle de décision choisie sur l'échantillon d'apprentissage on calcule ensuite le taux de mal classées. Il s'appelle le *taux apparent d'erreur* et est fourni en standard par les logiciels. Il n'est pas coûteux en temps de calcul mais sous-estime souvent le taux d'erreur. Il est en effet biaisé car est estimé sur les données qui ont servi à définir la règle de décision. Il est d'autant plus faible que le modèle est complexe (sur-paramétrisation) et que la taille de l'échantillon est faible. Il est peu recommandé.

6.2 Échantillon test

Elle consiste à partager l'échantillon en deux parties : une partie de l'ordre de 80% sert d'échantillon d'apprentissage de la règle de décision, l'autre partie sert à la tester. Cette estimation est plus fiable (non biaisée) mais nécessite un échantillon plus important.

En effectuant plusieurs tirages aléatoires d'échantillons d'apprentissage, on améliore encore l'estimation du taux d'erreur en calculant la moyenne des valeurs obtenues à chaque tirage.

6.3 Validation croisée

Pour tout $i = 1, \ldots, n$ on considère les n échantillons d'apprentissage constitués en éliminant la $i$ème observation. La règle de décision qui en découle est utilisée pour affecter cette $i$ème observation. Le taux d'erreur est estimé en divisant le nombre de mal classés par n.

Les moyens de calcul actuels permettent d'estimer les taux d'erreur en des temps raisonnables, ces techniques itératives doivent être systématiquement mises en œuvre en pratique.

Chapitre 7

Positionnement multidimensionnel

1 Introduction

Considérons n individus. Contrairement aux chapitres précédents, on ne connaît pas les observations de p variables sur ces n individus mais les $1/2n(n-1)$ valeurs d'un indice (de distance, similarité ou dissimilarité) observées ou construites pour chacun des couples d'individus. Ces informations sont contenues dans une matrice $(n \times n)$ $\mathcal{D}$. L'objectif du *positionnement multidimensionnel* (multidimensional scaling ou MDS ou ACP d'un tableau de distances) est de construire, à partir de cette matrice, une représentation euclidienne des individus dans un espace de dimension réduite q qui approche au "mieux" les indices observés.

Exemple : Considérons un tableau avec, en ligne, les individus d'un groupe et en colonne les pays de la C.E. La valeur 1 est mise dans une case lorsque l'individu de la ligne a passé au moins une nuit dans le pays concerné. Il est alors facile de construire une matrice de similarité avec un indice qui compte le nombre de 1 apparaissant dans les mêmes colonnes de tous les couples d'individus. L'objectif est ensuite d'obtenir une représentation graphique rapprochant les individus ayant visité les mêmes pays.

Les preuves des propositions sont omises dans cet exposé succinct, elles sont à chercher dans la bibliographie. Voir par exemple Mardia et col. (1979).

2 Distance, similarités

2.1 Définitions

Définition 7.1. —

- *Une matrice* $(n \times n)$ $\mathcal{D}$ *est appelée matrice de distance si elle est symétrique et si :*

$$d_j^j = 0 \text{ et } \forall (j,k), j \neq k, d_j^k \geq 0.$$

- *Une matrice* $(n \times n)$ $\mathcal{C}$ *est appelée matrice de similarité si elle est symétrique et si*

$$\forall (j,k), c_j^k \leq c_j^j.$$

Une matrice de similarité se transforme en matrice de distance par :

$$d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}.$$

DÉFINITION 7.2. — *Une matrice de distance est dite euclidienne s'il existe une configuration de vecteurs* $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ *dans un espace euclidien* E *de sorte que*

$$d_j^{k\,2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle \ .$$

On note $\mathbf{A}$ la matrice issue de $\mathcal{D}$ de terme général $d_j^k = -1/2 d_j^{k\,2}$ et $\mathbf{H}$ la matrice de centrage :

$$\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D},$$

qui est la matrice de projection sur le sous-espace $\mathbf{D}$-orthogonal au vecteur $\mathbf{1}$ dans l'espace euclidien F des variables muni de la métrique des poids.

PROPOSITION 7.3. —

- *Soit* $\mathcal{D}$ *une matrice de distance et* $\mathbf{B}$ *la matrice obtenue par double centrage de la matrice* $\mathbf{A}$ *issue de* $\mathcal{D}$ *:*

  $$\mathbf{B} = \mathbf{HAH},$$

  alors $\mathcal{D}$ *est une matrice euclidienne si et seulement si* $\mathbf{B}$ *est positive (toutes ses valeurs propres sont positives ou nulles).*

- *Si la matrice de similarité* $\mathcal{C}$ *est positive alors la matrice de distance* $\mathcal{D}$ *déduite est euclidienne.*

3 Distances entre variables

L'un des intérêts pratique du positionnement multidimensionnel est d'aider à comprendre, visualiser, les structures de liaison dans un grand ensemble de variables. On obtient ainsi des indications pour guider le choix d'un sous-ensemble de variables, par exemple les plus liées à une variable à expliquer. Cette approche nécessite la définition d'indices de similarité entre variables. Beaucoup sont proposés dans la littérature. Nous en retenons trois pour différents types de variables.

3.1 Variables quantitatives

On note X et Y deux variables statistiques dont les observations sur les mêmes n individus sont rangées dans les vecteurs $\mathbf{x}$ et $\mathbf{y}$ de l'espace euclidien F muni de la métrique des poids $\mathbf{D}$. On vérifie facilement :

$$\begin{aligned} \mathrm{cov}(X,Y) &= (\mathbf{x} - \overline{\mathbf{x}})'\mathbf{D}(\mathbf{y} - \overline{\mathbf{y}}) = \mathbf{x}'\mathbf{D}\mathbf{y} - \overline{\mathbf{x}}\overline{\mathbf{y}} \\ \sigma_X &= \|\mathbf{x} - \overline{\mathbf{x}}\|_{\mathbf{D}} \\ \mathrm{cor}(X,Y) &= \frac{\mathbf{x}'\mathbf{D}\mathbf{y} - \overline{\mathbf{x}}\overline{\mathbf{y}}}{\|\mathbf{x} - \overline{\mathbf{x}}\|_{\mathbf{D}} \, \|\mathbf{y} - \overline{\mathbf{y}}\|_{\mathbf{D}}}. \end{aligned}$$

La valeur absolue ou le carré du coefficient de corrélation définissent des indices de similarité entre deux variables quantitatives. Il est facile d'en déduire des distances. On préfère par la suite utiliser le carré du coefficient de corrélation qui induit une distance euclidienne :

$$d^2(X,Y) = 2(1 - \mathrm{cor}^2(X,Y)).$$

PROPOSITION 7.4. — *La distance entre variables quantitatives* $d^2(X,Y)$ *est encore le carré de la distance* $\|\mathbf{P_x} - \mathbf{P_y}\|_{\mathbf{D}}$ *entre les projecteurs* $\mathbf{D}$*-orthogonaux sur les directions engendrées par les vecteurs* $(\mathbf{x} - \overline{\mathbf{x}})$ *et* $(\mathbf{y} - \overline{\mathbf{y}})$.

Démonstration. — Un projecteur de rang 1 s'écrit :

$$\mathbf{P_x} = \frac{(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})'}{\|\mathbf{x}-\overline{\mathbf{x}}\|^2_{\mathbf{D}}}\mathbf{D}.$$

$$\begin{aligned}\|\mathbf{P_x}-\mathbf{P_y}\|^2_{\mathbf{D}} &= \mathrm{tr}(\mathbf{P_x}-\mathbf{P_y})'\mathbf{D}(\mathbf{P_x}-\mathbf{P_y}) \\ &= \|\mathbf{P_x}\|^2_{\mathbf{D}} + \|\mathbf{P_y}\|^2_{\mathbf{D}} - 2\mathrm{tr}\mathbf{P_x}'\mathbf{D}\mathbf{P_y}.\end{aligned}$$

Comme un projecteur est de norme son rang c'est-à-dire ici 1 et que :

$$\begin{aligned}\mathrm{tr}\mathbf{P_x}'\mathbf{D}\mathbf{P_y} &= \mathrm{tr}\frac{(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})'}{\left\|\mathbf{x}-\overline{\mathbf{x}}^2\right\|_{\mathbf{D}}}\mathbf{D}\frac{(\mathbf{y}-\overline{\mathbf{y}})(\mathbf{y}-\overline{\mathbf{y}})'}{\left\|\mathbf{y}-\overline{\mathbf{y}}^2\right\|_{\mathbf{D}}}\mathbf{D} \\ &= \frac{(\mathbf{x}-\overline{\mathbf{x}})'\mathbf{D}(\mathbf{y}-\overline{\mathbf{y}})}{\|\mathbf{x}-\overline{\mathbf{x}}\|_{\mathbf{D}}\|\mathbf{y}-\overline{\mathbf{y}}\|_{\mathbf{D}}}\frac{(\mathbf{x}-\overline{\mathbf{x}})'\mathbf{D}(\mathbf{y}-\overline{\mathbf{y}})}{\|\mathbf{x}-\overline{\mathbf{x}}\|_{\mathbf{D}}\|\mathbf{y}-\overline{\mathbf{y}}\|_{\mathbf{D}}} \\ &= \mathrm{cor}^2(X,Y)\end{aligned}$$

alors,

$$\|\mathbf{P_x}-\mathbf{P_y}\|^2_{\mathbf{D}} = 2(1-\mathrm{cor}^2(X,Y)).$$

∎

3.2 Variables qualitatives

Considérons maintenant deux variables qualitatives, X à r modalités et Y à c modalités. De nombreux indices de similarité ont été proposés : la "prob value" du test du χ^2 d'indépendance, le V de Cramer, le Φ^2 de Pearson, le T de Tschuprow (cf. T1) ... Ce dernier a une signification particulière. Soit $\mathbf{X}$ et $\mathbf{Y}$ les matrices contenant les variables indicatrices des modalités des variables et $\mathbf{P_X}$, $\mathbf{P_Y}$ les projecteurs $\mathbf{D}$-orthogonaux sur les sous-espaces engendrés par ces indicatrices. On montre (cf. Saporta 1976) alors la

Proposition 7.5. — *Dans le cas de 2 variables qualitatives,*

$$\|\mathbf{P_X}-\mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1-T^2(X,Y)).$$

Ainsi, en utilisant comme indice de similarité le carré du T de Tschuprow entre deux variables qualitatives, on définit une distance euclidienne entre ces variables.

3.3 Variables quantitative et qualitative

La même démarche s'adapte à l'étude d'une liaison entre une variable quantitative X, son projecteur associé $\mathbf{P_x}$ et une variable qualitative Y représentée par le projecteur $\mathbf{P_Y}$. On montre alors (cf. Saporta 1976)

Proposition 7.6. — *Dans le cas d'une variable quantitative X et d'une variable qualitative Y,*

$$\|\mathbf{P_x}-\mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1-R^2_c(X,Y))$$

où R_c désigne le rapport de corrélation.

Le rapport de corrélation (Cf. T1) est, dans ce cas, l'indice de similarité qui conduit à la construction d'une distance euclidienne entre variables de types différents.

On aboutit ainsi à une certaine généralisation de la notion de similarité entre variables conduisant, quelque soit le type des variables, à des distances euclidiennes. Néanmoins, en pratique, il n'apparaît pas simple de comparer, sur la même échelle entre 0 et 1, des liaisons entre variables de types différents. Les coefficients de corrélations se répartissent plus communément sur toute l'échelle alors que les indices de Tschuprow sont souvent confinés sur des petites valeurs.

4 Recherche d'une configuration de points

Le positionnement multidimensionnel est la recherche d'une configuration de points dans un espace euclidien qui admette $\mathcal{D}$ comme matrice de distances si celle-ci est euclidienne ou, dans le cas contraire, qui en soit la meilleure approximation à un rang q fixé (en général 2) au sens d'une norme sur les matrices. Nous ne nous intéressons dans ce chapitre qu'à la version "métrique" du MDS, une autre approche construite sur les rangs est développée dans la bibliographie.

Ainsi posé, le problème admet une infinité de solutions. En effet, la distance entre deux vecteurs $\mathbf{x}_i$ et $\mathbf{x}_k$ d'une configuration est invariante par toute transformation affine $\mathbf{z}_i = \mathbf{F}\mathbf{x}_i + \mathbf{b}$ dans laquelle $\mathbf{F}$ est une matrice orthogonale quelconque et $\mathbf{b}$ un vecteur de $\mathbb{R}^p$. Une solution n'est donc connue qu'à une rotation et une translation près.

4.1 Propriétés

La solution est décrite dans les théorèmes (Mardia 1979) ci-dessous :

THÉORÈME 7.7. — *Soit $\mathcal{D}$ une matrice de distance et $\mathbf{B} = \mathbf{HAH}$ la matrice centrée en lignes et colonnes associée.*

- *Si $\mathcal{D}$ est la matrice de distance euclidienne d'une configuration $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ alors $\mathbf{B}$ est la matrice de terme général*

$$b_j^k = (\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}})$$

  qui se met sous la forme

$$\mathbf{B} = (\mathbf{HX})(\mathbf{HX})'.$$

  Elle est donc positive et appelée matrice des produits scalaires de la configuration centrée.

- *Réciproquement, si $\mathbf{B}$ est positive de rang p, une configuration de vecteurs admettant $\mathbf{B}$ pour matrice des produits scalaires est obtenue en considérant sa décomposition spectrale $\mathbf{B} = \mathbf{U}\Delta\mathbf{U}'$. Ce sont les lignes de la matrice centrée $\mathbf{X} = \mathbf{U}\Delta^{1/2}$.*

Ainsi, dans le cas d'une matrice $\mathcal{D}$ euclidienne supposée de rang q, la solution est obtenue en exécutant les étapes suivantes :

i. construction de la matrice $\mathbf{A}$ de terme général $-1/2{d_j^k}^2$,

ii. calcul de la matrice des produits scalaires par double centrage $\mathbf{B} = \mathbf{HAH}$,

iii. diagonalisation de $\mathbf{B} = \mathbf{U}\Delta\mathbf{U}'$;

iv. les coordonnées d'une configuration, appelées *coordonnées principales*, sont les lignes de la matrice $\mathbf{X} = \mathbf{U}\Delta^{1/2}$.

Dans le cas euclidien, ACP et MDS sont directement connectés.

PROPOSITION 7.8. — *Soit $\mathbf{Y}$ la matrice des données habituelles en ACP. L'ACP de $(\mathbf{Y}, \mathbf{M}, 1/n\mathbf{I})$ fournit les mêmes représentations graphiques que le positionnement calculé à partir de la matrice de distances de terme général $\|\mathbf{y}_i - \mathbf{y}_j\|_{\mathbf{M}}$. Si $\mathbf{C}$ désigne la matrice des composantes principales, alors les coordonnées principales sont $\sqrt{n}\mathbf{C}$.*

Démonstration. — Posons $\mathbf{X} = \mathbf{HY}$. Les composantes principales de l'ACP sont données par

$$\mathbf{C} = \mathbf{XMV} = \mathbf{U}\Lambda^{1/2}$$

où $\mathbf{V}$ est la matrice des vecteurs propres de la matrice $1/n\mathbf{X}'\mathbf{X}\mathbf{M}$ et $\mathbf{U}$ ceux des vecteurs propres de la matrice $1/n\mathbf{X}\mathbf{M}\mathbf{X}'$ associés aux mêmes valeurs propres Λ. De son côté, le MDS conduit à considérer la matrice des produits scalaires $\mathbf{HYM}(\mathbf{HY})' = \mathbf{X}\mathbf{M}\mathbf{X}'$ qui amène aux mêmes vecteurs propres et aux valeurs propres $\Delta = \sqrt{n}\Lambda$. ∎

L'intérêt du MDS apparaît évidemment lorsque les observations $\mathbf{Y}$ sont inconnues ou encore si l'on cherche la meilleure représentation euclidienne de distances non-euclidiennes entre les individus ; c'est l'objet du théorème suivant. En ce sens, le MDS "généralise" l'ACP et permet, par exemple, de considérer une distance de type robuste à base de valeurs absolues mais la représentation des variables pose alors quelques soucis car le "biplot" n'est plus linéaire (Gower 19xx).

THÉORÈME 7.9. — *Si $\mathcal{D}$ est une matrice de distance, pas nécessairement euclidienne, $\mathbf{B}$ la matrice de produit scalaire associée, alors, pour une dimension q fixée, la configuration issue du MDS a une matrice de distance $\widehat{\mathcal{D}}$ qui rend $\sum_{j,k=1}^{n}(\{d_j^k\}^2 - \widehat{d_j^k}^2)$ minimum et, c'est équivalent, une matrice de produit scalaire $\widehat{\mathbf{B}}$ qui minimise $\left\|\mathbf{B} - \widehat{\mathbf{B}}\right\|^2$.*

5 Exemple

Cet exemple s'intéresse aux distances kilométriques par route (Source : IGN) entre 47 grandes villes en France et dans les pays limitrophes. Toutes ces valeurs sont rangées dans le triangle inférieur d'une matrice carrée avec des 0 sur la diagonale. La structure du réseau routier fait que cette matrice de distance n'est pas euclidienne, mais, comme le montre le graphique issu d'un positionnement multidimensionnel, l'approximation euclidienne en est très proche.

6 Application au choix de variables

La sélection d'un sous-ensemble de variables pour la mise en œuvre de techniques factorielles (Jolliffe 19xx) n'est pas aussi claire que dans le cadre de la recherche d'un modèle linéaire parcimonieux. Le problème vient souvent de la confusion de deux objectifs :

- supprimer des variables très liées, donc redondantes, et dont la multiplicité vient renforcer artificiellement l'influence de certains phénomènes,
- supprimer des variables afin de simplifier l'interprétation des axes tout en conservant au mieux les représentations graphiques.

Le premier objectif modifie donc les représentations en visant à être plus proche de la "réalité" ou au moins d'une réalité moins triviale tandis que, par principe, le deuxième objectif recherche le sous-ensemble restreint de variables susceptibles d'engendrer le même sous-espace de représentation.

Il n'existe pas de solution miracle néanmoins, les outils présentés dans ce chapitre : indices de similarité entre variable et positionnement multidimensionnel, peuvent aider à ces choix surtout lorsque l'analyse d'un grand nombre de variables nécessite de segmenter l'analyse en sous-groupes. Les algorithmes de classification (hiérarchique ou centres mobiles) appliqués sur les mêmes tableaux de distance apportent un éclairage complémentaire.

D'autres techniques sont également disponibles pour aider à l'interprétation des axes. Elles ont été développées dans le cadre de l'analyse en facteurs communs et spécifiques (factor analysis) mais sont transposables en ACP. L'objectif est la recherche de rotations orthogonales (varimax) ou obliques des axes dans le sous-espace retenu pour la représentation de sorte que ceux-ci soient le plus corrélés avec les variables initiales. Ils n'ont plus les mêmes propriétés optimales d'axes de

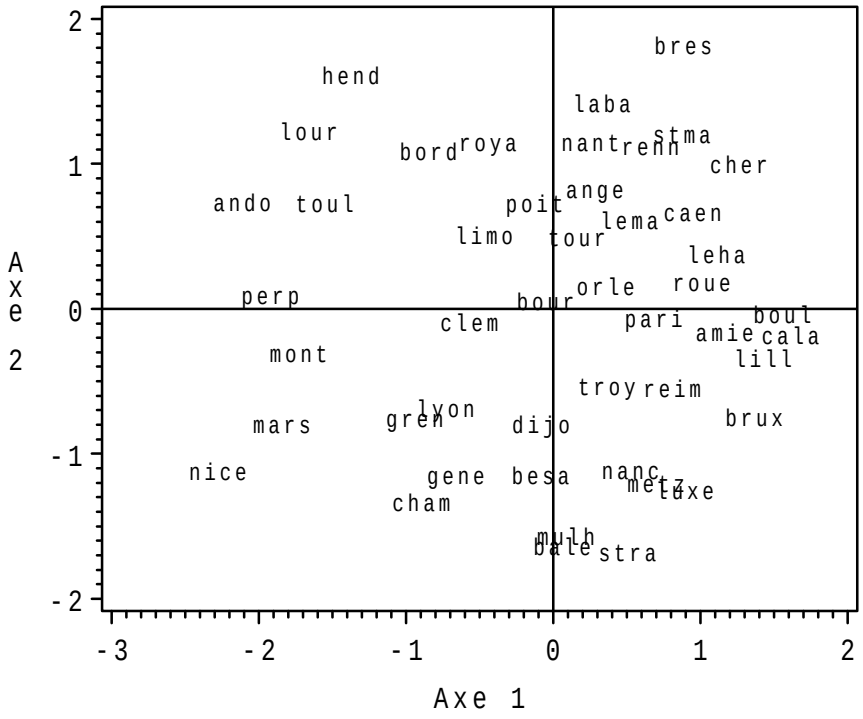


FIG. 7.1: Positionnement de 47 villes à partir de la matrice de leurs distances kilométriques.

plus grande dispersion mais, dans le sous-espace qui globalement est de plus grande dispersion, ils peuvent être plus simples à interpréter à partir des variables initiales.

Un algorithme (`varclus` dans SAS) de classification des variables dans le cas quantitatif suit ce même type d'objectifs et fournit des résultats sous une forme identique à la recherche d'une rotation oblique. Il procède par classification hiérarchique descendante de l'ensemble des variables et réalise à chaque étape les traitements suivants :

- sélection du sous-groupe de variable dont l'ACP conduit à la plus faible part de variance expliquée par le premier axe ou (en option) la plus forte du 2ème axe,
- rotation des deux premiers axes de l'ACP pour les rapprocher des variables et segmentation des variables en deux groupes par affectation à l'axe avec lequel elles sont le plus corrélées.

L'algorithme s'arrête lorsque la dimension dans chaque groupe est jugée être égale à 1. Par défaut, lorsque dans chaque groupe, une seule valeur propre est plus grande que 1.

Bibliography

Bouroche, J. and Saporta, G. (1980). *L'Analyse des Données*. Que Sais-je, PUF.

Bry, X. (1995). *Analyses factorielles simples*. Economica.

Bry, X. (1996). *Analyses factorielles multiples*. Economica.

Caillez, F. and Pages, J. (1976). *Introduction à l'Analyse des Données*. SMASH.

Droesbeke, J., Fichet, B., and Tassi, P. (1992). *Modèles pour l'Analyse des Données Multidimensionnelles*. Economica.

Everitt, B. and Dunn, G. (1991). *Applied Multivariate Data Analysis*. Edward Arnold.

Goodman, L. (1991). Measures, models, and graphical displays in the analysis of cross–classified data. *Journal of the American Statistical Association*, 86:1085–1138.

Jobson, J. (1992). *Applied Multivariate Data Analysis*, volume II : Categorical and multivariate methods. Springer-Verlag.

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

Lebart, L., Morineau, A., and Piron, M. (1995). *Statistique exploratoire multidimensionnelle*. Dunod.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Saporta, G. (1990). *Probabilités, Analyse des Données et Statistique*. Technip.

SAS (1989). *SAS/STAT User's Guide*, fourth edition. version 6.

Annexe A

Outils algébriques

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 :

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

Déterminant

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

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

Cette dernière propriété se montre en considérant les matrices :

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

puis en comparant les déterminants $|\mathbf{B}\mathbf{A}\mathbf{B}'|$ 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 :

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

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 $\mathbf{A}\mathbf{A} = \mathbf{A}$,

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

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 $F = \mathbb{R}^n$.

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

Rang

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

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 $\mathbf{B}'\mathbf{AB}$ 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 ($\mathbf{P}'\mathbf{M} = \mathbf{M}\mathbf{P}$) est une matrice de projection $\mathbf{M}$-orthogonale et réciproquement.

Propriétés

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

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

- $\mathrm{tr}\mathbf{P} = \dim(W)$.
- $\mathbf{P} = \mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-1}\mathbf{B}'\mathbf{M}$, où $\mathbf{B} = \left[b^1, \ldots, b^q\right]$.
- Dans le cas particulier où les b^j sont $\mathbf{M}$-orthonormés :

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

- 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 :*

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

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* Λ*.*

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 :*

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

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* Λ*.*

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 Λ 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 $\mathbf{A}$:

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

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,	$\lambda_p > 0$;

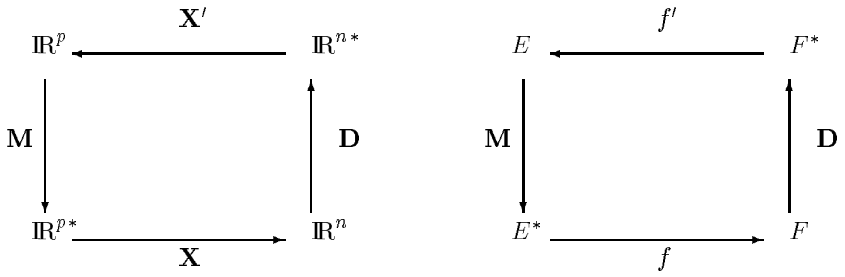


FIG. A.1: *Schéma de dualité*

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 :*

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

$\mathbf{U}$ $(n \times r)$ *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* $\Lambda(r \times r)$ *;* $\mathbf{V}$ $(p \times r)$ *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,*

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

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 $\mathrm{vec}(\mathbf{X}) = \left[x^{1'}, \ldots, x^{p'}\right]'$ 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 $\mathbf{X}$:

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

Le minimum atteint est :

$$\|\mathbf{X} - \mathbf{Z}_q\|^2_{\mathbf{M},\mathbf{D}} = \sum_{k=q+1}^{r} \lambda_k .$$

Les matrices $\mathbf{U}_q, \Lambda_q$ 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 reformuler 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 $F_q = \mathrm{Im}(\mathbf{U}_q)$) :

$$\begin{aligned} \widehat{\mathbf{P}_q} &= \sum_{k=1}^{q} v^k v^{k'} \mathbf{M} = \mathbf{V}_q \mathbf{V}_q' \mathbf{M} \\ \widehat{\mathbf{Q}_q} &= \sum_{k=1}^{q} u^k u^{k'} \mathbf{D} = \mathbf{U}_q \mathbf{U}_q' \mathbf{D}, \\ \mathbf{Z}_q &= \widehat{\mathbf{Q}_q}\mathbf{X} = \mathbf{X}\widehat{\mathbf{P}_q}' . \end{aligned}$$

Proposition A.6. — *Avec les notations précédentes :*

$$\begin{aligned} \widehat{\mathbf{P}_q} &= \mathit{arg} \max_{\mathbf{P}_q} \left\{ \left\| \mathbf{X}\mathbf{P}_q' \right\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\ & \qquad \left. \mathbf{P}_q \ \mathit{projection}\ \mathbf{M}\text{-}\mathit{orthogonale\ de\ rang}\ q < r \right\}, \\ \widehat{\mathbf{Q}_q} &= \mathit{arg} \max_{\mathbf{Q}_q} \left\{ \|\mathbf{Q}_q\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\ & \qquad \left. \mathbf{Q}_q \ \mathit{projection}\ \mathbf{D}\text{-}\mathit{orthogonale\ de\ rang}\ q < r \right\}. \end{aligned}$$

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 A.F.C. des exploitations agricoles

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 A.F.C.M. 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